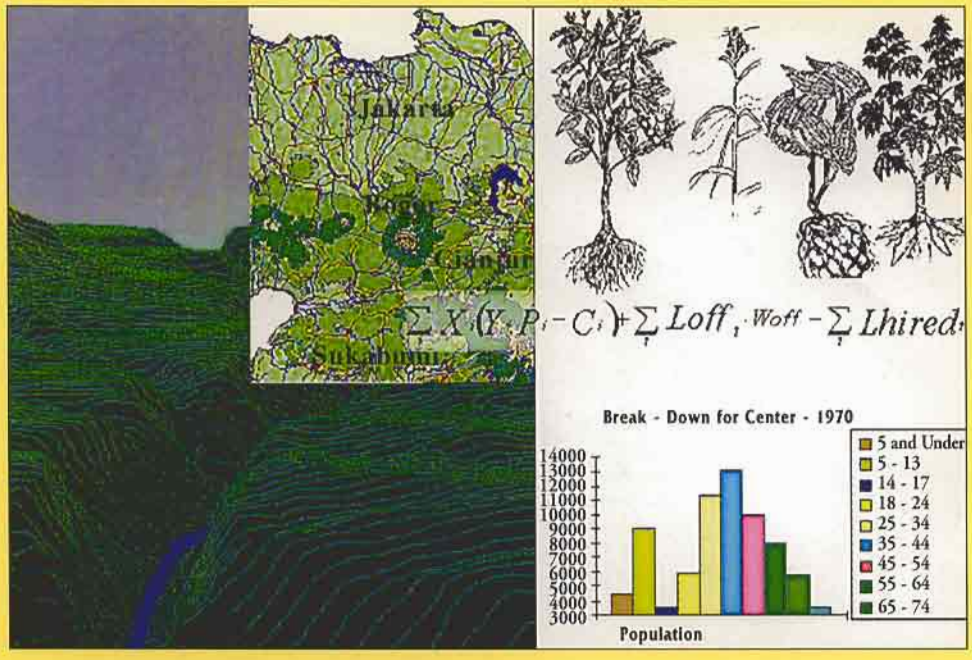
CGPRT NO. 35

Database Management and Analytical Techniques for Agricultural Planning:

A Course Manual



The CGPRT Centre

The CGPRT Centre

The Regional Co-ordination Centre for Research and Development of Coarse Grains, Pulses, Roots and Tuber Crops in the Humid Tropics of Asia and the Pacific (CGPRT Centre) was established in 1981 as a subsidiary body of UN/ESCAP.

Objectives

In co-operation with ESCAP member countries, the Centre will initiate and promote research, training and dissemination of information on socio-economic and related aspects of CGPRT crops in Asia and the Pacific. In its activities, the Centre aims to serve the needs of institutions concerned with planning, research, extension and development in relation to CGPRT crop production, marketing and use.

Programmes

In pursuit of its objectives, the Centre has two interlinked programmes to be carried out in the spirit of technical cooperation among developing countries:

1. Research and development which entails the preparation and implementation of projects and studies covering production, utilization and trade of CGPRT crops in the countries of Asia and the South Pacific.
2. Human resource development and collection, processing and dissemination of relevant information for use by researchers, policy makers and extension workers.

CGPRT Centre Monographs currently available:

CGPRT No. 16 *Maize Production in Sri Lanka*
by N.F.C. Ranaweera, G.A.C. de Silva, M.H.J.P. Fernando and H.B. Hindagala

CGPRT No. 17 *Sistem Komoditas Kedelai di Indonesia*

CGPRT No. 18 *Socio-Economic Constraints to Pulse Production in Nepal*
by M.K. Khatiwada, S.K. Poudel and D.K. Gurung

CGPRT No. 19 *Agricultural Marketing in a Transmigration Area in Sumatra*
by Yujiro Hayami, Toshihiko Kawagoe, Yoshinori Morooka, Henny Mayrowani and Mat Syukur

CGPRT No. 20 *Sensitivity of Soybean Production to Price Changes: A Case Study in East Java*
by Heriyanto, Ruly Krisdiana, A. Ghozi Manshuri and Irlan Soejono

CGPRT No. 21 *Potato in Indonesia: Prospects for Medium Altitude Production*
by J.W. Taco Bottema, Hoky Siregar, Sahat M. Pasaribu, Govert Gijsbers and Rofik S. Basuki

CGPRT No. 22 *Upland Economy in Java: A Perspective of a Soybean-based Farming System*
by Yoshinori Morooka and Henny Mayrowani

CGPRT No. 23 *Role of Secondary Crops in Employment Generation: A Study in a Rain-fed Lowland Village in Java*
by Toshihiko Kawagoe, Koichi Fujita, Shigeki Yokoyama, Wayan Sudana and Amar Kadar Zakaria

CGPRT No. 24 *Sweet Potato in Viet Nam, Production and Markets*
by J.W. Taco Bottema, Pham Thanh Binh, Dang Thanh Ha, Mai Thach Hoanh and H. Kim

(Continued on inside back cover)

Database Management and Analytical Techniques for Agricultural Planning:

A Course Manual

CGPRT NO. 35

Database Management and Analytical Techniques for Agricultural Planning:

A Course Manual

Edited by
J.W.T. Bottema
Siemon Hollema
Mohammad A.T. Chowdhury

CGPRT Centre
Regional Co-ordination Center for
Research and Development of Coarse Grains,
Pulses, Roots and Tuber Crops in the
Humid Tropics of Asia and the Pacific

CGPRT No. 25 *Marketing Innovation for Vegetables: Conditions of Diversification in Upland Farming*
by Yujiro Hayami, Toshihiko Kawagoe, Shigeki Yokoyama, Al Sri Bagyo and Amar Kadar Zakaria

CGPRT No. 26 *Rural employment and Small-scale Rural Food Processing in Asia*
edited by Aida R. Librero and Charles van Santen

CGPRT No. 27 *Local Soybean Economies and Government Policies in Thailand and Indonesia*
by Pattana Jierwiriyapant, Hermanto, Frederic Roche and J.W. Taco Bottema

CGPRT No. 28 *Changes in Food Consumption in Asia: Effects on Production and Use of Upland Crops*
edited by J.W.T. Bottema, G.A.C. De Silva and D.R. Stoltz

CGPRT No. 29 *Marketing and Processing of Food Legumes and Coarse Grains: Effects on Rural Employment in Asia*
Edited by T. Napitupulu, J.W.T. Bottema and D.R. Stoltz

CGPRT No. 30 *Upland Agriculture in Asia: Proceedings of a Workshop Held in Bogor, Indonesia April 6-8, 1993*
Edited by J.W.T. Bottema and D.R. Stoltz.

CGPRT No. 31 *Farmers and Traders in a Changing Maize Market in East Java*
Hitoshi Yonekura

CGPRT No. 32 *Integrating Seed Systems for Annual Food Crops*
Edited by H. van Amstel, J.W.T. Bottema, M. Sidik and C.E. van Santen

CGPRT No. 33 *Women in Upland Agriculture in Asia*
Edited by C.E. van Santen, J.W.T. Bottema and D.R. Stoltz

CGPRT No. 34 *Market Prospects for Upland in Asia*
Edited by Sotaro Inoue, Boonjit Titapiwatanakun and D.R. Stoltz

This series is published by the CGPRT Centre, Bogor. For further information, please contact: Publication Section, CGPRT Centre, Jl. Merdeka 145, Bogor 16111, Indonesia

CGPRT CENTRE
Publication Section

Editor: Douglas R. Stoltz

Production: Agustina Mardyanti
S. Tayanih (Yayan)

Distribution: Fetty Prihastini

Printer: SMK. Grafika Desa Putera

Table of Contents

Foreword

The importance of databases and their management in agriculture is well recognized. The significant gains in agricultural productivity in the last few decades have been achieved in part through improvement in the transfer of knowledge and information to farmers, traders, researchers and policy-makers. The provision of accurate and timely information enables and rationalizes the decision-making processes of managers in a cost effective manner. Advancement of information systems is, therefore, of vital importance in agricultural planning and development.

The primary reason for this importance derives from the fact that management is the key to success or failure of an organization. Deliberate sourcing, organization and production of information is critical for making the managerial decisions required in business, research and government enterprises. Effective database management is essential to any research system from the level of the research institute up to the national level. Without an effective database management system, there will be lack of adequate coordination within the research system, duplication of effort, lack of continuity in building a knowledge base, and inefficiencies in utilization of limited resources. International Services for National Agricultural Research (ISNAR) determined that investing resources in strengthening database management could have a high payoff for national agricultural research systems (NARS). It is believed that advances in microcomputer technology and the growing availability of microcomputers in NARS offer opportunities for providing NARS managers with improved information systems and better management tools.

The CGPRT Centre has implemented the following eight projects since 1989 under the human resources development and information services programme (HRD/IS):

1.	Strengthening of the national and regional statistical database system of CGPRT crops (RSDS)	(1989 ~ 1992)
2.	Strengthening the human resources development and information services programme of the CGPRT Centre (HRD/IS)	(1991 ~ 1995)
3.	Agricultural economist: Strengthening the CGPRT Centre's research and training activities	(1992 ~ 1995)
4.	Training course in market research and survey relating CGPRT crop development	(1993 ~ 1995)
5.	Training in socio-economic methodologies for agriculture research, with specific reference to upland agriculture	(1994 ~ 1995)
6.	Training in socio-economic research and policy planning in CGPRT crops	(1994 ~ 1998)
7.	Strengthening training activities of the CGPRT Centre	(1995 ~ 1998)
8.	Hands-on training in database management and application relating to CGPRT crop development	(1995 ~ 1998)

Course materials were prepared for each training activity to meet its curriculum. Addition and revision were repeated to improve the materials. They are now compiled in a comprehensive syllabus under the title **Database Management and Analytical Techniques for Agricultural Planning: A Course Manual**. Consequently, this manual is to be considered a collective product of the above projects. I hope this manual will be useful in relevant activities.

Many staff were involved in the projects and in drafting, compiling and revising the manuscripts all along the way. The authors of the corresponding chapters are deeply acknowledged firstly. Many thanks go to Dr. Ir. J.W.T. Bottema, the former programme leader of HRD/IS, Mr. Siemon Hollema, the operations research specialist, and Dr. Mohammad A.T. Chowdhury, the assistant training officer, for their devoted effort in completing the manuscripts. Thanks are also due to Dr. Douglas R. Stoltz for editing and Ms. Augustina Mardyanti for her typing.

Finally, I thank the Government of the Netherlands and the Government of Japan for funding the projects.

September 1, 1998

Haruo Inagaki
Director
CGPRT Centre

Introduction

Considering the growing importance of agricultural information systems, this manual was compiled to include topics ranging from database management to appropriate analytical techniques with special reference to agriculture. The objectives of this syllabus are, therefore, two-fold. The first is to provide a general introduction to the important information management tools in a manner that permits nonmathematically-inclined students to easily grasp basic database concepts. This introduction should prove quite useful and interesting not only to agricultural economists, agronomists and statisticians, but also to other natural and social scientists. The second, and equally important objective, is to provide numerous examples for application of database management tools and analytical techniques in the context of agriculture.

To achieve these objectives, this manual contains sections on database management from an institutional point of view, relational database systems, geographic information systems, use of spreadsheets, the mathematical backgrounds of sets, and statistical techniques as inferential and descriptive measures. This line of presentation is continued with the development of hands-on training modules in agriculture, drawing from a wide range of software including setting up a relational database (in Access); creating a spatial database (in MapInfo); using spreadsheets in agriculture (in Excel); and introducing operations research and simulation modeling techniques (using Excel and GAMS). Throughout the course, attempts have been made to provide the reader with real world examples. It is expected that students will be able to recognize problems in their jobs that can be solved using the techniques explained.

This syllabus is primarily designed as a common training manual of the UN/ESCAP CGPRT Center for a two-week course usually conducted at the Center. However, it can be well adapted even for a one-week course. Unless the students have an excellent mathematical and computer background, instructors may wish to omit certain sections, for example, linear programming and mathematical programming in GAMS. If the time available for study of linear programming is limited, it may be desirable to concentrate more on the use of spreadsheets or to extend the database and GIS modules vertically. Notes on integrated database management could be used as opening remarks of the course. Sections on sets and the algebra of sets could be integrated with the database concepts, and can be studied simultaneously with the thematic mapping and GIS overlay concepts. The section on statistics could either be used as background information or be deferred to a refresher course in quantitative methods.

Integrated Database Management for Agricultural Planning and Research

J.W.T Bottema[*]

Introduction

This paper concerns management. The first section will focus on concepts and definitions. The second will deal with broader issues of management and more specifically with management in Asia. The third section will deal with the role of information in state activities and the final section will deal with data availability and the use of integrated database management to improve efficiency in public organizations.

The state plays an important role in Asian economies and especially in developing Asian economies. However, there are substantial pressures on the government apparatus and its bureaucracies in developing countries. The state apparatus is not very strong in these economies and yet it is the major channel for public investment. Development implies investment, and investment implies project selection and project implementation. When projects are implemented they require monitoring and measurement of effects. It follows that the speed of investment and development is constrained by the capacity of the bureaucracies to select, implement and monitor programs. In order to facilitate efficient investment, we must strengthen the government agencies which guide investments. This is where integrated database management is involved.

What is management?

Management is of course a much wider concept than integrated database management. Management is the process of planning, organizing, leading and controlling the resources and activities of an organization in order to fulfill its objectives most cost effectively. It can be defined as the ability to get things done efficiently.

This is an essentially inward looking definition of management. Another vital aspect of management is the ability to understand the place of one's organization in the larger scheme of things, while at the same time remaining fully abreast of what is going on inside. Management requires imagination, technical competence and social skills. Good management shows transparency of procedure and performance standards.

What is integrated database management?

Integrated database management is simply the application of management on the arrangement and production of information. Integrated database management means the deliberate sourcing and organization of information with the objective of providing the information needed for decisions. Integrated database management can be used to improve efficiency and cost effectiveness.

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

Integrated management systems versus integrated database systems

It may be useful to devote a few words to the concept of integrated management systems versus integrated database systems. An integrated management system allows a manager to have instant access to information concerning his business and to make decisions with a minimal time lag. It concerns how the organization is structured in order to equip management to make timely decisions. For example, in manufacturing this means direct access to production and marketing figures, which enables the manager to adapt production and or sales in accordance with the technical options.

An integrated database system is one part of an integrated management system that deals with the data involved in decision making. The final product of an integrated database system is information used by management. The final product of an integrated management system is a decision made by management using the information provided.

Integrated management systems are used in the public sector. An example would be commodity procurement and sales systems. In general, however, the public sector concentrates more on collecting and processing data for the public good. In this case an integrated database system is more important. One can imagine, for example that crop production data would be collected then made available to the public. The data would also be stored to build up a time series of historical data.

The best example of an integrated management system is in fact the total sum of state activities enabling itself and its citizens to make decisions on activities and allocation of resources. Similarly, one could apply the term integrated database system to the process of aggregation of farm, village, sub-district, district, province and national production figures of selected commodities. Putting the figures together one serves a vast number of users at various levels of the state hierarchy and in various research and development agencies. The following section focuses on the key aspects of research management with reference to database systems.

What is research management?

Management applies to all organizations, and also to research. Regarding agriculture, one often encounters the words “management” and “research” in conjunction. The meaning is not always clear; sometimes one means simple personnel administration, or budget expenditure control, in other cases it means relating expenditures to substantive issues, organizational objectives and targets.

Research management is concerned with decisions about the allocation of scarce resources for multiple ends. Cost and time saving strategies are very important for research and programming agencies, both on the factor as well as on the product side, i.e., on the data source and the report targeting sides. In order to increase efficiency, members of research sections need many skills. Aside from knowledge of basic theory, they should possess a broad knowledge about data availability and the time and costs involved with generation of new data. They should be capable of making the distinction between primary and secondary data, producing time series for production and consumption patterns, applying data matching techniques to assess the validity and reliability of data, assessing the credibility of data sources, and making a judgment on the overall and the specific relevance of these data in relation to the research problem. The general adagium which should guide this exercise is “save time first, save costs later”.

It may be pointed out that an improved flow of data and information helps public and private sector organizations to make decisions on the allocation of scarce resources. This is where information management is important. Before we begin to discuss what information

management is all about, it is useful to pay some attention to the nature and characteristics of information.

What is information?

When we speak of information as a product, we are simplifying somewhat. Information actually has many characteristics, which cause it to behave in complex ways. It behaves sometimes as a good, say a diskette, and sometimes as a service, for example the address of a person. Information can be duplicated; simple information at a very low cost, more complex information carries quite substantial price tags. The most peculiar characteristic of information is that it only assumes its value when someone uses it. This means that the ability of the user determines the value of the information.

An obvious management task, getting the right person for the job, is the main principle of getting things done properly. The first rule of management then is that it is all about people, and especially so in the management of information, where technological development goes very fast. Information management is not just automation; it requires thinking, interpretation and a clear mind.

Another peculiar characteristic of information is that it is always unique. Information reflects the countless situations people and organizations find themselves in when seeking solutions to problems. This is the reason why standardizing information is a near impossible task. Standardized information is problem specific; we usually refer to science as the broad method to create it. The scientific cycle: problem - hypothesis - fact finding - solution, is naturally not limited to scientists, everyone uses such a cycle in solving everyday problems. We refer to the ability to solve everyday problems as common sense.

The development and budgets of science show that one can only at a great cost standardize information to the extent that the same information has the same meaning for everyone. However, it is relatively easy to standardize the form (or format) of information into printed material, tables, pictures, diagrams, maps and so forth. The new electronic technology is making a big difference in this field. One should refer to it as a revolution in information technology, not only as a revolution in information itself.

Management described

Management of information is not different from management of a carpentry shop or an automobile factory. Information is simply a product like furniture or automobiles. Management deals with organizing tasks and people to facilitate efficient production. Of course, there are many differences of tasks and tools between a car factory and a rattan business, but the point is that deliberate organization is vital to the success of both. The carpenter will make sure where he left his hammer since he last used it, likewise an automobile factory will invest only once in the expensive molds to press the car bodies. Users of information will wish to know where the data, their raw material, and their tools are, so as to minimize time and costs. Deliberate organization is of equal importance in government work, which encompasses programmes, projects, planning and services. Deliberate organization is a central characteristic of bureaucracies and agencies so obviously management is important in government.

Box 1 Monitoring, socio-economics and standardization.

Project monitoring and measurement activities virtually always include, next to specialized parts, a socio-economic component. These socio-economic components relate activities, investments and programmes and projects to people. Used in this way, socio-economics basically means a general type assessment involving some specified market development and/or state activity, and people.

It is wise not to lose sight of the fact that socio-economic studies are necessary elements in assessment of all groups and categories in society: urban and rural people, male and female workers, farmers and non-farmers, and poorer and richer segments of the population. It goes without saying that the national statistics are the most powerful numerical tool for the measurement of indicators of a society through time. The point is, however, that socio-economic components are virtually always location, time and programme (activity) specific. It is therefore not surprising that every year in Asia many thousands of assessment and appraisal reports for public and private activities are completed. These many reports can not - and should not - be expected to be standardized. Yet, there is substantial pressure for the standardization of programmes, reporting and assessment methodology. In many cases public agencies, carrying a number of programmes like to standardize procedures, operational as well as decisional. Some part of socio-economic assessments are standardized. In agriculture assessment procedures, developed to zoom in on people, farmers and families, have been standardized.

It is wise to realize that there are always good reasons to improve the socio-economic component of any study, but that standardization is only a part of it. One can not standardize everything, and it may not be wise at all to standardize too much. The merit of standardization is that it allows ready comparison of studies; a problem is that it might inhibit innovation.

People

Normally we define people by name, gender, personality and looks. In organizations people are defined by the tasks they are asked to perform. The professional staff are defined by the specific skills they possess and the specialized tasks they perform. Managers usually do not require much specialized knowledge, but rather a familiarity with organizing: delegation of tasks, task definition and control. Although we have treated the professional and the manager as two different types of people, everyone combines to some extent managerial capabilities and professional capabilities.

Another characterization concerns the focus of the people in an organization; inward looking people and outward looking people. Although in practice the distinction is not really exclusive because some people may look outward while others may look inward, the two orientations are in fact the basic variable along which tasks are defined in the organization. For example, the marketers are typically looking to the outside whereas personnel administration and production have inward orientation. Problems arise when orientations are overly concentrated in one direction neglecting the other.

In organizations under stress, one often encounters inward looking people. Staff members are aware of problems and may safeguard positions while an outward looking person may simply leave, resulting in an unbalanced situation. The most common situation in healthy organizations is that the senior people maintain an outward orientation while the younger staff perform the inward looking tasks.

Tasks

A task can be defined as a set of activities which belong together, and the performance of which is measurable and tangible. Tasks of professionals do not necessarily stay stable through time. Some items and responsibilities may be dropped, other responsibilities may be added. Information technology may change the way some jobs are done. One can encounter good examples of this last point in data-processing organizations. Aggregation of local data into larger blocks was sometimes done by hand, which required substantial manpower. The introduction of computerized data-processing technology has had important consequences for

task definition and the definition of requirements for positions. As well, it has created different hierarchies.

The definitions of a task may differ between people in an organization. From a management point of view one should strive towards harmony between supervisor and professional in definition of tasks. When differences arise consultation is necessary.

A special problem occurs when managers and professionals receive large numbers of special requests for urgent assistance. Uncontrolled expansion of largely undefined tasks add to the routine tasks and may create tension, simply because of lack of consultation time, which prohibits calm definition of tasks.

Tasks differ from professions. Everyone who finished education in a specific field has found out that holding a diploma in say, agronomy, is different from being an agronomist.

Organization

In the previous paragraphs we have spoken about people and tasks, and it has already become clear that one can not really separate people and tasks from their organizations. We define an organization as the deliberate allocation of tasks to individuals, to achieve a goal, measurable as a defined product. This definition is very close to the classic definition of economics, "economics concerns the allocation of scarce resources to multiple ends". Efficiency and choices concern managers and professionals as much as economists.

We can distinguish between formal and informal organizations. Formal organizations have fixed tasks and objectives, and are legally registered. Examples are manufacturing industries and government organizations. Informal organizations do not have legal registration, nor do they have fixed objectives. Fraternities and social groups are examples. Within the category of the formal organizations we can distinguish short term organizations, "project organizations", and continuous organizations such as bureaucracies.

Project organizations have a set time scope and narrowly defined objectives. Bureaucracy refers to long existing, usually complex organizations. States have bureaucracies, headed by a government, but also long existing private companies (such as General Motors) have bureaucracies. Note that the word bureaucracy carries no negative meaning. A bureaucrat is someone, usually a professional or a manager, who performs tasks in a complex organization. Ideally speaking, it is an honor to serve the public in government, and the requirement of efficiency applies very much to bureaucracies and bureaucrats.

Flexibility in formal organizations

In government, we encounter long term organizations together with short and medium term organizations. Bureaucracies often use short term, temporary, organizations - commonly called "project organizations" - to perform specified tasks. Projects are usually attached to directorates, divisions or sections and seek to accomplish a task. Examples of medium term organizations are special programmes, say on introduction of a new technology in agriculture. Once such a task is completed one would expect such a programme organization to dissolve. This does of course not mean that staff are laid off, although this is not unthinkable.

Many professionals do not realize that substantial changes have taken place in the agricultural government bureaucracy. If one looks into the history of agricultural ministries in Asia, one encounters substantial change from the pre-World War II years to the present, both in terms of number of directorates as well as in terms of manpower. There is every reason to assume that changes will continue to take place, but that the years of broad expansion of manpower and services are coming to an end.

Let us return to the implications of the above for management's need to define the tasks. It is obvious that in fact a fairly continuous process of performance and need assessment takes place at high level in government bureaucracy. From an individual point of view one may perceive a great deal of continuity in organizational structure, yet from some distance with a longer time frame one perceives dynamic change in goal and task definition.

Management in the public sector, difficulties and opportunities

In seeking to explain a lack of overall management efficiency in Asia some critics have charged that Asian managers, whether in public or private settings, consider data under their control as their personal asset. A letter (Far Easter Economic Review, July 13, 1995) by an associate Professor in Business and Management of the University of Hong Kong says:
".... Confucianism and other cultural influences have created management systems that do not favour the use of formal and integrated management information systems. East Asian organizations make great use of personal and verbal communication and practice highly centralized decision-making. For Asian managers, information really is power and it becomes a personal asset rather than an organizational resource."

The notion that what is good for an individual is not necessarily good for an organization is useful, but it seems that this sort of behavior happens the world over. It may be more important to look at this issue in the larger context of management in the public sector. It is easy to imagine that an opportunist company manager negotiating contracts with various parties, may shift companies, or set up their own business taking clients and expertise with him. Such a thing can not really happen in the public sector, because information, which is the major asset of state organization, always stays where it is, in the databanks, archives and the files of the organization. Of course, one can make commercial use of such information, but one can not easily imagine that the information can be taken away.

A better explanation for lack of efficiency may be the small use of computers made by private and public managers. Integrated management systems are not yet popular in many parts of Asia. By comparing Asia with the United States and Western Europe, we find the major reasons for this are low public budgets and in the countryside a low level of services in Asian countries.

In general, the difficulties in management of public agencies engaged in data collection and registration are somewhat different from those in the private sector. They usually concern low budgets and related delays in adoption of new technology, difficulties in "proving" that financial investment is useful and has a payback, and finally keeping the staff interested and pro-active.

One reason it is difficult to provide evidence of the financial benefits of modernization of data collection and processing is that government agencies always fit into a larger scheme. It is for example very difficult to distinguish between agricultural extension services and agricultural research in calculating their respective impacts. Any impact would also have to account for the influence of the private sector on the performance of agriculture. If one looks inside the state organizations it is equally difficult. To take just one example, how can one calculate the benefits of remeasurement of triangularization and improved accuracy of maps? To be sure, one would not wish to question the value of high quality maps; however, to calculate the utility of accurate maps one would have to make estimations of the time and accidents saved by users consulting improved maps as compared to the same when using old maps. Not an impossible task, but by no means simple. It is thus clear that the quest for

improvement of performance of public services in collection and processing of information has its own peculiar difficulties.

However, there are also very specific opportunities to improve efficiency and scope of work in public data collection and processing services. These opportunities relate directly to the continuity of indicators through time, and to - what is generally considered to be the weakness of the state apparatus - the complexity of state services, and the undeniable dependence of the state services on one another. Before we discuss these in detail it is important to discuss one very popular way organizations seek to increase efficiency.

Reorganization in complex organizations, some do's and don'ts

With shrinking budgets and increased pressure to 'do more with less', reorganization is a constant theme in organizations around the world. There are many points of view on reorganization; the personnel point of view, the task and goal point of view, the efficiency point of view, and so on. It is very important to understand that reorganization, though seemingly rational, is in practice not always a rational process, or, sometimes not even a rational decision. One can distinguish a number of reasons for reorganizations:

- a quantitative change in objective and task range of the organization. In Asian agriculture, many coordinating agencies have gone through substantial expansions since the 1960s. The present trend may go towards contraction.
- a qualitative change in objective and tasks of the organization. Examples in agriculture in Asia are the changes from the commodity centred programmes of the 1960s through the 1980s, towards resource defined zonal programmes of the early 1990s.
- the creation of new agencies, which are to address objectives and tasks which earlier were only an aspect of tasks. Examples in Asia (but also the west) in agriculture concern the creation of resource and pollution research agencies, or the creation of gender focused agencies.
- the dissolution of agencies. This variety of organizational change is relatively rare. Usually these are agencies which found initial funding by outside donors, and which were to be continued by the government. With scarce budget allocations these agencies are the first to suffer. These sleeper agencies may contain highly useful information, which can usefully be incorporated in better endowed agencies.
- the administrative reorganization. This type of reorganization happens when an agency is growing too big or too small for its place in the ministerial hierarchy, and when by regulation or consensus steps need to be taken for adjustment. One can speak of "cutting-up" or "lumping" organizations.

Reorganization is very difficult to complete elegantly with care for both tasks and people. It requires clear insights into the substantive issues, good personnel management and steady relations with the management of the agency as a whole. All too often internal stress in an agency leads too diminished capacity, diminished quality and quantity of output, and a reduction in customer relationships. It may be true that many reorganizations are the result of earlier attempts to deal with difficulties in management, lack of concern for personnel and other substantive matters. It therefore happens that agencies get caught in a cycle of reorganizations. It may be true that many reorganizations can be avoided with clearer task description and terms of reference for staff.

In a bureaucracy under pressure - development bureaucracies are under heavy pressure to achieve targets - one encounters a relatively high proportion of on-going reorganization. There are good and understandable reasons for attempts to create ever more creative agencies

with the aim of bringing about development. Yet, it is wise to realize that once organizations begin structural change that trivial leadership mistakes can cause total failure of the attempt, which then, in turn, can call forth new attempts to restructure other or related parts of the bureaucracy. A "chain" of reorganizations rarely delivers the desired result. The keys to management, we repeat, are people, tasks and organizations. One can not solve problems concerning people by making an organizational change only. If organizational change is necessary, the wise course of action is to make everyone understand both the necessity for and the direction of the change, clarify professional consequences, and listen carefully to suggestions from the task managers on how to actually go about task change.

The following section provides a very brief review of the role of information in the modern state in Asia, and focuses on data collection and measurement issues in agriculture.

The role of information in state activities: regulation, registration, and participation

There are many lines one can follow when thinking about agricultural services and planning for agriculture. One can start from scratch, with the concept of the state, and its necessary functions in monitoring, market participation and the provision of services. One can then construct a blueprint of the agencies and departments that will make up the state and proceed from there. We will take a different approach.

We will start with the tasks of the state agencies as a given, and approach the overall picture by making lists of state organizations, departments, agencies and so on, to build up a directory of tasks and performances from the bottom up. We will take this route because it starts with the current situation and therefore offers the chance to take immediate action.

The state and its apparatus performs three types of activities:

(i) it regulates by law and rule,
(ii) it registers, and,
(iii) it participates in the economy.

We will concentrate here on registration and participation, and we will leave law/rule outside our scope. In virtually any country of the world one encounters government organizations which engage in registration of assets and participation in agriculture for one or another purpose. Examples of such organizations are:

1. department of land registration
2. department of people registration
3. department of land taxation
4. line departments measuring use of land, usually split up in forestry, food crops, fisheries, livestock, industrial crops, etc.
5. line departments of public works measuring land, water and roads
6. line departments covering natural resources measuring extraction of minerals
7. departments and institutes engaged in measurement of land / weather / climate / soils, etc.
8. central statistical agency, performing independent measurement and compiling data for redistribution
9. marketing boards/agencies participating in trade.

All the above agencies perform measuring and registration of resource use which is directly connected with taxation and with measurement of performance. The structure of states into hierarchical levels means that the information generated must be aggregated at various

levels. In general the level of aggregation and abstraction increases the higher one steps on the hierarchical ladder. This is best reflected in the spatial organization of the state. In any country one encounters people, households, villages, sub-districts, districts, divisions, regions or provinces, and finally a national level in government. The data needs of the national government are far more abstract that the more specific data needs of a household

Each of the above agencies organizes data collection and processing in its specific way. The actual choice of indicator and technique of data collection depends, of course, on the tasks of the agency concerned. There is wide variation: weather stations measure precipitation, sunlight and temperature daily, or even hourly, whereas estimations on the area planted to certain crops are made on a monthly or seasonal basis, to be finally reflected in annual reports as a yearly figure. The land taxation services usually measure land only when a transaction is completed or a split in a plot is made; they use registered land as a norm for annual taxation. Cartographic and soil measurement is costly and therefore rather rare. It is self evident that measurement techniques, costs and time intervals differ.

It is most important to observe that there is an astonishing continuity in measurement of the same indicators through time. Before going deeper into the costs and time-intervals of measurement, a brief diversion to the continuity of measurement of the same indicator through time is important, and, as will be seen, invariably linked to the one immovable asset of any economy, land.

Continuity of major indicators in registration by the state

Observations on the continuity of measurement of the major economic indicators start with the choice of indicators when institutionalized measurement started in Asia. The major indicators were land, houses, animals and people. These made up the major assets in a given economy under a given rule. In earlier years economies were largely agricultural and it stands to reason that indicators used for taxation covered agriculture. The foremost among the indicators used in Asia was cultivated land. This indicator continues to be used today. Likewise the number of livestock is still in use as an indicator for taxation. This continuity of measurement occurs because food is a basic need and its method of production has not changed greatly in thousands of years.

We can conclude that the continuity in the type of indicator used by the state offers opportunities to apply improvement in data processing. Before turning to this issue we have to take a look at the actual measurement techniques.

Progress in technology, engineering and measurement

Whereas the major economic indicators show continuity through time, the history of measurement shows a rather more dynamic development. Measurement itself is directly connected with science and technology. The advances in measurement are directly related to progress in engineering, and both private and public investment. The significance of improved measurement goes beyond improved engineering and investment, because it relates to consumer utility and lower transaction costs inherent to any economy. This issue is very important and not very well researched, probably because of the many complexities involved. Here we will stay with measurement cum registration for the time being.

There are many examples of improved measurement. Measuring the exact location on the globe is now very easy with modern radio satellite supported technology. In earlier years measurement required extensive knowledge of algebra and mathematics and in fact made up the stock-in-trade of captains and navigators, etc. Measurement of weather, waterflows and soil

properties has improved tremendously over the years; in general one can say that physical measurement of immovables has vastly improved. This is reflected in the on-going public investment in mapping, resource inventories and actual engineering projects generating utilities. All these investments however, do not necessarily result in a changed relationship between people and the state. Measurement of the number of people has not changed. The only way is to count them, and write down the number. Improved measurement and engineering does not necessarily result in a more efficient or better administration, but it seems to be a basis for improved public services.

The institutional organization of data collection

It is logical to start with the observation that the type of data determines the way they are collected and processed, and therefore also the structure of the agency concerned. We have demonstrated above that the state organizes its registration of information in various departments and agencies, and that the major indicators remain the same through time. It would seem to follow that the organization of registration and data collection remains constant through time.

To verify this hypothesis, one would have to conduct longitudinal investigation (through time) with the hypothesis that data covering a specific group of indicators are collected by one and the same agency through time. This is not always the case, for instance departments registering land and people are politically and financially very important and these may shift ministry in times of revolution and uncertainty to remain close to central authority. Yet, the actual way of measurement and registration rarely changes. The consequence is that the actual organizational features of data collection and processing tend to remain the same. This is now changing throughout the world, in some countries faster than others.

There are, of course, very good reasons for the concentration of specific tasks in specific departments and agencies; the collection and procession of specific information needs skills, standards of performance and management. If one would scan old yearbooks and organizational charts of, for example, climatic/cartographic/ agricultural/soil research agencies one would surely find that the actual internal division of tasks centered around the technology of data collection and processing, and changed with the introduction of new measuring technologies.

There are two major technological changes which influence the actual organization of data collection: the wider use of the telephone and the introduction of computerized data processing. The telephone increases the speed of information transfer among areas, whereas the computer facilitates more comprehensive analysis and improved storage of information.

For agencies involved in market participation, quick and timely access to information is essential. The state agencies occupy a unique position in terms of access to information from the various producer and consumer areas; they differ from private companies in the spatial and commodity coverage of accessible information. Market information is essentially different from annual bookkeeping because of the time requirements of the information involved. For market-oriented organizations, the telephone and local access to information is the major requirement.

In the keeping of annual statistics data usually flow from the producer areas to higher administrative levels, and data usually become more aggregated in the course of their travel from locality to the national level. Finally, the thousands of local estimations over the months of the year are brought together in one figure: national annual production. There is a great deal of effort required to arrive at an aggregate estimation, and usually users of aggregates have no direct opportunity to disaggregate the data. It is of course totally obvious that the use of

personal computers can solve the problem of simultaneously holding disaggregated data sets and performing analysis.

In the cartographic aspects of registration, the introduction of digitization has led in many cases to the establishment of new sub-departments. The introduction of new data processing technology through computers is still leading to new skill and job requirements. Yet, if one would scrutinize the main function of the departments/agencies concerned, one would most likely encounter a remarkable stability in task and goals. Maybe one can summarize that the stability hangs together with the necessary internal functions of the state agencies, whereas internal change within these agencies reflects technological progress and task definition. This paradoxical situation requires rather specific management approaches and skills.

Data availability and the opportunities for improved monitoring, planning and operations research by public agencies

Every state plans, predicts and monitors. These activities require large amounts of data. Through computerization and the Internet, data availability is less of a problem today. The problem is organizing the data into useable information.

An example of predicting in agriculture is forecasting for commodities, such as rice. Short and medium term production targets are set. The state performs these jobs through specialized planning agencies. These agencies are fed with information from various line-departments and sub-departmental statistical units, as well as statistical bureaus. Usually plans, whether agricultural, industrial or financial, are made on an annual basis. The reason why stems from the weather, which completes the seasonal cycle in one full year. One can use annual performance in agriculture as the basis of next year's performance; in actual fact this very simple short cut method is the most popular forecasting approach in use. In this way one can establish trends, and extrapolate these, with or without making assumptions. However, every agriculturist, climatologist, agricultural trader, and hopefully also economist, knows that for successful forecasts and business operation one needs higher temporal data density. There is much information in daily or weekly data that is not included in yearly amalgamations. The commencement of rainfall indicates planting time and determines likely harvesting time. For a trader this means that the daily weather reports indicate where and when operations are possible; it means for agricultural staff of local offices that monitoring of area planted needs to start, and so on. However, with these admittedly somewhat trivial observations, we have only scratched the surface of the wealth of information available.

Let us take a closer look at the time dimension, agriculture and available data. The basis of plant growth is photosynthesis, basically a day-night, light induced cycle of plant activity. Weather stations provide data on a daily basis on temperature, precipitation, radiation and wind speed. Plant scientists are not likely to be satisfied with daily data and may need to establish hourly data to test hypotheses; in such a case they are likely to generate their own data. Farmers usually monitor their crops on a daily basis, while agricultural district staff commonly monitor area planted and harvested on a monthly basis, going by reports of farmers or local functionaries. Prices are of paramount importance, yet difficult to approximate satisfactorily. In off-seasons sales prices simply do not exist, however, distribution prices may be available. Monitoring of collection prices is common for major commodities, but for minor crops may not be available. Usually local functionaries make some type of weekly construct price; daily recording of prices is somewhat rare in Asia.

One could go on like this, but it will be clear that within the public services engaged in the collection of data a vast variety and volume of data are handled, and the data has different time intervals. The weather is measured daily, whereas soil properties may be measured only once in several decades.

Likewise, if we look at the spatial dimension we encounter substantial variation. Farmers obviously look at their own businesses, and those of their neighbours; local functionaries report on larger spatial entities, hamlets or villages, whereas staff of line departments usually report on small administrative entities, sub-districts, districts and larger spatial entities, divisions, provinces, watersheds and so on. It should be well appreciated that if for a given spatial entity information is available, this does not imply that the information is indeed valid for the whole area. Indeed the problem of validity of local weather data, for example, is quite difficult to solve. It is well known that in mountainous areas weather variations within valleys can be quite substantial. Likewise, reports on collection prices of a commodity registered in a local centre do usually not reflect the actual collection price as received by farmers living in localities farther away from the centre and road. Inferences on data validity, however, are possible. Usually one spots the data density by plotting the data on a local map.

Spatial and temporal data frameworks

The rapidly expanding availability of statistical information on agriculture, demography and infrastructure in virtually all countries in Asia offers very basic possibilities for inter-disciplinary research in general and agricultural socio-economic studies in particular. In addition, as resource and land information also expands rapidly, this information is available at meteorological, soil, geological and cadastral agencies.

There is still large variation in data coverage within countries and among countries in Asia. Thinner populated areas and upland areas are usually the last parts to be mapped. Nevertheless, data availability has rapidly increased and it is now possible to make overlays of areas using various data, each providing a theme.

In practice it is possible to transfer the vast mass of data available to computers of sections and divisions, etc. However, it is still a good idea to make a start at compiling the basic data. In agriculture these concern area allocation to land, input output data for major crops and prices of farm inputs and outputs. It is very useful to collect the data on a disaggregated basis to achieve some degree of homogeneity in overlays. This may not always be possible in areas where farm and seasonal diversity are very high, such as in young volcanic areas and horticultural production centres.

There are many advantages in setting up a spatially disaggregated time series on chosen key indicators. For research the major advancement of such an approach is that it offers a framework for assessing spatial and temporal validity. In data assessment, validity and reliability is most important. Usually validity (do the data reflect, what they mean to reflect) and reliability (are the data measured properly) receive extensive attention in measurement for hypothesis testing. What rarely receive attention are the basic dimensions of time and place in assessment of validity of indicators. This is an essential step on which generalization in place and time is based.

Spatial and temporal data sets also offer the option of searching for reflection of project impacts in given areas over time. For example, overlaying data on income and infrastructure facilitates powerful analysis, both broad and deep in scope. This use of data on agriculture, people and resources is in its beginning steps in Asia. In using the data one has to be careful to

check the data, through central, provincial, state and district offices. Matching of statistics with reports on primary data will be quite important to gain insight regarding the reliability of data.

Census data if periodically repeated, offer very good potential for analysis through time and space. Census data are usually highly under-utilized; they are usually only analyzed at the aggregate level. Census data offer, however, excellent possibilities for spatial analysis, one simply needs to use the raw data, structured by the administrative (or spatial) frame. With the increasing attention for local and regional issues and development, census data offer very cost effective ways of periodical monitoring.

The construction of a database matrix

Every agency (even if its staff is not aware of it) has a database matrix. The simplest way to set this up is to make a list of all data handled, a data directory. One can go one step further and that step concerns setting up a database matrix. Figure 1 shows a way of spotting data along both time and spatial dimensions.

Figure 1 Spatial and temporal dimensions of data.

Spatial entity	Farm	Village	Sub-district	District	Zone/province	State/division	Country
Time interval							
hourly							
daily							
weekly							
monthly							
seasonal							
yearly							
multi-yearly							

This figure gives a schematic view of data, defined by time period and spatial entity. On the time continuum it shows a progression from observations covering short to longer periods, and on the spatial a similar progression is set up. This scheme facilitates a check on the basic time and space validity characteristics of data, and serves also to identify gaps in data sets and bases. It shall be clear that every cell in the above matrix can contain a wide range of information, which does not necessarily have validity for the spatial area and the time interval in the matrix. The matrix is just a very simple tool to structure available data; the validity and generalization issues require specific solutions, short cuts, assumptions, and indeed, investigations.

Any manager dealing with data collection and processing is well advised to set up such a matrix, simply charting availability of data. If one strives for completeness in agriculture, the task is large but by no means impossible. A good look at the various types of available information then easily discloses the many options one has for creative processing, and the many options one has to tackle research and development problems. Figure 2 shows a typical set up of a database matrix represented by rows (spatial observation units) and columns (attributes).

Figure 2 Setting up a database matrix.

Area harvested, production and yield of rice by district January -April, 1997.

Fields: Indicators/ Attributes

	Area Harvested (ha)	Production (mt)	Yield (kg/ha)
Records: Cases			
District 1	232,756	899,754	3,866
District 2	352,134	1,295,909	3,680
District 3	140,988	634,382	4,563

In our view it is the vast availability of information from state data collection and processing agencies which offers many opportunities of added value of information and improved data management and processing. One can think of examples, many of which are well known. For instance, a soil research and mapping institute by itself can only become effective if the information on soils can be usefully combined with other information such as agricultural production and weather, to generate insights which may lead to more cost efficient public and private investment. Likewise, it is clear that agricultural research planners need to have a high density of climatological and soil data to identify probabilities of crop success. These examples are obvious and well known. There are many more possibilities if one expands one's view and includes more indicators along the two dimensions of the globe, time and space. Getting data density to required levels takes time and money; nevertheless the problems to be addressed by the state are always pressing. People are always looking for solutions, not necessarily for long term investments.

There is a danger in compiling endless data sets covering many indicators. The human brain has limits to its capacity, and a wise manager never demands performance beyond capability. There are, however, ways to bring about some clarity. The best thing to do is to recall theory from the various disciplines available, and to start from defined problems and operationalised goals. Such an exercise, however, should not be confused with task definition; it is part of it. It is regrettable but true that at the academic level many disciplines have more or less stagnated around the object of study and a few key concepts. Demography is the numeric science of people; however, anthropologists claim they are also studying people. Economics is the science studying households, national and/or otherwise. A lot of energy seems to go into drawing boundaries between disciplines, referring to the unique products - in terms of information - of the various disciplines.

It is when problems are being solved that true inter-disciplinary cooperation occurs. Agriculture, as a field, actually shows the truth of the foregoing statement. It encompasses a truly vast range of basic sciences and concepts. At any rate, one must be careful not to fall victim to fashionable inter-disciplinarities, and to keep the practical goal of both supplying good quality information to users and to perform sound research, with the aim of improving performance.

Integrated database management and operations research

The word operation in the context of agriculture can mean many things; operation means simply planned activity. This means that farmers, traders and also government agencies perform operations. Usually these parties have to sequence their activities in relation to one another. They are mutually dependent.

The concept of operations research is usually reserved for bigger organizations where units (or sections and divisions) work together to achieve an explicitly set quantitative goal (or set of sub goals). Operations research refers then to systematic investigation in the allocation of resources, usually with the aim of minimizing costs or maximizing impacts of measures. It can be thought of as using scientific methods to research how to manage your resources more efficiently.

It is obvious that one can view the services provided by an agricultural ministry with operations research in mind, and that these can also benefit from operations research. This thought is by no means new. When operations research gained in popularity after the Second World War, it also gained a foothold in agriculture, especially in crop market agencies. Yet, the concept of operations research has never become really popular in agriculture. There may be some reasons for this, such as the complexity and the somewhat overdone attention for mathematics in the research process. Another reason may be that the traditional structure in agricultural departments shows some resistance against sweeping and rather complex approaches. An associated reason may also be that disciplinary orientation follows segmentation within agricultural services (for example the tendency that livestock related agencies are manned by veterinarians/livestock specialists, and so on).

All the above may account in variable degree for the lack of popularity of operations research in agriculture. In our view though, a major reason is the slow adaption of new technology in data collection, transfer and analysis. If state agencies are slow to adapt to technological change, they will also be slow to adapt a concept like operational research.

The above should not be misunderstood as a cheap argument against institutional inertia or old fashioned competence in disciplinary fields, but rather as an explanation why perfectly useful concepts such as operations research have not received more attention in the context of agriculture. The possibilities of integrated database management are likely to offer a new stimulus to older but still highly relevant concepts such as operations research.

Examples of operations research in (agriculture) are: production forecasting and its derived uses for forecasting and operational planning of state participation in markets; purchasing, storing, pricing, and state participation in grain trade; firm analysis and budgeting and also in performing budget allocation analysis for departments and agencies.

One can also think of mixtures between operations research and scenario formulation where one uses techniques from operations research as a tool to calculate approximations: for example land use assessments using crop models, or analysis of trade regimes in agricultural products with the same, or to relate trade regime analysis and prices in assessment of land use. These combinations of operations research tools in static form are very useful in explorations.

Conclusion

We have seen that management involves using the resources of an organization efficiently. The key resource is people and managers use their skills to get people to complete the tasks of an organization in the most effective way. For many managers this is an inward looking process as they concentrate mainly on what happens in their organization. But managers must also be outward looking to be able to find and adopt new technologies to improve efficiencies in their organization. For various reasons managers in Asian governments have been slow to adopt new technologies. Integrated database management is one technology which can prove especially useful to agricultural planning and research. There are large amounts of data available from government organizations and elsewhere that can be gathered and analyzed

with available technology to provide profound new insights and improve the efficiency of our organizations in times of decreased funding.

Introduction to Relational Database Management Systems

Gary Timoshenko[*]

Every organization collects data and maintains records on various facets of its organization. How effectively it uses these data helps determine how productive the organization is. Almost every agricultural and research organization is facing cutbacks. This makes it even more important for these organizations to make efficient use of their data. Improving an organization's data management is a cost-effective way to increase its information output and the quality of decision-making

An organization's data are organized and stored in collections of related data called databases. A database need not be stored on a computer. Any logically coherent collection of data may be considered a database. This can mean a collection of papers in a file folder, the names, telephone numbers and addresses you have stored in an address book, or the data on agricultural production you have stored on your computer. In order to support informed and effective decisions, these data must be accurate, reliable and easily accessible. A computerized database management system makes this possible.

Why use a database?

There are many advantages to using a computerized database system. For the individual user advantages include:

- Compactness - no need for a lot of paper files or for repetition of data.
- Speed - the machine can retrieve and change data much faster than a human can. In particular, ad hoc, spur of the moment inquiries can be answered quickly without the need for time-consuming manual or visual searches.
- Less drudgery - much of the tedium of maintaining files by hand is eliminated.
- Currency- accurate up-to-date information is available on demand at any time.

In a multi-user environment, these benefits apply even more, as the database is likely to be much larger and more complex. However, the one overriding additional benefit in the multi-user case is that a database system provides the organization with centralized control of its data. Specific benefits accrue from this centralized control:

- Redundancy can be reduced. In a non-database system each application has its own private files. This can lead to redundancy. For example at a large company the person who keeps track of incoming employees will have their names and addresses on file; so will the person who looks after pay cheques, as will the person who looks after health insurance. A shared database system will eliminate these redundancies.
- Inconsistency can be avoided. This is related to the previous point. If one piece of data is represented in two or more places, there will be occasions when the data do not

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

agree. If an employee has moved, one set of files may contain the employee's new address while one contains the old address. It is often difficult to tell which piece of data is the correct one. In a database the information is stored only once so inconsistencies are avoided.

- Data can be shared. If the database is connected to a network, data can easily be shared across a network and more than one person can access the data at one time. Also new applications using the existing data can be developed without the need for any additional data.
- Standards can be enforced. The database can be structured so that each piece of data is represented in a standard form. For example all dates are written the same way or names are entered with first name in one column and last name in another column. This makes exchanging data between systems much easier.
- Security restrictions can be applied. Having jurisdiction over the complete database can ensure that only designated people have access to all or part of the database and only designated people can update or change the database. However, if these restrictions are not in place, security of the data may actually be at greater risk.
- Integrity can be maintained. Integrity means ensuring the data in the database are accurate. Inconsistency between two pieces of information that purport to represent the same fact is an example of this. Even without redundancy the database may still contain incorrect information. An employee may be shown as working 400 hours in one week, or belonging to a non-existent department. Centralized control can help avoid such problems because the integrity of data only needs to be checked in one file. Conflicting requirements can be balanced. The database system can be structured to provide overall service that best fits the companies needs.

Most of these advantages are fairly obvious. One further point, which is not so obvious, is the point of data independence. Data independence means that we can access the data in different ways, we can present the data in different ways, we can do calculations on the data, and manipulate the data but these do not effect the original organization of the data.

Relational databases

There are different kinds of database software but relational databases are the most common and on personal computers, they are used almost exclusively. In a relational database system, the data are organized by subject. For each subject a table is created to store all the information on that subject. Later we define how the information in each of the tables is related so that it is easy to bring related data together.

A table is a collection of data about a particular subject. All data in the table describe the subject of the table. The table is set up in the usual fashion, with rows and columns. In this case the columns represent fields and the rows represent records.

Field 1	**Field 2**	**Field 3**	**Field 4**
Record 1			
Record 2			
Record 3			

For demonstration purposes we will create two sample databases. We will be using the example of a research centre in a ministry of agriculture. The first task for this ministry is to

keep track of personnel and payroll. The second task is to track rice production, rice prices and climate variables for every province in the country. To track personnel and payroll, we will first need a table with employee information. Let's look at the fields and records that might make up this table.

- **Fields** - Fields are the smallest unit in a database. Fields contain single pieces of information about something. In our example we want to create a table that contains all of our employee names, employee numbers, addresses, and other personal information. Individual pieces of data that we might want to include would be:

  Employee name
  Employee number
  Address
  Phone number
  Job title
  Hourly wage

  Each of these categories is a field.
- **Records** - Fields make up records. All of the information about one entity (our "something") is a record. Using our employee table example, all of the information about one employee will be one record. For example:

  Robin Jones
  098567
  1234 Camilla Way; St. Cloud, Florida, 12345
  904-555-1234
  Researcher
  $12.00

All of this information about Robin Jones is one record. In other databases, records might include books, inventory items, etc.

- **Tables** - The final part of the database structure is the table. All of the records make up the table. In our example, the records of all the employees at the company would comprise the table. In other words, all of the information about an entire group is the table.

Keys

We have seen that relational databases organize data into tables. But what do we do if we want to combine information from one table with information from another table. We need some way to relate the tables. This is done through keys and key fields.

In a relational database each table must have one field that provides a unique identification for each record in that table. To demonstrate we will add data to our table of employee information.

Table: Employee

Last Name	First Name	Employee Number	Address	Job Title	Wage
Evans	Mark	04-234	21 Elm St.	Researcher	10.00
van Druemel	Terry	07-456	45 Jalan Merdeka	Supervisor	12.00
Nagase	Yoshi	01-637	87 Jalan Sempur	Secretary	7.00
Cooper	Charlotte	04-734	29 Spagnum St.	Researcher	10.00
Evans	Michelle	03-346	21 Elm St.	Accountant	9.00

We have the fields: Last Name, First Name, Employee Number, Address, Job Title and Wage. If we look at the Last Name field we see that there are two people named Evans. Therefore the Last Name field is not unique for every record. Looking at the Address field we see two people with the same address of 21 Elm St. so Address cannot be used as the key field. The Job Title and Wage fields also have duplicate entries. Finally we see that First Name is unique for every employee, as is Employee Number. They can both be used as the key field but which one should we choose? We should remember that the key field must always be unique for every record. Suppose we chose First Name as the key field. What would happen if in the future we hired another employee named Mark. Once we added him to the database the first name field would not be unique for every record. The key field must always be unique no matter how large the table gets and how many records we add. Employee number will always be unique for every record in the table. No matter how many employees we add, we can ensure that no two employees will have the same employee number. Therefore, employee number is our key field. We will see why uniqueness is important when we use key fields to link tables.

Table: Climate

Date (dd/mm/yy)	**City**	**City Code**	**High Temp. (°C)**	**Low Temp. (°C)**	**Precipitation (mm)**
05/02/98	Jakarta	01	32	23	0
05/02/98	Yogyakarta	02	33	24	.15
05/02/98	Bandung	03	30	22	.08
06/02/98	Jakarta	01	33	25	.04
06/02/98	Yogyakarta	02	32	24	.11
06/02/98	Bandung	03	31	23	.07

Sometimes there is no field that has a unique entry for every row in a table. Take an example of daily weather for selected cities in Indonesia. We can see that all of the fields except for Precipitation have duplicate entries. Should we use Precipitation as a key field? No we shouldn't because it is possible (in fact probable) that as we add new records, we will get duplicate data in the Precipitation field. Since we have no field guaranteed to always be unique we will have to create one. This can be done in two ways. The easiest way is to simply number the records. Most relational database packages can create these numbered records automatically for use as a key field. The other more complicated way is to combine fields. We know we are getting one set of weather data for each city each day. We will never get two sets of data from a city on the same day or we know the data are not correct. Therefore if we combine the Date and the City Code, we will have a field that is unique for every record and we can use it as our key field. (Microsoft Access supports complex keys directly so we do not have to create a new column.)

Notice that in the above table we have a filed City code that provides a numeric code for each city. Using a numeric code for character fields limits typing mistakes and instances where two spellings may arise like Yogyajarkta and Jogyakarta. It also makes it easier to sort by region and combine fields. Having a field for city name and city ID in a table about climate is a redundancy. We should also create new tables dealing exclusively with data on cities. This will be demonstrated in the section on database design.

Table: Climate

Record Number	Date (dd/mm/yy)	City	City Code	High Temp. (°C)	Low Temp. (°C)	Precipitation (mm)
01-050298	05/02/98	Jakarta	01	32	23	0
02-050298	05/02/98	Yogyakarta	02	33	24	.15
03-050298	05/02/98	Bandung	03	30	22	.08
01-060298	06/02/98	Jakarta	01	33	25	.04
02-060298	06/02/98	Yogyakarta	02	32	24	0
03-060298	06/02/98	Bandung	03	31	23	0

Let's look at another table that tells us how many hours employees worked in a given week. The fields will be Record Number, Employee Number, Week, and Hours. The key field is record number since it is the only field guaranteed to be unique for every record. We could have combined Employee number and week to provide a key field but in the case we will just automatically number the records.

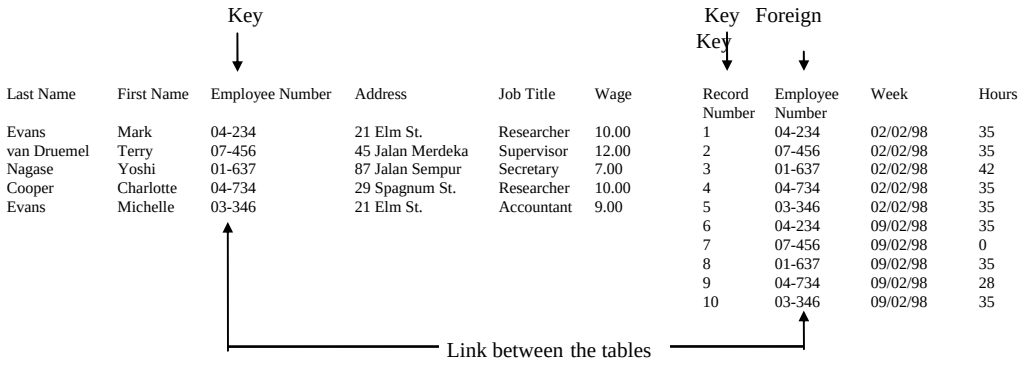
Table: Time Sheet

Record Number	Employee Number	Week	Hours
1	04-234	02/02/98	35
2	07-456	02/02/98	35
3	01-637	02/02/98	42
4	04-734	02/02/98	35
5	03-346	02/02/98	35
6	04-234	09/02/98	35
7	07-456	09/02/98	0
8	01-637	09/02/98	35
9	04-734	09/02/98	28
10	03-346	09/02/98	35

We are now ready to combine data from two different tables.

Queries

The real power of a database is the ability to see the data you want, in the order you want to see it. Queries allow us to ask questions of the data in our tables and to combine data from different tables.



Suppose we want to calculate our employees' pay for the week of April 20th to April 27th and mail them their check. We will need data from both the employee table and the payroll table. In order to link data from two tables we require some of the data from the first table to also be in the second table. This is done by having the key field from the first table also found in the second table. The combined data from the two tables are shown in the Payroll query:

Query: Payroll

Emp No.	Last Name	First Name	Address	Wage ($)	Week	Hours	Pay ($)
04-234	Evans	Mark	21 Elm St.	10.00	02/02/98	35	350.00
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	02/02/98	35	420.00
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	02/02/98	42	294.00
04-734	Cooper	Charlotte	29 Spagnum St.	10.00	02/02/98	35	350.00
03-346	Evans	Michelle	21 Elm St.	9.00	02/02/98	35	315.00
04-234	Evans	Mark	21 Elm St.	10.00	09/02/98	35	350.00
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	09/02/98	0	0
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	09/02/98	35	245.00
04-734	Cooper	Charlotte	29 Spagnum St.	10.00	09/02/98	28	280.00
03-346	Evans	Michelle	21 Elm St.	9.00	09/02/98	35	315.00

From the employee table we have use the fields: Name, Address and Wage. From the payroll table we have used the fields Week and Hours. The field Employee number is found in both tables. It is the key field in the employee table but it is not the key field in the payroll table. When a key field from one table is used as a field in another table we call it a foreign key. The key field and the foreign key field provide the link between the two tables and allow us to combine records. Pay is a calculated field that we create when we create the query (Pay=Wage*Hours).

Database design

Overview

Database design is a complex subject, no matter how easy some people think it is. This session only scratches the surface, but it is a good introduction. A properly designed database is a model of a business, or some "thing" in the real world. Like their physical model counterparts, data models enable you to get answers about the facts that make up the objects being modeled. It's the questions to be answered that determine which facts need to be stored in the data model. In the relational model, data are organized in tables that have the following characteristics: every record has the same number of facts; every field contains the same type of facts in each record; there is only one entry for each fact; no two records are exactly the same; the order of the records and fields is not important.

Why design?

Accurate design is crucial to the operation of a reliable and efficient information system. Computing technology is now so advanced that the impact of poor design may not show up as early as in the past; however, when the problems appear they can be severe.

The design of a database has to do with the way data are stored and how that data are related. The design process is performed after you determine exactly what information needs to be stored and how it is to be retrieved. The more carefully you design, the better the physical database meets your needs. In the process of designing a complete system, you must consider user needs from a variety of viewpoints.

Problems resulting from poor design

A myriad of problems can manifest themselves as a result of poor database design:

- The database and/or application may not function properly.
- Data may be unreliable or inaccurate.
- Performance may be degraded.
- Flexibility may be lost.

The following section explains some common problems resulting from poor database design. The problems can be grouped under two categories: redundant data and modification anomalies.

Redundant data

In an earlier section we used a query combine information in the Employee table with information from the Payroll table. Consider what would happen if we had the information combined in a table. The following table contains data about employees and how long they have worked each week. This seemingly harmless table contains many potential problems.

Table: Anomalies

Emp-No.	Last Name	First Name	Address	Wage ($)	Week	Hours
04-234	Evans	Mark	21 Elm St.	10.00	02/02/98	35
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	02/02/98	35
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	02/02/98	42
04-734	Cooper	Charlotte	29 Spagnum St.	10.00	02/02/98	35
03-346	Evans	Michelle	21 Elm St.	9.00	02/02/98	35
04-234	Evans	Mark	21 Elm St.	10.00	09/02/98	35
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	09/02/98	0
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	09/02/98	35
04-734	Cooper	Charlotte	29 Spagnum St.	10.00	09/02/98	28
03-346	Evans	Michelle	21 Elm St.	9.00	09/02/98	35

First, disk space is wasted by duplicating data about the employee. Every time a new week is entered for a particular employee, all of the employee data (name, address, wage) have to be repeated. Imagine the problems after several months of information are entered.

Modification anomaly

What if Charlotte Cooper moves to a new address? How many rows have to be changed in order to ensure that the new address is recorded?

Table: Anomalies

Emp-No.	Last Name	First Name	Address	Wage ($)	Week	Hours
04-234	Evans	Mark	21 Elm St.	10.00	02/02/98	35
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	02/02/98	35
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	02/02/98	42
04-734	*Cooper*	*Charlotte*	*29 Spagnum St.*	*10.00*	*02/02/98*	*35*
03-346	Evans	Michelle	21 Elm St.	9.00	02/02/98	35
04-234	Evans	Mark	21 Elm St.	10.00	09/02/98	38
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	09/02/98	36
09-002	Dharma	Sangit	32 Main St.	10.00	09/02/98	0
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	09/02/98	35
04-734	*Cooper*	*Charlotte*	*29 Spagnum St.*	*10.00*	*09/02/98*	*28*
03-346	Evans	Michelle	21 Elm St.	9.00	09/02/98	35

Again, imagine the issues surrounding modifications of hundreds of rows of data for one employee. When changes are made, they must be made to all copies of the data. Think about the confusion that results from changing only a subset of the duplicate data.

Deletion anomaly

Suppose that since Sangit Dharma is no longer working at our organization, we decide to delete his record from our database.

Table: Anomalies

Emp-No.	Last Name	First Name	Address	Wage ($)	Week	Hours
04-234	Evans	Mark	21 Elm St.	10.00	02/02/98	35
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	02/02/98	35
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	02/02/98	42
04-734	Cooper	Charlotte	29 Spagnum St.	10.00	02/02/98	35
03-346	Evans	Michelle	21 Elm St.	9.00	02/02/98	35
04-234	Evans	Mark	21 Elm St.	10.00	09/02/98	38
07-456	van Druemel	Terry	45 Jalan Merdeka	12.00	09/02/98	36
01-637	Nagase	Yoshi	87 Jalan Sempur	7.00	09/02/98	35
04-734	Cooper	Charlotte	29 Spagnum St.	10.00	09/02/98	28
03-346	Evans	Michelle	21 Elm St.	9.00	09/02/98	35

Now, looking at the remaining data, what is Sangit Dharma's address? We may still need to contact him. A deletion anomaly means that we lose more information than we want. We lose facts about more than one subject with one deletion.

Domain/key normal form

Relational theorists have classified database schema that have inconsistencies based on the anomalies to which they are susceptible. To keep a database free of these anomalies we must understand the terms: dependency, key, domain, and restriction.

Dependency

A dependency is a relationship that may exist between two fields. Given the value of one field, you are able to determine the value of another field. Let's use the table in the previous examples. Given the employee number, we are able to determine the employee's last name. This is a dependency: last names are dependent on employee's number. Given an employee number, are we able to determine the hours worked? No, for every employee number there can be different results for hours worked. It depends which week is specified. Therefore, hours worked is not a dependency of employee number.

To detect a dependency, ask this question:

In this table, does the value of one field determine all possible values of another field?

Employee Number
    determines Last Name?
        Yes
Last Name
    determines Salary?
        No

Key

Most tables should have a field or a combination of fields that uniquely identifies each row of data. A field is key if all other fields in a row are dependent on it.

At first glance, it may appear that the Employee Number in our example uniquely identifies a row of data. However, the key field must be unique for each row, so values cannot be repeated. Employee number 04-234 occurs in the first row, then 5 rows down it repeats again. Therefore, employee number 04-234 identifies how many hours Mark Evans worked in the week of February 2nd and how many hours he worked in the week of February 9th. The field EmployeeID is not the key. In this table, we have no fields which would qualify as a key field. We must create a key field by creating a record number field or by creating a complex key derived from EmployeeID and Date.

In good database design, every field in a table should be determined by the key field and there should be no other dependencies in the table.

Domain

A domain is the set of values a field can have. Every field has a domain, which has both physical and logical properties. The physical part of a domain is the type of information about that field. In our example, Last Name is defined as TEXT 25. Because of this definition, the physical description of the domain is the set of TEXT data with 25 or fewer characters. Similarly, the physical description for the domain of Hours is expressed as INTEGER. This results in data of nine or fewer numbers. The logical part of the domain is the set of information associated with that fact. First Names are not in the same domain as customer addresses, although they have the same physical property of TEXT 25.

Consider the value Mark. Is this value in the domain of First Name? To be in this domain it must have fewer than 25 characters and be a first name.

Restriction

A restriction is a limitation of some type on the values in a table. A dependency is a type of restriction. Stating that Last Name is dependent on Employee number is a restriction. Keys are a type of restriction. When a field is a key, it means that all other fields in that table are dependent on the key. Remember that a key can be a combination of fields. A domain is another type of restriction. When defining the physical and logical properties of a field, we restrict the data in that field.

Restriction is a general term. There are many other ways to restrict data in a table. Below are some examples:

- Week must be formatted as DD/MM/YY.
- Employee Number must be formatted as ##-####.
- Address must be TEXT with 40 or fewer characters.
- Wage must be CURRENCY with values between $0.00 and $9,999,999.99.

The normalizing process

Normalizing the database ensures that the structure of the database allows changes to be made without incurring unexpected consequences. The role of normalization is to maintain stable, reliable data through good database design. Tables, like paragraphs, should have a single theme. The table in the previous examples has two themes:

- Information about products

- Information about the suppliers of products.

The way to manage this information most efficiently is to split the table into two tables: a table of employees and a table of weekly hours worked. Now you can add employees even if they have not started work, change employee addresses without changing several rows, and not lose information if you delete a part. If you wish, you can always bring the original table back using a query to join the tables using Employee Number.

A method of database design

As you have seen, database design plays a major role in the stability and the reliability of your data. In this section, we show the process of designing a database. To help illustrate the design process, a database named Rice will be created for Indonesia. This database will track rice production, prices and weather for all the provinces in a country.

Although there are a number of rules that can be followed in designing a database structure, the design process is as much an art as it is a science. Follow these rules when at all possible, but not to the point where the database loses the functionality that is so important to the user.

Doing a paper design first has several advantages:

- Saves time, money, and problems
- Makes the system more reliable; avoids potential data-modification problems
- Serves as a blueprint for discussion
- Helps in estimating costs and size.

A good design should have the following objectives:

- Meet the users' needs
- Solve the problem
- Be free of modification anomalies
- Have a reliable and stable database, where the tables are as independent as possible
- Be easy to use.

Design of the database model

The design of the database structure requires the following steps:

1. List the objects.
2. List the facts about the objects.
3. Turn the objects and facts into tables and fields.
4. Determine the relationship among objects.
5. Determine the key fields.
6. Determine the linking fields.
7. Determine the constraints.
8. Evaluate the design model.
9. Implement the database.

Step 1: List the objects. Make a list of all objects. An object is a single theme, similar to a paragraph. For our database the objects are: province, climate, output and prices.

Step 2: List the facts about the objects. There is a great deal of information associated with every object. In this step, you should list the facts about an object and then eliminate the facts that are not important to the solution of the problem. A province, for example, can have many facts associated with it: name, square miles, population, gross domestic product, health

data, miles of road, number of telephones and much more. We need only the information we will use now and possibly in the future.

- Object: important facts about the object
- Province: province name, area, population
- Climate: province name, month, rainfall, sunlight, temperature
- Output: province name, month, area planted, area harvested, production
- Prices: province name, month, wholesale, retail.

Step 3: Turn the objects and facts into tables and fields. Objects automatically become tables, and facts become fields (columns) once the field domains are determined. Recall that a domain is a set of values that a field can have. Every field has a domain, which has both physical and logical properties. For example, the field for province name is defined as TEXT 25. TEXT 25 is the physical property of the field. Because of this definition, its domain is the set of all province names with 25 characters or less.

If a field is used to link two or more tables, the domains must be the same and the fields should be given the same name. If the logical description differs (for example, Province Name and ProvinceID), the fields are not the same and should not share the same name. The following is a list of the preliminary tables, fields, domains and frequencies for our Rice database:

Table: Province

Field	Domain	Frequency
Province Name	Text 25	one-time
Area	Numeric	one-time
Population	Numeric	one-time

Table: Climate

Field	Domain	Frequency
Province Name	Text 25	onetime
Month	Date	Monthly
Precipitation	Numeric	Monthly
Sunlight	Numeric	Monthly
Ave. Temp.	Numeric	Monthly

Table: Production

Field	Domain	Frequency
Province Name	Text 25	one-time
Month	Date	Monthly
Area Planted	Numeric	Monthly
Area Harvested	Numeric	Monthly
Production	Numeric	Monthly

Table: Prices

Field	Domain	Frequency
Province Name	Text 25	onetime
Month	Date	Monthly
Wholesale	Numeric	Monthly
Retail	Numeric	Monthly

Often it helps in the design stages to draw boxes to represent the tables. In later steps you can then fill in key fields and draw the relationships among the tables.

Step 4: Determine the relationship among objects. To determine the relationship among the objects, take each object and look at how that object may be related to another. Keep in mind that not every relationship existing between objects is important. The relationships that are important are those that allow you to model the database after the real-world situation that the database represents.

In agriculture time plays a very important role in determining relationships between tables. The data we use must have the same time dimension if we are to join the tables. For example, weather information is usually collected daily, price information monthly and crop production also monthly. In order to link this data it would all have to be of the same time frame. We would have to create monthly averages for the weather data. The spatial dimension is also important. National data cannot be combined with provincial data unless the provincial data is aggregated to the national level. For convenience sake we have assumed all data are of the same time and spatial dimension.

One-to-one relationships. For any given row in Table A, there is only one row in Table B. For any given row in Table B, there is only one row in Table A. In our data we have a one to one relationship between climate and crop production. We have a set of weather information for each province each month and for each province each month we have crop production statistics.

One-to-many relationships. For any given row in Table A, there are many rows in Table B. For any given row in Table B, there is only one row in Table A. The relationships between province and weather and between province and crop production are one-to-many, because one province will have many entries for monthly weather averages and crop production.

Many-to-many relationships. For any given row in Table A, there are many rows in Table B. For any given row in Table B, there are many rows in Table A. In our example we do not have any many-to-many relationships.

The first step in determining the type of relationship between tables is to list every table and to see how it relates to any others:

- Province relates to Climate and Output.
- Climate is related to Output.
- Output is related to Prices.

An effective method to find the type of relationship is to ask whether a specific record in Table A can point to (is linked to) one or to many rows in Table B, and then reverse the tables and ask the question again.

- Does a province record point to one or many climate records? Many
- Does a climate record link with one or many provinces? One
- The relationship between the tables is one-to-many.
- A monthly average of provincial climate statistics relates to one or many monthly provincial rice production reports?
- A monthly provincial rice production reports relates to one or many monthly average of provincial climate statistics?
- The relationship between the climate and the output tables is one-to-one.
- Not all tables have to be related.

Step 5: Determine the key fields. Keys were discussed extensively earlier. A key can be an account number, social security number, part number, license number, or any other numeric value or combination of characters that are unique. A complex key is one that is derived from more than one field. Microsoft Access supports complex keys directly so we do not have to create a new field (column).

The key field cannot have duplicate values. It is also useful if the key field is also a field in other tables, since this allows the linking of tables. If a province name is universally unique, it is used as a unique row identifier. However, if there is any possibility another province could have the same name, then it is not unique and must not be employed as a key field. Do not use any field as a key where the possibility of a duplicate exists. A key field cannot contain null values.

Text names as key fields may cause problems. Often text names are not unique. In the case of Province there may be alternative spellings or abbreviations. Jakarta and DKI Jakarta would be classified as two different provinces. Text fields also can not be used in numeric calculations. To alleviate these problems it is useful to make the key field a numeric value. In this case we will create a field Province ID to provide a numeric identifier to each province. Since this is a key field and a foreign key field in other tables we will add ProvinceID to other tables. If we want automatic numbering for ProvinceID, COUNTER data type in Microsoft Access is a good choice for a physical description for the domain of a key field.

We still need key fields for the three other tables. In the Climate Table no field is suitable for a key field. ProvinceID and Province Name will be repeated for every month of data and Month will be repeated for every province. There is also no guarantee that the variables (Precipitation, Sunlight and Ave. Temp.) will be unique for every row. This is the same situation for the Production and the Price tables. We could create a key field for each of these tables by numbering the records but it would be more useful to create a combined key using ProvinceID and Month. This combined key is a unique identifier for every row and it alows us to combine The Climate, Production and Prices tables because the combined key will be the same field in each of these tables. The database now looks like this:

Table: Province

Field	Domain	Frequency
ProvinceID	Numeric	one-time
Province Name	Text 25	one-time
Area	Numeric	one-time
Population	Numeric	one-time

Table: Climate

Field	Domain	Frequency
ProvIDMon	Combined	
ProvinceID	Numeric	one-time
Province Name	Text 25	one-time
Month	Date	Monthly
Precipitation	Numeric	Monthly
Sunlight	Numeric	Monthly
Ave. Temp.	Numeric	Monthly

Table: Production

Field	Domain	Frequency
ProvIDMon	Combined	
ProvinceID	Numeric	one-time
Province Name	Text 25	one-time
Month	Date	Monthly
Area Planted	Numeric	Monthly
Area Harvested	Numeric	Monthly
Production	Numeric	Monthly

Table: Prices

Field	Domain	Frequency
ProvIDMon	Combined	
ProvinceID	Numeric	one-time
Province Name	Text 25	one-time
Month	Date	Monthly
Wholesale	Numeric	Monthly
Retail	Numeric	Monthly

Each table in our database contains a key field. Each key is also indexed, and duplicates are not allowed.

Step 6: Determine the linking fields. If you have been careful about designating key fields, you also have determined the linking fields. Links provide a way to tie information (rows) in one table to another table. If a table has a key field, that field can generally serve as the link. However, the placement of the key is important, and where the link is placed depends on the type of relationship between the tables.

To determine the placement of the links, you must first know the type of relationship among the objects or tables. Once you know the type of relationship among tables, it is much easier to determine where to place the linking field to tie two tables together. Note that not all tables need to be linked relationally.

Linking in a one-to-one relationship: In one-to-one relationships the link should be the most stable field or should be from the table where the key field is created. The most stable is the field least likely to change. If an automatic numbering system is being used, then use that field as the linking field.

Table: Climate

Province Name	Month	Ave. Temp
West Java	02/98	28
Bali	02/98	29
Aceh	02/98	27
West Java	03/98	29

Table: Production

Province Name	Month	Hectares Planted
West Java	02/98	657
Bali	02/98	475
Aceh	02/98	398
West Java	03/98	456

One-to-one relationships

Linking in a one-to-many relationship: In one-to-many relationships the linking field should come from the one table. The key field from the employee table (the "one" side) should be placed in the dependent table (many side). When the key ProvinceID is placed in the Output Table, it is referred to as a foreign key in the dependent table.

Table: Province

Province Name
West Java
Bali
Aceh

Table: Climate

Province Name	Month	Ave. Temp.
West Java	02/98	28
Bali	02/98	29
East Java	02/98	27
West Java	03/98	29
Bali	03/98	30

One-to many relationships

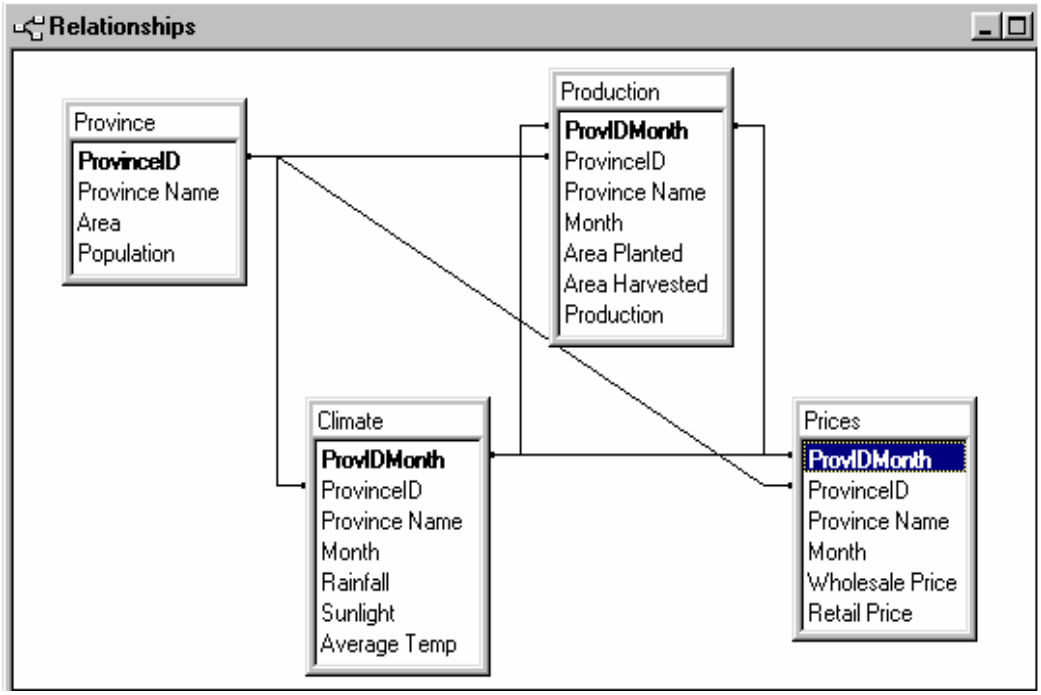
Linking in a many-to-many relationship: The many-to-many relationship causes problems when attempting to retrieve data and when relating a value in one table to its corresponding value in the other table. It is important to understand this relationship to be able to recognize and control this situation when it arises.

A classic many-to-many relationship is product and invoice. A product can be an item on many different invoices and an invoice can have many products associated with it. Linking causes problems because there will be redundant data, and performance may suffer. The solution to many-to-many relationships is to create an intersection table. This table should contain the key fields from both tables.

In our database we link the key field ProvinceID from the Province table to ProvinceID in the other three tables and we link the three key fields of ProvIDMonth in the Climate, Output and Prices tables.

Step 7: Determine the relationship constraints. Often the information we get from a database comes from more than one table, for example, if we want to compare rice production in a province to its population. Obviously we need to ensure every province entered in the Production Table is also found in the Province Table. To ensure the integrity of the data in our database, our model should require, for example, that no row can be added to the Production Table, unless there is already a corresponding row in the Province Table. This requirement is known as a relationship constraint. In this case, a constraint must exist on the Production Table that ensures that the province exists. Microsoft Access has certain referential integrity constraint mechanisms built into the engine to ensure these constraints.

In Microsoft Access, rules at the database or form level can be employed to enforce field domains (for example, to accept values less than 200, or text value must be F or M) or in any other operation where you want a data entry test to be performed.



Step 8: Evaluate the design. The next step in the design process is the evaluation of the design. In this step, you should look for any design flaws that could cause the data to be unreliable, unstable, or redundant. Every table should be evaluated by asking the following questions:

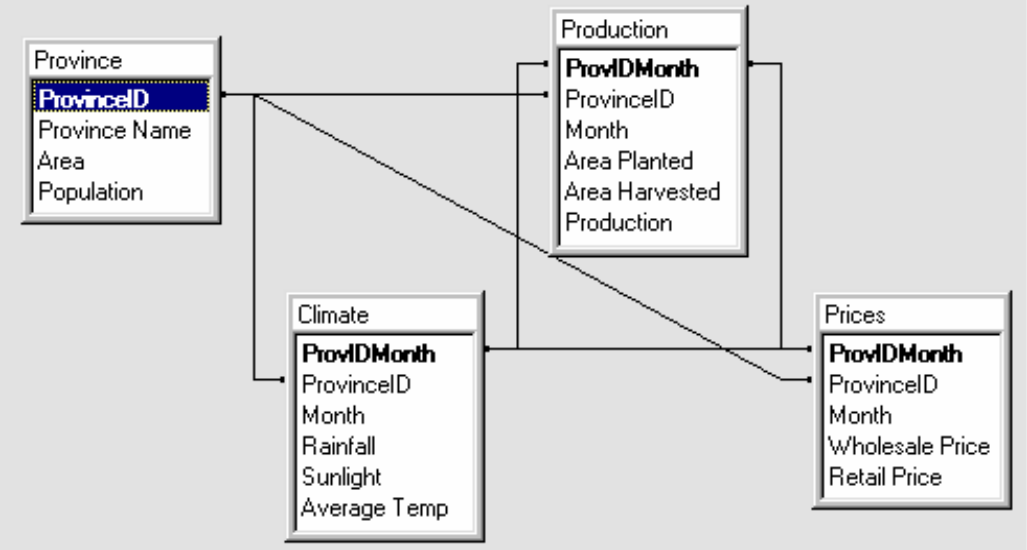
- Does each table have a single theme? It should. Each field should be a fact about the key.
- Does each table have a key filed(s)? It should.
- Are there any dependencies? Only logical consequences of the key should exist.
- Are the domains unique among tables? Do not mix domains unless the field is common between tables.
- Are the restrictions domain or key?
- Is the table easy to use?

In evaluation of the Production table, it is clear that the table has a single theme: rice Production. The table has a key: a combined key of province and month (ProvIDMonth). Given the combined key of province and month, all other fields can be determined. This means all fields are dependent on the key field, which is want we want. However there is another dependency in the table. Given ProvinceID we can determine Province Name, so Province Name is dependent on ProvinceID. Since this dependency does not involve the key field, Province Name is redundant. Since we have already defined which province ID goes with which province in the Province Table, Province Name is redundant in the Production Table and should be deleted. We can do the same in the Climate and Prices tables.

Step 9: Implement the design. Once the database had been designed on paper, the next step is to implement the design in Microsoft Access. When defining tables in Microsoft Access, it is extremely important to keep your paper design in mind. Designing a database on the fly can

cause problems that may be quite difficult to recover from. (Remember the anomalies earlier in this chapter.)

The following is a list of the final tables and fields for our Rice database, including linking fields:



Summary

Database design seems easy; however, there are many possible problems that result from careless design. By following the nine-step design process, the problems of data redundancy, changing multiple occurrences of data, and deletion and insertion anomalies can be avoided. It is well worth the time spent in the design process to ensure a reliable and flexible system. As you design more systems, many of the rules stated here become intuitive.

Design to the point where redundancy is eliminated or controlled. As you design your database, keep in mind the following list of common database errors to avoid:

- trash-table putting everything in the same table
- no unique row identifier (key field or columns)
- no linking or common fields
- mixing logical and physical descriptions of domains
- putting the linking field in the wrong table
- restrictions not enforced
- many-to-many relationships without intersecting tables.

Bibliography

Date, C.J. 1995. An Introduction to Database Systems. Vol. I, Vol, II., Addison-Wesley Publishing Co.

Kroenke, David M.; and Dolan, Kathleen A. 1988. Database Processing, Fundamentals, Design, Implementation, Third Edition. SRA.

Martin, Daniel. 1985. Advanced Database Techniques. MIT Press.

Access Relational Database Management System

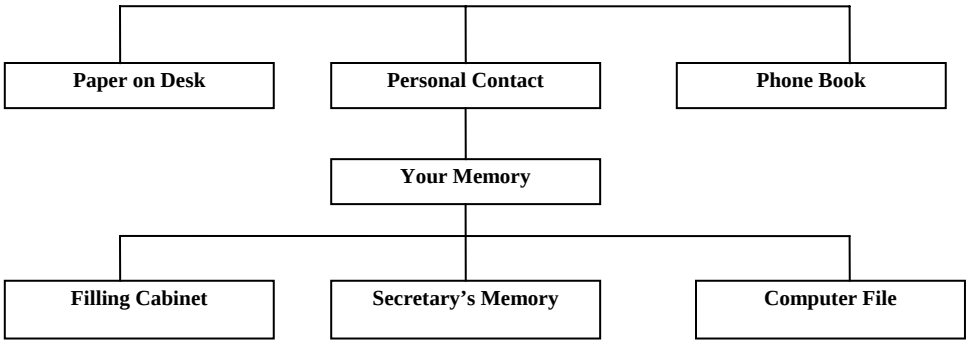
Terry van Druemel, Hasrat Madiadipura, Muhamad Arif and Gary Timoshenko [*]

Relational database

Quick and flexible access to information is essential for increasing the effectiveness of an organization. Once this is clear, organizations must decide which tool will best help them to manage their data and achieve this goal. Many options are available, but relational database management systems (RDMS) provide the only sensible choice for data management.

What is a relational database? A relational database is a collection of related information organized into separate tables (listings). This allows storage of related information in one place. That may sound cryptic, but consider the following real life example of a relational database. You will find in your office many collections of information. The filing cabinet is one collection, and a phone directory, a computer hard drive, and a pile of paper on your desk are others. The data that you and your secretary store in your heads are two more collections of information. To demonstrate the 'relational' aspect of this database, think about what you do when you make a phone call to one of your personal contacts. You don't have to memorize a phone number to call a personal contact. Instead, you remember his last name, a way in which you relate to him. Your contact is related to the phone book by his last name, too. You can access his phone number by looking for his last name in the phone book, which reduces the amount of information you need to keep in your head. To use the analogy further, keeping data in a spreadsheet is like memorizing all of the numbers in the phone book.

Figure 1 Relational database (real life).



An RDMS is a database that stores information in tables - rows and columns - of data and conducts searches by using data in specified columns of one table to find additional data in another table. In the diagram above, each box might represent a separate table. In a relational database, the rows of a table represent records (information about separate items, such as one

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

employee, or a crop in a given year) and the columns represent fields (particular characteristics of a record).

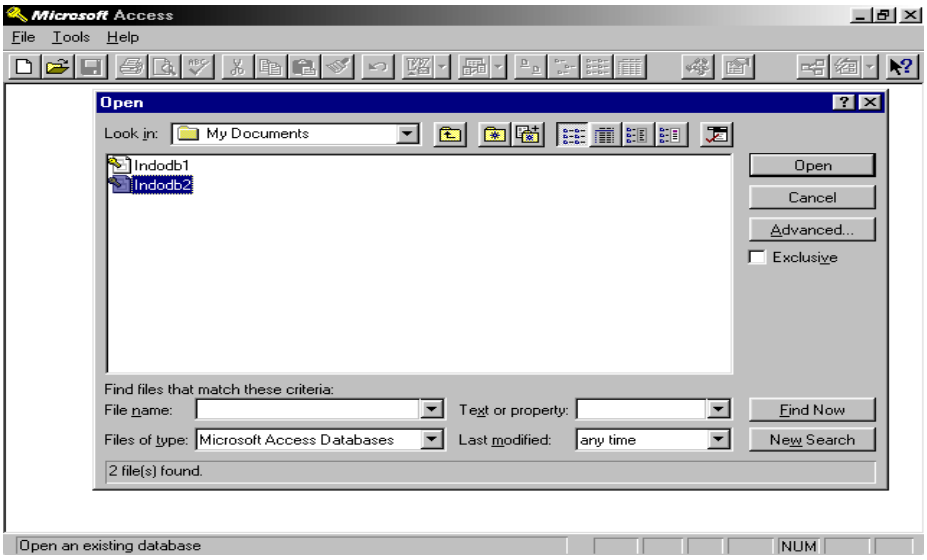
The ability to search for any combination of data in a database (known as a query) is the single most important feature of any RDMS. In conducting searches, a relational database matches information from a field in one table with information in a corresponding field of another table to produce a third table that combines requested data from both tables. For example, if one table contains the fields District ID, District Name and District Population and another table contains the fields District ID, Year and Consumption, a relational database can match the District ID fields in the two tables to find, say, per capita consumption of a given crop in a given year. In other words, a relational database uses matching values in two tables to relate information in one to information in the other.

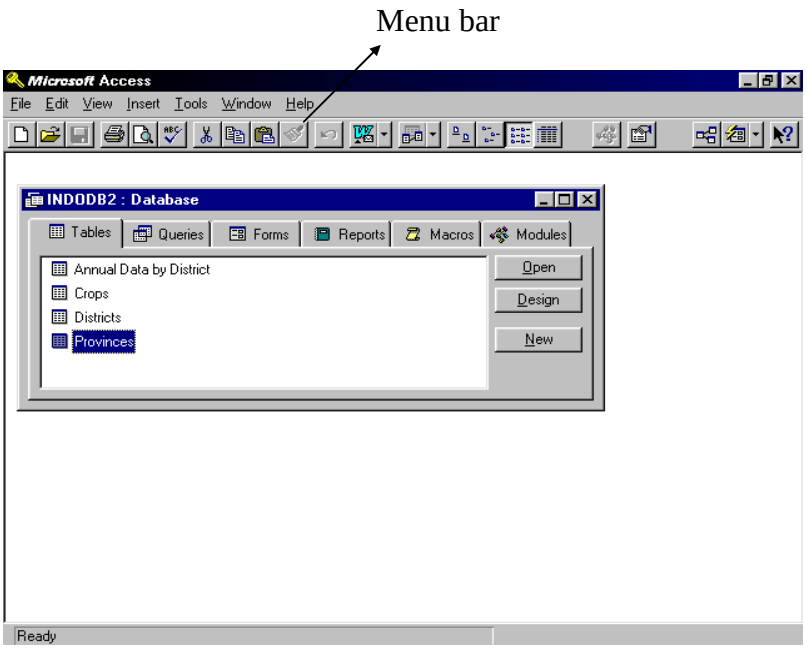
Relational database systems also help you to enter data quickly and correctly. Using forms, input masks and data validation rules, anyone can input data into the system with limited opportunity for making mistakes, even if they have never used an RDMS before.

Introduction to Microsoft Access version 7. 0

We have chosen to use Microsoft Access 7.0 as the RDMS software we will use to create our sample database. Microsoft Access is compatible with other Microsoft products, including Word and Excel. Data can also be imported from and exported to other relational database systems (e.g. Paradox, Dbase). The user interface and on-line help make Access quite user-friendly, at least as far as database software goes. Let's begin by taking a look at the different parts of the Access screen:

- Start the Access program from the Start menu.
- From the first screen click **Cancel**.
- From the **File** menu, choose **Open Database**.
- Double click to open the **Indodb2.mdb** (file available from the CGPRT Centre website http://www.cgprt.org.sg).





Select each menu, and familiarize yourself with the functions that each one performs. If you have used other Windows software, you will recognize most of the options in the **File** menu, from which you can open, close, import, and export files (databases), and print. You will also be acquainted with the **Edit** menu, since you will have seen the *Cut, Copy and Paste* commands in many other programs. The top section of the **View** menu allows you to select which part of the database you want to work with. You can also select each tab in the database window to perform the same action. The **Tools** menu contains special functions like spell checking, setting relationships between tables, and a **Security** menu. The **Security** menu allows you to control who has the ability to use and modify the database. If you have more than one window open at a time, you may arrange them using the **Window** menu. The **Help** menu is an excellent source of information regarding the software, and also lets you use 'cue cards' to walk you through the various steps used in creating and using databases. If you are ever stuck or confused, simply press the **Fl** key, and the program will automatically display on-line help for the screen you are looking at. Many of the menus such as **View** and **Insert** change when we are creating tables or doing queries. Also new menus like **Format** and **Records** will appear.

Access objects

Microsoft Access uses ‘objects’ to perform various functions, as we might use a pen ‘object’ to write, or a scissors ‘object’ to cut. An object is something you can select and move, or change. Here are some brief descriptions of the objects you will use in Microsoft Access:

- Table - stores data in rows and columns.
- Query - asks a question about the data in your tables.
- Form - allows you to enter or view records of one or more tables.
- Report - lets you customize the way you print your records (i.e. in groups, or with associated totals).

- Macro - a series of pre-set instructions for the program to carry out.
- Modules - one or more Access Basic procedures.

During this course, we will be largely concerned with the first four object types. Macros and modules are objects that will be useful to advanced Access users who wish to automate their databases.

Now that you are familiar with basic Access concepts, let's examine the most important step of database creation.

Database design

Before we can begin putting our database together, we need to design it. Here are five steps to follow when designing a database:

- Step 1: Determine the purpose of your database - What will your database contain? Who will use the data, and how will it be used?
- Step 2: Determine the tables you need in your database - Each table represents a different 'subject' contained in your database. For instance, one table may contain all information about employees of an organization and another may contain information related to a computer inventory.
- Step 3: Determine the fields you need - Each field in a table is one category of information that describes the subject. For instance, an employee's last name and salary are two distinct characteristics of an employee, and each requires its own field.

Last Name	**First Name**	**Salary ($ Monthly)**
Evans	Mark	2000.00
Williams	Hank	2001.99

- Step 4: Determine the relationships between tables - Once you have put all of the subjects of your database into tables, you need to show how those subjects are related to one another.
- Step 5: Refine your design - After you have finished the first four steps, you may find that you need to add some information to one of the tables, or that other tables have duplicate and redundant information. You may have to do considerable 'tweaking' of the database design before you have something that is efficient and clear.

Creating a statistical database

The problem defined

Assume you have been given the task of creating and managing a database that will contain various statistics for many crops. These statistics are gathered annually, at the district level. For instance, one record in your database might contain data on maize for District X in 1994, and another record may contain data on cassava in District Z, for 1996.

How do we go about creating this database? Let's follow the five steps of database design one at a time.

The first step was taken in the definition of the problem above. We already know what data will be contained in the database, and we can guess that researchers will analyze the data to answer any number of research questions.

The second step requires that we divide our database into subjects. It helps to restate the main task of our database. We will be gathering data on various crops for different provinces and districts. So we can say that crops are one subject, districts are another and provinces yet another. Of course, The actual crop statistics per time period are also an important subject in our database. It is important that for each subject, the time and spatial factors are the same. For instance if you have annual provincial data, this cannot be in the same table as monthly national data. You would write these different subjects (tables) down on paper.

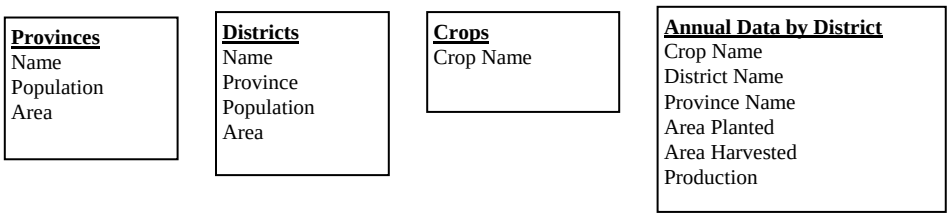
Provinces
Districts
Crops
Annual data by district

The third step: we need to determine which fields are needed in each table. For instance, under district we will have to know the district name so we include a District Name field. Since we know that population data are available for each district and population relates to labour availability, consumption, etc, we will also include a Population field for the Districts table. For the same reason we will include an Area field. It will also be useful to know which province the district is in so we will include a Province field. The Province table will contain Name, Population and Area fields.

For the Crop table the only information to include would be Crop Name. This is an important table, however, because it sets out the crop categories. There are countless forms and varieties of rice (shelled, unshelled, wetland, dryland, broken, sticky) and we have to know if each of these will be a separate crop or if we will only have one field for all rice.

For the Annual Data by Province table the fields will have to include the indicators we want (in this case Area Planted, Area Harvested and Production) and which crop, district and province the data is for.

So on our first try our tables will look like this:



These tables may be confusing because the Area field in the Province table and the Area field in the District table have the same name but do not contain the same data. To avoid confusion we will name them Prov Area and Dist Area. We should also do the same for the Population fields and the Name fields.

It is often useful to have numeric fields to identify character fields. We will be using the fields Crop Name, Province Name and District Name often within our database. These fields are character data. This can cause problems. Is Jogjakarta the same as Yogyakarta, is Corn the same as Maize. To avoid confusion we assign ID codes to these fields. We will create the numeric fields Crop ID, Province ID and District ID. In the Annual Data by District table we have used Crop Name, Province Name and District Name. These should be changed to Crop ID, Province ID and District ID. Our database now looks like this:

Provinces
Province ID
Province Name
Prov Population
Prov Area

Districts
District ID
District Name
Province ID
Dist Population
Dist Area

Crops
Crop ID
Crop Names

Annual Data by District
Crop ID
Year
District ID
Province ID
Area Planted
Area Harvested
Production

Every table in a relational database must have a primary key, which ensures that you don't accidentally enter the exact same data twice. The primary key is a field or combination of fields that uniquely identifies each record in a table. As the main index for the table, it is used to associate data between tables. You can have MS Access create this key field for you when you set up your table but most of the time it is best to designate the primary key yourself.

Looking at the Provinces table we can see that both Province ID and Province Name uniquely identify each record in the table. We will choose Province ID as the primary key because it is best to choose a numeric field as the primary key. The same is true for District ID and Crop ID. Finally for the Annual Data by District table, there is no field that uniquely identifies each row so we will add a field called Record Number so that each record has a unique number. This becomes our primary key.

The next step in design is to check our database for redundancies. If we look at the Annual Data by District table we find the fields District ID and Province ID. If we know which district the data is from we can look in the District table to find which province the district is in so we do not need the Province ID field. Now, our database design would look something like this (the fields in Bold represent primary keys):

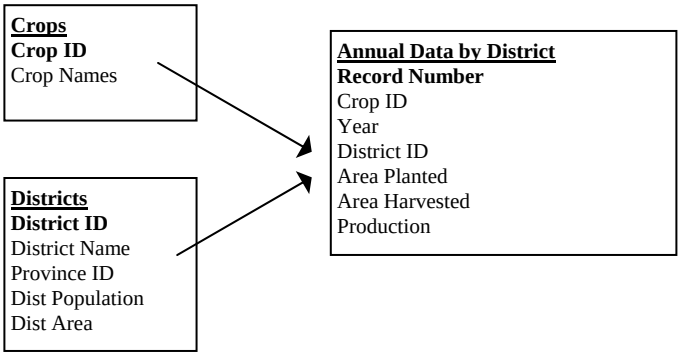
Provinces
Province ID
Province Name
Prov Population
Prov Area

Districts
District ID
District Name
Province ID
Dist Population
Dist Area

Crops
Crop ID
Crop Names

Annual Data by District
Record Number
Crop ID
Year
District ID
Area Planted
Area Harvested
Production

Now that we have an idea of which fields to include in each table, we can begin the fourth step, defining the relationships between tables. Notice the foreign keys in some of the tables. A foreign key is just a primary key that finds itself away from home. For instance, the Crop ID is a primary key at home in the Crops table, but it is a foreign key in the Annual Data by District table. These keys enable you to link one table to another, as shown in the following diagram.

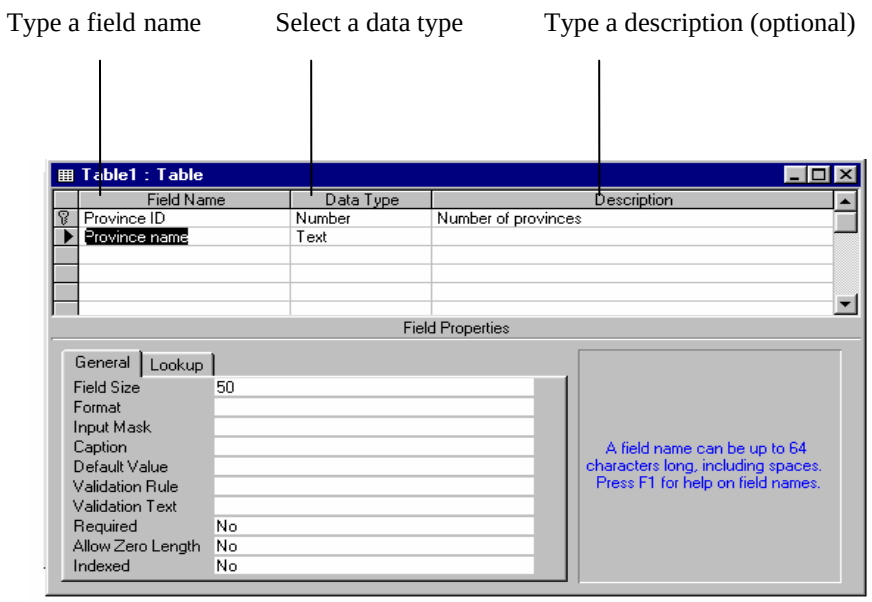


The advantages to linking the tables together in this way may not be immediately obvious. Imagine that you accidentally misspelled the name of an exotic crop. You would have made that same mistake thousands of times in the Annual Data by District table. If you had your data stored in spreadsheets, you would have to change the spelling in thousands of cells. With a search and replace function, this may not be so difficult, because you only have to search for the mistake and replace it with the proper spelling. Keep in mind, however, that so far, we are only working with a very small sample database. A more realistic database would probably have many more tables to search for errors. In our sample database, however, we would only have to fix the spelling once in a single place, the Crops table. The correction would carry through automatically to any other table that uses a crop name.

The fifth step involves refining your database design. This step actually takes place all the way through design and may need repeating many times. With a relational database, it is simple to make this and other refinements which will have a large effect on the whole database.

Data types and descriptions

We now have the design of our database but when we enter the design in Access we will also need to know data types, item description and any special field properties like field size or format. This is demonstrated in the Table Design Window presented below. When a table is designed or modified, specify the fields that it should contain in the upper portion of the Table Design window. In the lower portion of the window, enhance the table by setting properties for each field. The records in a table contain several categories of information. A field can be added to the table for each category of information to be stored. When you create a table, you start by specifying fields in the first row of the upper portion of Table Design window. In each row enter the field name and data type. An optional field description can also be included.



A field name can contain a maximum of 64 characters. These characters can include letters, numbers, spaces (except leading spaces), and special characters except the period (.), exclamation mark (!), backquote (‘), either the left or right brackets ([ ]), or printer control codes.

The default data type is text with a default field size set at 50 characters. All default settings can be reset to reflect the field contents. If the selection of a data type or any of its related properties does not agree with the data entered, Access will not accept that entry. The Data Type property uses the following settings:

Setting	**Type of data**
Text	(Default) Text or combinations of text and numbers. Maximum of 255 characters.
Memo	Lengthy text or combinations of text and numbers. Up to 64,000 characters.
Number	Numeric data used in mathematical calculations.
Date/Time	Date and time values for the years 100 through 9999.
Currency	Currency values and numeric data used in mathematical calculations involving data with one to four decimal places. Accurate to 15 digits on left side of the decimal separator.
AutoNumber	A unique sequential (incremented by 1) number or random number assigned by Microsoft Access whenever a new record is added to a table. AutoNumber fields can’t be updated.
Yes/No	Yes and No values and fields that contain only one of two values (True/False, On/Off).
OLE Object	An object (such as a Microsoft Excel spreadsheet, a Microsoft Word document, graphics, sounds or other binary data) linked to or embedded in a Microsoft Access table.

Lookup Wizard… Creates a field that allows you to choose a value from another table or from a list of values using a combo box. Choosing this option in the Data Type list starts the Lookup Wizard to define the data type.

The following is a list of field names, data types and descriptions for our database:

Table: Provinces

Field Name	Data Type	Description
Province ID	Auto Number	Province code
Province Name	Text	
Prov Population	Number	Population in thousands
Prov Area	Number	Area in sq. km

Table: Districts

Field Name	Data Type	Description
District ID	Auto Number	District code
District Name	Text	
Province ID	Number	
Dist Population	Number	Population in thousands
Dist Area	Number	Area in sq. km

Table: Crops

Field Name	Data Type	Description
Crop ID	Auto Number	Crop code
Crop Names	Text	

Table: Annual Data by District

Field Name	Data Type	Description
Record Number	Auto Number	
District ID	Number	District code
Crop ID	Number	Crop code
Year	Number	
Area Planted	Number	Area planted (ha)
Area Harvested	Number	Area harvested (ha)
Production	Number	Production (mt)

Now that we have our design in place, we can go ahead and build the database using Microsoft Access 7.0

Building the database in Access 7. 0

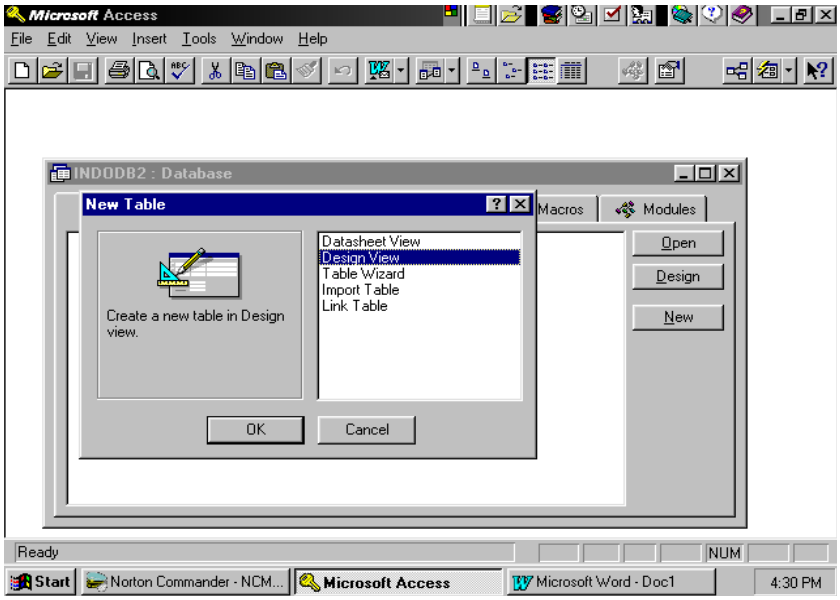
Creating the tables

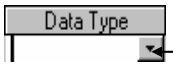
Return to MS Access, and close any database you may have open. To create a new database, do the following:

1. From the **File** menu, choose New **Database,** then click on the "Blank Database" icon.
2. Where it says Filename, type the name of your database - **test**.
3. Click **Create**.

You are now presented with an empty database window. Let's begin filling it with the tables that we have designed.

5. The **Tables** tab is selected by default, so simply click the New button.
6. The Microsoft Access displays the **New Table** dialog box.



7. Select **Design View** and click the **OK** button.
8. Since we always need a primary key, let's enter that **field** first. Type **Province ID** under the Field Name, and then hit the **Tab** key.
9. Select Auto Number from the Data Type drop down list. 

Drop down list button
10. Under **Description,** type "Province Code"
11. To set this field as the primary key click on **Province ID** in the Field Name column then click the Primary Key button on the toolbar.  A picture of a key should appear next at the left of the **Province ID** row.
12. Under Field Name in the next row, type **Province Name**, then hit **Tab**.
13. Select **text** from the drop-down list for Data Type.
14. You can leave the description blank if you like.
15. In the next row under Field Name type **Population** and select data type **Number**. Type the description as "Population in Thousands".
16. Finally, in the next row under Field Name type **Area** and select data type **Number**. The description will be "Area in Sq. Km".

We have completed designing our first table. Once we close the table, we can enter some data.

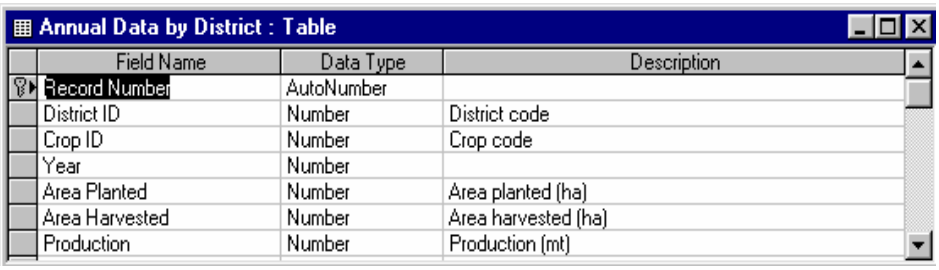
17. Select **Close** from the **File** menu, and choose **Yes** to save the table.
18. Type **Province** when asked for a Table Name, and click **OK.** There is now a table object in the database window, called Province.
19. Double-click the **Province** table, and this opens the table for viewing and data entry. A field with a counter data type doesn't need to have any data entered - it will advance the counter each time you enter a new record (or Province Name in this case).
20. Hit the **Tab** key on your keyboard to advance to the next field, and type "Aceh".

21. Hit the **Tab** key again to move to the next field, and type "3,546" for the population of Aceh. Hit one more tab and enter "57,070" for the area of Aceh.
22. Hit tab to go to the next record, and tab again to move to the next field, and type the name of the next district, "North Sumatra".
23. Hit tab again and enter "11,867" for the population of North Sumatra and enter "72,570" for the area.
24. You can close the table again, but notice that it doesn't ask you if you want to save the changes. Access automatically saves any data immediately after you enter it.

Now that you are familiar with the process for creating new tables, finish creating the tables that we have already designed.

Here are some helpful hints:

- Foreign keys in tables (e.g. the Province ID field in the Districts table) are number data type, not counter, because their value is determined elsewhere. As a reminder, you can type the description "Links this table to the _ table."
- Remember that periods (.) are not allowed in the field names. If using abbreviations like Prov or Dist do not use a period at the end.
- Under description, include the units of measurement for your fields (eg. hectares, kilograms, millions). For example when creating the Annual Data by District table we will enter the following:



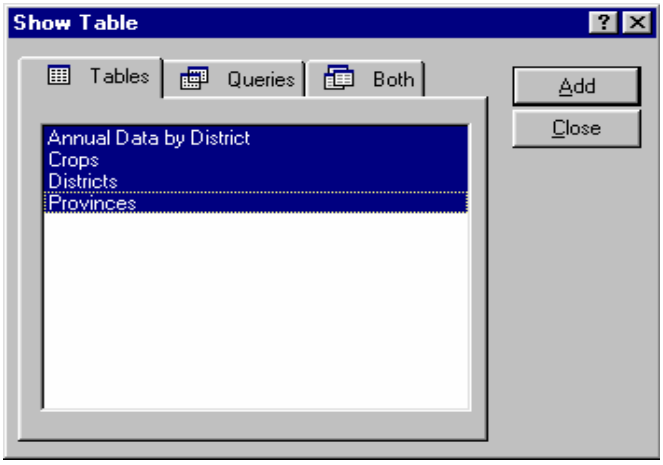
- In the bottom have of the screen you can determine the field size, the Format and Input Mask for the data you are entering. For instance, for monthly data, we would have a data type **Date/Time,** use **mm/yy** for the format and ##/## for the input mask. Now users will be able to quickly key in 11 97 for November, 1997. It is worthwhile experimenting with different formats, input masks, and validation rules, since these features will significantly contribute to the validity of your database.

Creating links between tables

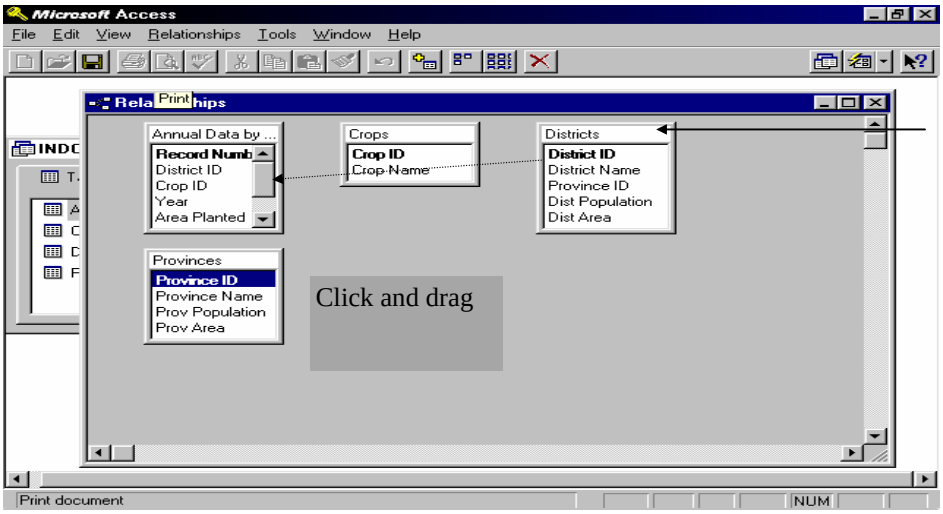
Now that we have finished creating our tables, we need to specify the relationships (links) between tables. Relationships are used to link data found in separate tables. A link is possible when the same field is found in two or more tables. In our database we see that Crop ID is found in the Crops table and in the Annual Data by District table. District ID is found in the Districts table and in the Annual Data by District table and Province ID is found in the Provinces table and in the Districts table. These will be our relationships. We are relating fields with the same name. Related fields do not have to have the same name but if the fields are using the same data they should have the same name (like the ones in our database). Likewise, even if fields in different tables have the same name this does not mean that they can be linked because they may not use the same data.

To create relationships, from the **Tools** menu, choose **Relationships** or click the Relationships button on the toolbar  . This opens the *Relationships* window.

Before you can work with the Relationships window, you need to choose which tables you want to relate to one another. We want to include all tables in our relationships diagram. Hold down the **Shift** key while tapping the **Down Arrow** to select **all** tables in the Add Table window. Then click **Add** and **Close.**

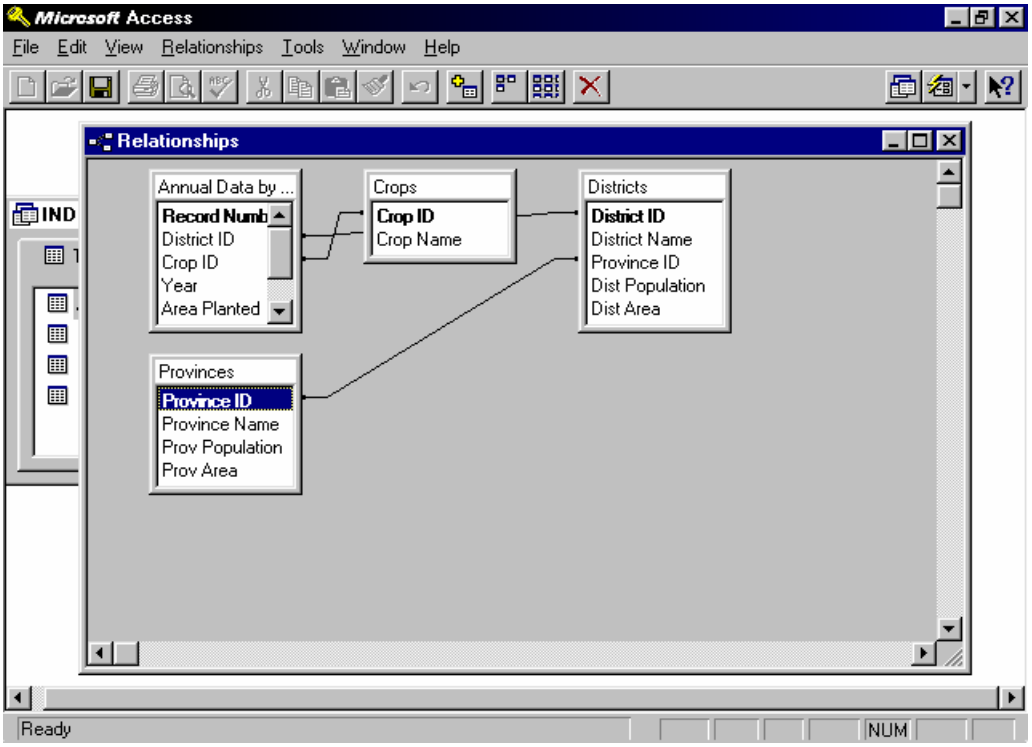


All of the tables are placed in the Relationships window. Your screen should look like the one pictured below. We can arrange the tables by clicking and dragging the title bars of each table.



To create relationships click and hold down the left mouse button on one field and drag it to the field you want to relate it to. Click and drag the bold primary key **District ID** from the **Districts** table, and drop it onto the foreign key *District ID* in the **Annual Data by District** table. Simply click the **Create** button on the window that pops up to create a link. Do this again

for the other primary and foreign keys that you have in the tables. **Crop ID** in the **Crops** table will link to **Crop ID** in the **Annual Data by District** table and **Province ID** in the **Provinces** table will link to **Province ID** in the **Districts** table. You should end up with something similar to the following picture.



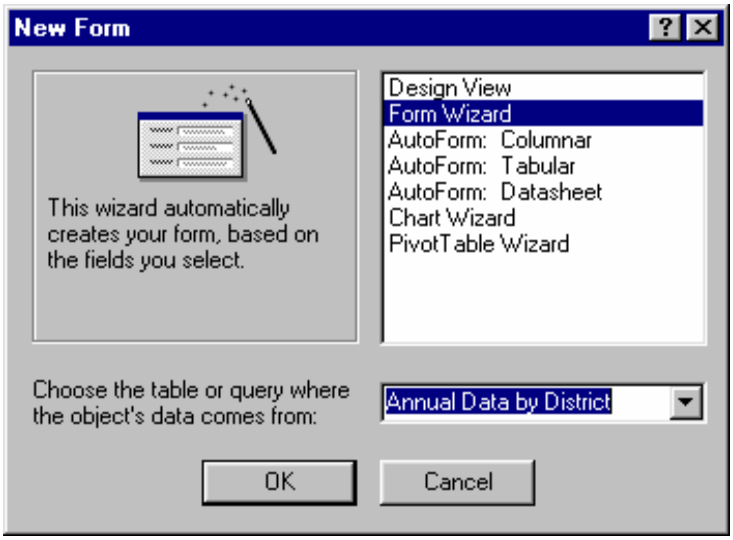
You can close the Relationships window and save the changes that you have made. Now we are ready to start entering data.

Creating and using forms

One of the quickest ways of entering data that span more than one table is to create a form. If you are only going to enter a few items into a table, like the names of crops, you may simply wish to open the table and enter the crop names directly, like we did in the Province table in an earlier section. However, if you will be entering large volumes of data on an on-going basis, you will likely want to create a form.

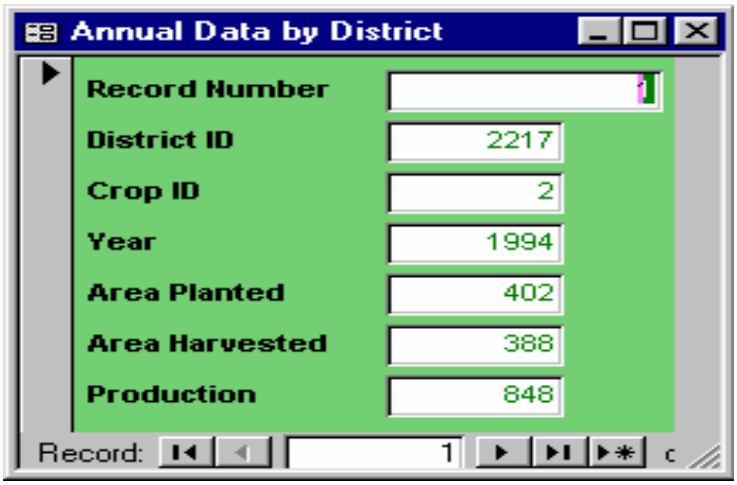
We will be using a database similar to the one we have just created, but to save time, some data has already been entered.

Open the database called **Indodbl.mdb.** Click on the **Forms** tab in the database window. Click the **New** button, and Access displays the **New Form** box:



1. Select **Form Wizard** and below that choose **Annual Data by District** where it says to "Choose the table or query where the object's data comes from:" Click OK.
2. Click the double arrows [>>], to include all fields in your form, then click Next.
3. When asked what layout you would like for your form choose "Columnar" then click Next.
4. When asked what style you would like the chart click "Evergreen" then click Next.
5. You should be on the last screen of the Form Wizard (which shows a checkered flag). Under the question "What do you want to do?", turn on the option "Open the form to view or enter information". Now click **Finish.**

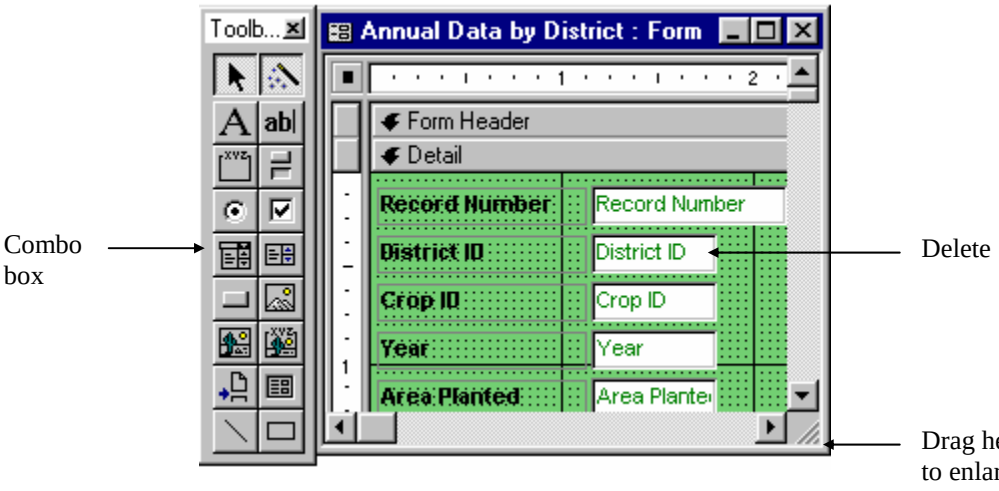
The resulting form should look like this:

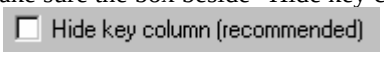


In this form we are required to enter the District ID. District ID will be difficult to remember because it is an arbitrary number. It would be easy if we could just enter the district name instead of having to remember each district ID number. Even better would be to select the district we wanted without typing it out.

Here is how we can do that:

1. From the **View** menu click **Form Design.** Notice that a toolbar appears either on the right or left hand side of the screen. Also the Annual Data by District Form may not be large enough to see the entire form. Enlarge the form by moving the mouse pointer to the edge of the form then clicking and dragging.



2. **Select** and **delete** the **District ID** field and label.
3. **Select** the **Combo Box** button  from the toolbar to the left or right of the form.
4. Move the mouse pointer into the spot where the **District ID** field and label used to be and click.
5. The Combo Box Wizard appears. Click **Next** so that we can have the combo box look up values in a table.
6. Select the **Districts** table to provide the values for the combo box, and click **Next.**
7. Include both **District ID** and **District name** in the combo box values, and click **Next.**
8. Make sure the box beside "Hide key column (recommended)" does not have a check in it. 

Click **Next**.
9. In the next window choose **District ID** as the field that uniquely identifies a row, then click **Next**.
10. Click **Store that value in this field,** and select **District ID** from the pull down menu**.** Click **Next.**
11. Enter **District ID** for the combo box label. Now click **Finish.**
12. The form design window appears but the label and input box for **District ID** are not in the correct place. You may want to move and resize the **District ID** combo box to make the form look nice. Do this by clicking and dragging the sides of the boxes.
13. Now click **View**, **Form** to view your new form. The new form appears with an arrow beside the **District ID** box. When you click the arrow the list of districts appears and

you can select one of these districts from the list. This automatically enters the **District ID**.

14. When you are finished, **Close** the form, and save it as **Annual Data by District**.

Creating and using queries

Queries are simply questions about information contained in a database. They can ask questions about a single table or from multiple tables. (With data stored in a spreadsheet, you are limited to questions about single tables.) This is where the linking of tables we performed earlier becomes important and where the power of relational databases becomes apparent.

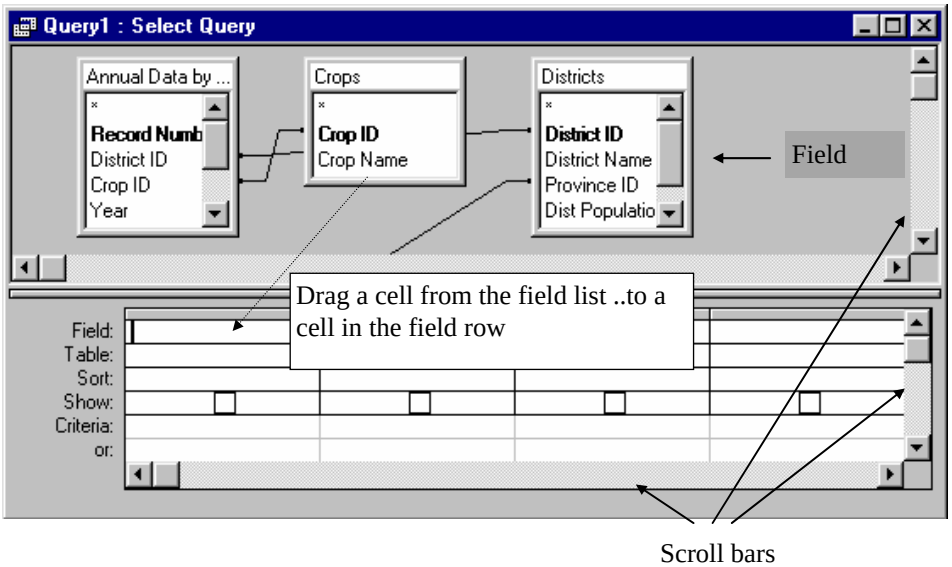
We will create two queries. The first one will provide us with data we can export to mapping software and the second one will demonstrate calculated fields.

For the first query we want to find maize production for the province of West Java in 1995.

1. If you haven't already done so, close any database you have open.
2. Open the database entitled **Indodb2.mdb,** which is our original database, but with data already entered. The Indodb2.mdb file is available from the CGPRT website http://www.cgprt.org.sg.
3. Click on the **Query** tab in the database window, and then click the **New** button.
4. In the **New Query** box choose Design View.
5. In the Show Table Box choose all the tables. To select all the tables hold down the **Ctrl** key and click on each of the tables. Once you have selected these tables, click the **Add** button, followed by the **Close** button.

Choosing fields

After you have added tables to our query, you are ready to select the fields to include in the query. The fields shown in the field list belong to the tables or queries you selected in the Add Table dialog box. The fields you select determine the data you see when you view query in Datasheet view. You should now see the **Select Query** window.



You will have to use the scroll bars to view all the tables in the top half of the window.

To add a field to a query

Drag the field from the field list to a cell in the Field row of the bottom grid. To add more than one field at a time, hold down the CTRL key and click the fields you want to add; then drag the group to a cell in the Field row. To select a block of fields, select the first, hold down the SHIFT key and click the last field, and than drag.

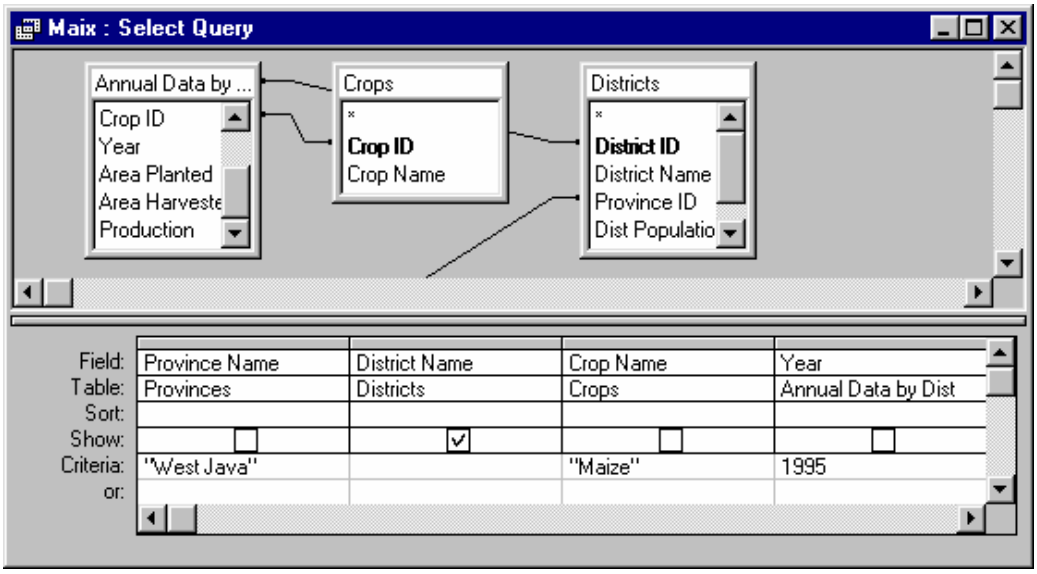
1. Click on the field **District Name** from the **Districts** table and drag it to the Field row in the first column.
2. Click on **Province Name** from the **Province** table and drag it to the Field row in the second column.
3. Click on **Crop Name** from the **Crop** table and drag it to the Field row in the third column.
4. Click on **Year** from the **Annual Data by District** table and drag it to the field row in the fourth column.
5. Finally click and drag **Area Harvested** and **Production** from the **Annual Data by District** table and drag these fields to the field row of the fifth and sixth columns. To view the fifth and sixth columns you will have to use the scroll bars at the bottom of the Query box.

To view a query

When you're finished your query, you're ready to see the results. The quickest way is to click the **View** menu and choose **Datashee**t or click the Datasheet View button on the toolbar  .

Now we want to modify our query. We only want data on Maize production in West Java in 1995. To change the query, choose **Query Design** from the **View** menu (or click the Design View button on the toolbar). We will make the following changes.

1. In the Province Name column, in the Criteria row type "West Java". In the Show row click the box so the check mark disappears. This means the Province Name will not be seen when we view our query.
2. In the Crop Name column type Maize in the Criteria row and also click the box Show row so the check mark disappears.
3. Finally, in the year column type 1995 in the Criteria row and click the box in the Show row so the check mark disappears.



Choose **Datasheet** from the **View** menu to see the new query (dynaset).

Saving a query

When you save a query, it becomes a part of the database. Microsoft Access provides a default query name, but it's a good idea to give your query a more descriptive name. To save and name a query:

- From the file menu, choose **Save** (or click the Save button on the toolbar).
- If you are saving the query for the first time, type a name for the query, in this case Maize95 and choose **OK.**

Our final exercise will be to add a calculated field to our query. From the Area Harvested and Production fields we will calculate the yield. From the database window choose Queries, click on Maize95 and click Design. In a blank column in the field row type:

[Annual Data by District]![Production]/[Annual Data by District]![Area Harvested]

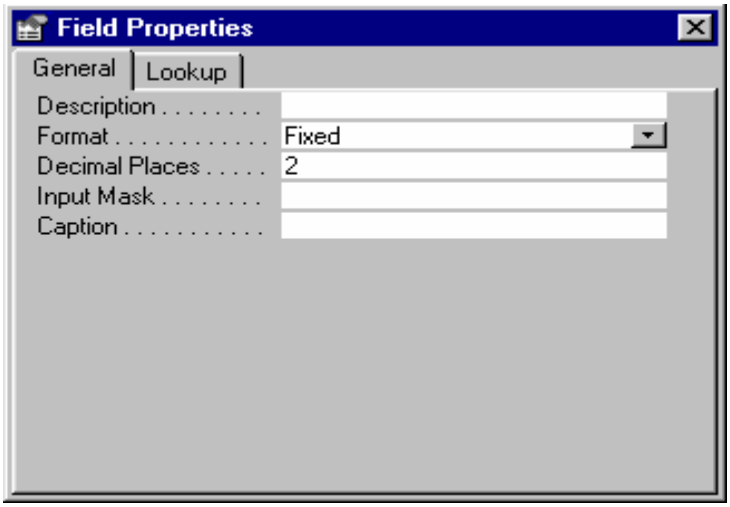
When you press Enter the expression changes to:

Expr1: [Annual Data by District]![Production]/[Annual Data by District]![Area Harvested]

"Expr1:" is the name of the field. Use the mouse to click on "Expr1:" and change it to "Yield:".

Now click **View**, **Datasheet** from the menu bar to see the results of your query. The yield column has too many decimal places to be viewed properly. To change this click on **View**, **Query Design.** With the mouse pointer in the Yield column we have just created, click the right mouse button. In the menu that pops up click **Properties**. Now beside Form type **Fixed** and

beside Decimal Places type **2**. Close the window and click **View**, **Datasheet** again to see the results.



Now under the **File** menu click **Save As/Export** and then "within the current database as" **Yield95**. Now click **File, Close.**

Exporting

You can export data from tables or queries to spreadsheets, or to any database format. The data can then be used for various applications like mapping or statistics.

To export from Access

You can use this procedure to export data to Excel, Lotus 1-2-3, Paradox, FoxPro, or dBASE files. To export proceed as follows:

- Open the Access database containing the table or query you want to export. In this case we will open our query **Maize 95**.
- In the Database window, choose **Save As/Export** from the **File** menu. Access displays the Save As dialog box. Click "To an external File or Database" and click OK.
- We now have a choice of saving our query in many different formats. We will save the query as Maize95 but in save as type Microsoft Excel 5-7.

Our data can now be used in Excel or easily imported in mapping or statistical software.

Geographic Information Systems: An Overview

Mohammad A.T. Chowdhury[*]

Introduction

Geographic information systems (GIS) are computer-based systems for the collection, storage, analysis and display of objects and phenomena where geographic location is an important characteristic or critical to the analysis. For example, GIS information mey be crucial for the location of a fire station or for identifying the locations where soil erosion is most severe. In each case, what it is and where it is must be taken into account.

Over the past two decades, GIS technology has developed so rapidly that it is now accepted as an essential tool for the effective use of spatial (land-based) information. The recent and widespread introduction of GIS has created a sudden need for users of geographic information to become knowledgeable about this technology. Managers within public and private organizations are being called upon to make decisions about the introduction of GIS technology and to establish policies for its use. Politicians are being asked to support expensive programs to convert map data into digital form suitable for use with GIS.

The technology has provided an exciting potential for geographic information to be used more systematically and by a greater diversity of disciplines than ever before. However, the ease with which GIS can manipulate geographic information has also created a major difficulty. Users unfamiliar with GIS techniques or the nature of geographic information can just as easily conduct invalid analysis as valid ones. Valid or not, the results have the air of precision associated with sophisticated computer graphics and volumes of numerical tabulations. A better understanding of GIS technology by users, managers, and decision-makers is crucial to the appropriate use of the technology. This section is addressed to those users who may have no previous experience with computer-based geographic information handling but who need to use or direct the use of GIS technology. It provides a concise introduction to the fundamentals of GIS, the capabilities of these systems, and some of the issues that arise when GIS is implemented.

The map: the original geographic database

Geographic information has traditionally been presented in the form of a map. Until computers were available, geographic data were represented as points, lines, and areas drawn on a piece of paper or film. They were coded using symbols, textures, and colors that were explained in the map legend or accompanying text. The map and its documentation constituted the original geographic database.

From the earliest civilizations, maps have been used to portray information about the earth's surface. Navigators, land surveyors, and the military used maps to show the spatial distribution of important geographic features. Land surveying and map-making were an integral part of the Roman government. It rose to prominence in Europe in the eighteenth century as

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

governments realized the value of mapping as a means of recording and planning the use of their lands. National institutes were commissioned to produce map coverage of entire countries. General purpose maps showing the topography of the land and boundaries of national or administrative units were produced. As the study of natural resources developed, thematic maps were used to portray the spatial distribution of such features as geomorphology, soils and vegetation.

In the twentieth century the pace of science and technology acceleterated. This increase created the demand for even greater volumes of geographic data to be presented in map form more quickly and more accurately. For example, with the development of satellite-based remote sensing technology, there has been an explosion of geographic data production, wider use and more sophisticated analysis. Geographic data are now being generated faster than they can be analyzed.

Emergence of GIS: a brief historical overview

Since the 1960s, decision-making has become increasingly quantitative and mathematical models have become common place. Two factors inhibited the full use of these techniques in natural resources management. First, spatial analysis involves tremendous volumes of data. Manual cartographic techniques allowed manipulation of these data: however, they were inherently limited by their non-digital nature. Traditional statistics and mathematics enable quantitative analysis, but the magnitude and detail of the digital data sets were prohibitive for the GIS of that time. However, the computer has provided the means for both efficient handling of voluminous data and the numerical analysis that is required. The current revolution in GIS is thus rooted in the digital nature of the computerized map.

It was only in the 1970s with the availability of suitable digital computers that the technology to handle spatial data leapt forward. Computer-based GIS was developed to provide the power to analyze large volumes of geo-referenced data. In the early 1970s, GIS focused on computer mapping to automate the cartographic process. The mapping programs of SYMAP and Odyssey developed at the Harvard Laboratory for Computer Graphics and Spatial Analysis are examples of this pioneering work.

During the early 1980s, the change in format and computer environment of map data were exploited. Spatial database management systems were developed that linked computer mapping capabilities with traditional relational database management systems. Two alternative data structures (raster and vector) for encoding maps were debated. By the mid 1980s, the general consensus within the GIS community was that the nature of the data and the processing desired determine the appropriate data structure. The increasing demands for mapped data focused attention on data availability, accuracy and standards. Hardware vendors continued to improve digitizing equipment, with manual digitizing tables giving way to automated scanners in many GIS facilities. A new industry for map encoding and database design emerged, as well as a marketplace for the sales of digital map products. Regional, national and international organizations began addressing the necessary standards for digital maps to ensure compatibility among systems. With the recognition of the digital nature of mapped data, applications of GIS became increasingly quantitative. Complex spatial models were developed using an approach analogous to traditional maths and statistics. The application of this new technique in natural resources is revolutionary.

The nature of GIS

GIS uses geographical position, or location, as a common thread to integrate and analyze information from a variety of sources. In essence, the system starts with a computerized topographic map as it's base, and then overlays other types of information from other databases. The major advantage of GIS is that it allows one to identify the spatial relationships between map features. New computer technologies now allow users to link this information and perform integrated analysis. It may be noted that GIS does not store a map in any conventional sense; nor does it store a particular image or view of a geographic area. Instead, GIS stores the data from which one can draw a desired view to suit a particular purpose.

Components of GIS

GIS is typically made up of the following components:

I. hardware,
II. software,
III. data,
IV. people, and
V. organization.

Here, we will focus on data, particularly, the way it is related to the construction of a spatial database, management, manipulation, analysis and output. The following is a set of capabilities that GIS can provide to handle geo-referenced data:

I. data input (construction of spatial data base),
II. data management (data storage and retrieval),
III. data manipulation and analysis (spatial operations), and
IV. data output (map production).

Data input

Central to the system is the database - a collection of maps and associated information in digital form. Since the database is concerned with earth surface features, it can be seen to be comprised of two elements, a spatial database describing the geography (shape and position) of earth surface features, and an attribute database describing the characteristics or qualities of these features.

The data input component converts data from their existing form into one that can be used by the GIS. Georeferenced data are commonly provided as paper maps, tables of attributes, electronic files of maps, air photos, and even satellite imagery.

The data input procedure can be as straightforward as a file conversion from one electronic format to another, or it can be complex. Data input is typically the major bottleneck in the implementation of GIS. Construction of large databases can cost five to ten times that of the GIS hardware and software.

It can take months to years to complete the initial data input. Thus, the expense and time needed to bring the GIS into full operation must be budgeted as part of the overall start-up plan. Once the inaccuracies (error in data input) have been rectified, the confidence of the users must then be rebuilt-- and the first impressions of users are remarkably resistant to change.

For this reason, data input methods and data quality standards should be carefully considered well before data entry is to begin. The various methods of data entry should be

evaluated in terms of the processing to be done, the accuracy standards to be met, and the form of output to be produced. The next important component is data management.

Data management

The data management component of GIS includes those functions needed to store and retrieve data from the database. The methods used to implement these functions affect how efficiently the system performs all operations with the data. There are a variety of methods used to organize the data into computer-readable files. The way the data are structured and the way files can be related to each other (the organization of the database) place constraints on the way in which data can be retrieved and the speed of the retrieval operation. The short- and long-term needs of the users should be identified and used in evaluating performance trade-offs. An expert at GIS database design and analysis procedures is needed to evaluate these trade-offs. The next section focuses on the most important component of GIS - the analytic functions.

Data manipulation and analysis

Data manipulation and analysis (also called as the GIS analytic functions) determine the information that can be generated by the GIS. A list of required capabilities should be defined as part of the system requirements. To anticipate the way in which the data in GIS will be analyzed requires that the users be involved in specifying the necessary functions and performance levels.

With geographic analytic functions, we extend the capabilities of a traditional database query to include the ability to analyze data based on their location. Perhaps the simplest example of this is to consider what happens when we are concerned with the joint occurrence of features with different geography. Suppose we want to find all areas of agricultural land on clay soil types associated with high rice productivity. This is a problem that a traditional data base management system cannot solve because clay soil types, land use divisions and productivity measures simply do not share the same geography. Traditional database query is fine as long as we are talking about attributes belonging to the same features. But when the features are different, it simply cannot cope. For this we need GIS.

The second set of tools that GIS will typically provide is that for combining map layers mathematically. Modeling in particular requires that we be able to combine layers according to various mathematical combinations. For example, we might have an equation that predicts mean annual temperature as a result of altitude. Or, as another example, consider the possibility of creating a soil erosion potential map based on factors of soil erodability, slope gradient and rainfall intensity. Clearly we need the ability to modify data values in our map layers by various mathematical operations (e.g. Boolean algebra) and transformations and to combine factors mathematically to produce the final results.

Data output

The output or reporting functions of GIS vary more in quality, accuracy, and ease of use than in the capabilities available. Reports may be in the form of maps, tables of values or text in hard copy or soft copy (electronic file). The functions needed are determined by the users' needs, and so user involvement is important in specifying the output requirements.

Data structure: raster and vector

The basic functions mentioned above are one aspect of the ways GIS vary. However, an even more fundamental distinction is how they represent map data in digital form.

GIS stores two types of data that are found on a map, i.e. the geographic definitions of earth surface features and the attributes or qualities that those features possess. Not all systems use the same logic to achieve this. Nearly all, however, use one or a combination of the two fundamental map representation techniques: raster and vector.

The raster

With raster systems, the graphic representation of features and the attributes they possess are merged into unified data files. In fact, one typically does not define features at all; rather, the study area is subdivided into a fine mesh of grid cells in which one records the conditions of attributes of the earth's surface at that point. Each cell is given a numeric value, which may represent either a feature identifier, a qualitative attribute code or a quantitative attribute value. For example, a cell could have the value "6" to indicate that it belongs to District 6 (a feature identifier), or that is covered by soil type 6 (a qualitative attribute) or that it is 6 meters above sea level (a quantitative attribute value). Although the data stored in the grid cells do not necessarily refer to phenomena that can be seen in the environment, the data grids themselves can be thought of as images - images of some aspect of the environment - or as layers - each one of which stores one type of information over the mapped region - that can be made visible through the use of a raster display.

In a raster display, there is also a grid of small cells called pixels. Pixel is a contraction of the term picture element. Pixels can be made to vary in their color, shape or gray tone. To make an image, the cell values in the data grid are used to regulate directly the graphic appearance of their corresponding pixels. Thus in a raster system, the data directly control the visible form we see.

The vector

The second major form of data representation is known as vector. With vector representation, the boundaries of the features are defined by a series of points that, when joined with straight lines, form the graphic representation of that feature. The points themselves are encoded with a pair numbers giving the X and Y coordinates in systems such as latitude/longitude. The attributes of features are then stored with a traditional database management software program. For example, a vector map of property parcels might be tied to an attribute database of information containing the address, owner's name, property valuation and land use. The link between these two data files can be a simple identifier number that is given to each feature in the map.

Raster versus vector

Raster and vector systems each have their special strengths and weaknesses. Raster systems are typically data intensive since they must record data at every cell location regardless of whether that cell holds information that is of interest or not. However, the advantage is that geographical space is uniformly defined in a simple and predictable fashion. As a result, raster systems have substantially more analytical power than their vector counterparts in the analysis of continuous space and are thus ideally suited to the study of data that are continuously

changing over space such as terrain, vegetation cover, rainfall and the like. The second advantage of raster is that its structure closely matches the architecture of digital computers. As a result, raster systems tend to be very rapid in the evaluation of problems that involve various mathematical combinations of the data in multiple layers. Hence they are excellent for evaluating environmental models such as soil erosion potential, and agricultural management suitability.

While raster systems are predominantly analysis oriented, vector systems tend to be more database management oriented. Vector systems are quite efficient in their storage of map data because they only store the boundaries of features and not what's inside those boundaries. Because the graphic representation of features is directly linked to the attribute database, vector systems usually allow one to roam around the graphic display with a mouse and inquire about the attributes associated with a displayed setup, such as the distance between points or along lines, the areas of regions defined on the screen, and so on. In addition, they can produce simple thematic maps of database queries.

Compared to their raster counterparts, vector systems do not have as extensive a range of capabilities for analysis over continuous space. They do, however, excel at problems concerning movements over a network space and can undertake the most fundamental of GIS operations. For many, it is the simple database management functions and excellent mapping capabilities that make the vector systems attractive. Regardless of the logic used for spatial representation of data, one can see that a geographic database is organized in a fashion similar to a collection of maps. Although there are subtle differences, for all intents and purpose, raster layers and vector coverages can be thought of as simply different manifestations of the same concept the organization of a database into elementary map-like themes.

Questions GIS can answer

So far, GIS has been described in two ways: i) through formal definitions, and ii) through its ability to carry out spatial operations, linking data sets using locations as the common key. One can, however, also distinguish GIS by listing the types of questions it can (or should be able to) answer. For any application there are five generic questions that a sophisticated GIS can answer.

Location: What is at.....?

The first of these questions seek to find out what exists at a particular location. A location can be described in many ways using, for example, a place name, a postal code, or a geographic reference such as longitude and latitude.

Condition: Where is it?

The second question is the converse of the first and requires spatial analysis to answer. Instead of identifying what exists at a given location, you want to find a location where certain conditions are satisfied (e.g., an unforested section of land at least 2000 square meters in size, within 100 meters of a road, and with soils suitable for multiple cropping).

Trends: What has changed since.....?

The third question might involve both the first two and seeks to find the differences within an area over time.

Patterns: What spatial patterns exist?
This question is more sophisticated. One might ask this question to determine whether high crop yield is a cause of fertilizer use in district X. Just as important, one might want to know how many anomalies there are that do not fit the pattern and where they are located.

Modeling: What if.....?
"What if..." questions are posed to determine what happens, for example, if as new road is added to the network, or if a toxic substance seeps into the local ground water supply. Answering this type of question requires geographic as well as other information.

Utility of GIS

As noted above, GIS is a powerful tool for handling spatial data. In GIS, data are maintained in a digital format. As such the data are in a form more physically compact than that of paper maps, tabulations, or other conventional types. Large quantities of data can also be maintained and retrieved at greater speeds and lower cost per unit when computer-based systems are used. The ability to manipulate the spatial data and corresponding attribute information and to integrate different types of data in a single analysis and at high speed are unmatched by any manual methods. The ability to perform complex spatial analysis rapidly provides a quantitative as well as a qualitative advantage. Planning scenarios, decision models, change detection and analysis, and other types of plans can be developed by making refinements to successive analysis. This iterative process only becomes practical because each computer run can be done quickly and at a relatively low cost.

GIS applications in agricultural planning

Today, the number and variety of applications for GIS are impressive. Local governments use GIS for planning, zoning, property assessment, land records, parcel mapping, land-use and environmental planning. Resource managers rely on GIS for fish and wildlife planning, management of forested, agricultural, and coastal lands, and energy and mineral resource management. The list can be exhaustive. We are concerned here about agricultural applications.

In the applied field of agricultural analysis, one deals first of all with biophysical characteristics of land: soil properties, topography, hydrology and so on. Moreover, climate is the second group of indicators. Here temperature distribution, rainfall, wind circulation, velocity and radiation are the main indicators. Now, one can construct the agricultural land types by combining a minimum of two indicators in a thematic map. The most basic step to take is to combine a climate and soil map in order to arrive at a group of land types with different characteristics. With GIS one can depict the various spatial distributions of these land types in a given area. With GIS one can calculate the distance between points, and can also calculate the exact surface area under study. This step is called the overlay.

Similar overlays can be used in other applications with a number of possible combinations. For example, GIS can be used to predict overall crop yields using information on soil conditions, rainfall patterns and yearly crop conditions and so on. In land administration, GIS can help in deciding what land could be developed and what portion should remain protected for agricultural use by assessing soil types and water availability. Land use

management and policy decisions are almost always based on the analysis of the interplay of factors pertaining to an issue. For instance, the preservation of prime agricultural land involves political, institutional and economic factors. GIS can provide better information to support this type of complex decision-making. Such basic information as the distribution of current land use activities, the relationship of these activities to agricultural capability, land prices, and urban demand for land can be developed quickly using GIS tools.

However, GIS does not operate in vacuum. To be successful, it must reside within a suitable organizational framework. The following section focuses on the spirit and purpose of GIS from an organizational perspective.

GIS from an organizational context

The GIS is operated by staff who report to management. Management is given the mandate to operate the GIS facility in such a manner as to serve some user community within an organization. Ultimately, the purpose and justification for the GIS facility is to assist the users in accomplishing the goals of their respective organizations.

As an organization becomes more familiar with a new system, people find new ways of getting a job done. They will develop analytical procedures different from those originally anticipated. While it is not possible to predict what these new methods will be, changes can be expected. The type and variety of functions provided by a specific system will influence the types of innovations that will occur.

For this reason, the management environment of the GIS facility is perhaps the most important single factor in determining its success or failure. It is the organization that in the end determines whether the physical equipment and human resources will function as an effective information system. The provision of effective user services, from training materials to qualified consultants, is critical to the effective utilization of the benefits that GIS can offer.

GIS is expensive to implement. Existing data must be converted to digital form, a task that is usually many times the cost of the hardware and software. GIS represents a significant overhead cost both to maintain the system and for the considerable degree of expertise required of the personnel who operate it. These costs are more easily justified if the data volumes are large, the data are frequently accessed, and can usefully be updated for a wide range of analyses. If these conditions do not apply, then GIS may not, in fact, be a cost-effective solution.

The major challenge in acquiring GIS is the development of analytic approaches that address the problem at hand and make effective use of the technology. Learning to operate GIS is relatively easy. It is far more difficult to learn how to apply this skill effectively and creatively to satisfy “real world” needs.

In many ways, learning GIS involves learning to think in spatial terms learning to think about patterns, about locations, areas, regions and about processes that act in space. As specific procedures are learned, they will often be encountered in the context of specific examples.

References

Aronoff, S. 1995. Geographic Information Systems: A Management Perspective. WDL Publications, Ottawa.

Berry, J.K.; and Ripple, W.J. 1994. Emergence and Role of GIS in Natural Resources. ASPRS, Bathesda, Maryland,

Eastman, J.R. 1995. Idrisi User's Guide. Clark University, Worcester, MA.

ESRI. 1995. Understanding GIS: ARC/INFO Method. John Wiley and Sons, Inc., New York,

Geographical Information Systems: MapInfo

Muhamad Arif, Siemon Hollema and Mohammad Chowdhury[*]

An introduction to MapInfo Professional

MapInfo has been selected as a representative geographical information system. MapInfo is relatively easy to learn and it does not require much in terms of hardware. It runs under Windows 3.1 or Windows 95. It is easy to import data from other programs such as databases and spreadsheets and to export the maps to other programs such as MS Word. This does not mean that other GIS software such as ArcInfo or IDRISI are not equally good. ArcInfo, for example, is a very complete GIS package used in many organizations throughout the world. For most applications however, MapInfo will suffice.

In MapInfo, all information is organized in tables. A table consists of files, either a map file or a database file, which contain the records and maps. A table consists of at least two files:

- *'filename'.tab* - a small text file describing the structure of the data.
- *'filename'.dat* - this file contains the actual data. If it is a DBASE, Lotus 1-2-3, or Excel file, the extension will be *.dbf*, *.wks*, and *.xls* respectively.

If the data contains graphic objects, i.e. items used to create and draw a map such as lines, points, polygons, rectangles, ellipses and text, two more files will be associated with the table:

- *'filename'.map* - a file describing the graphic objects of the map.
- *'filename'.id* - this file links the data with the graphical objects.

Finally, the table may include an index file. The index file is used to find different objects on the map. For example, you might want to search for a certain street, country or city on the map. If these fields are indexed, they can be located by using the *find command*. The index is stored in:

- *'filename'.ind.*

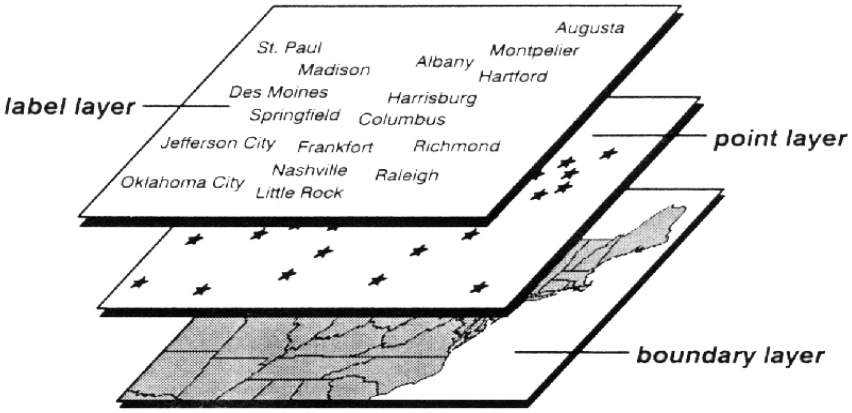
MapInfo offers different formats for viewing the data.

- a *map window* presents the information as an ordinary map.
- a *browser window* presents the information in rows and columns, just as in a spreadsheet or database.
- a *graph window* allows you to visualize the information as a graph (line, bar, pie or scatter).

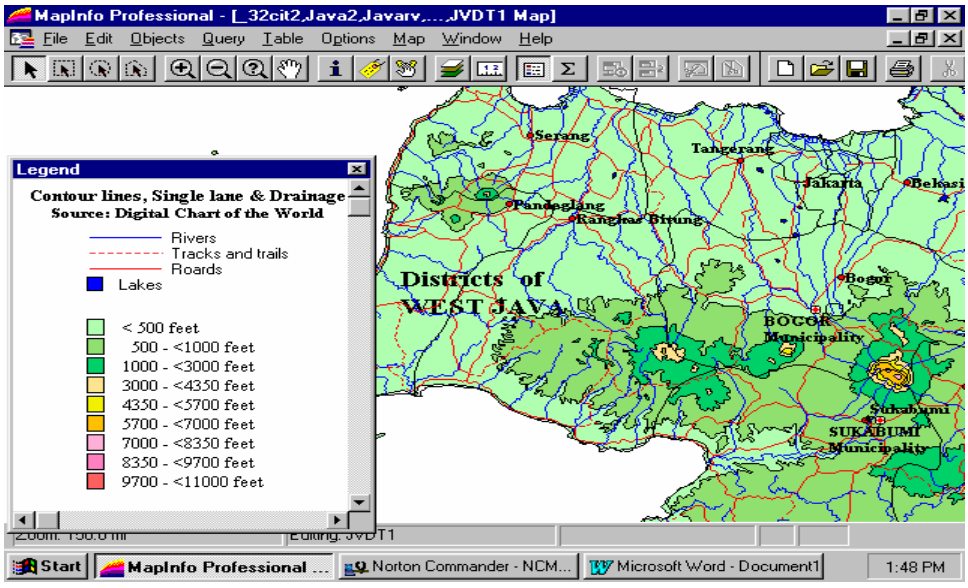
MapInfo allows the display of data in many different windows with different views at the same time. However, only one window can be active at a time. If changes are made in one window, they will automatically be reflected in the other. A *layout window* is available to combine the map, browser and graph windows into one layout, which can then be sent to the printer.

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

Computer maps are organized in layers. They can be thought of as transparencies that are stacked on top of one another. Each layer shows different aspects of the map. They contain the different map objects, such as regions, points, lines and text.



For example, one layer may contain the state boundaries, a second layer symbols, representing the capital towns, and the top layer text labels giving the names of the capital towns. By stacking these different layers on top of each other, a complete map can be created. Layers can be in different order. Layers can be deleted, added and edited. Each table is displayed as a separate layer. One, two or many tables can be opened at the same time. If possible MapInfo combines the tables in one single *map window*. For example, the following map of West Java is a combination of several tables. The open tables are displayed in the upper bar of the map window [Java2, Javarv, Javard1,…, JVDT1 Map]. The triple dots indicate that there is insufficient space to display all the open tables.

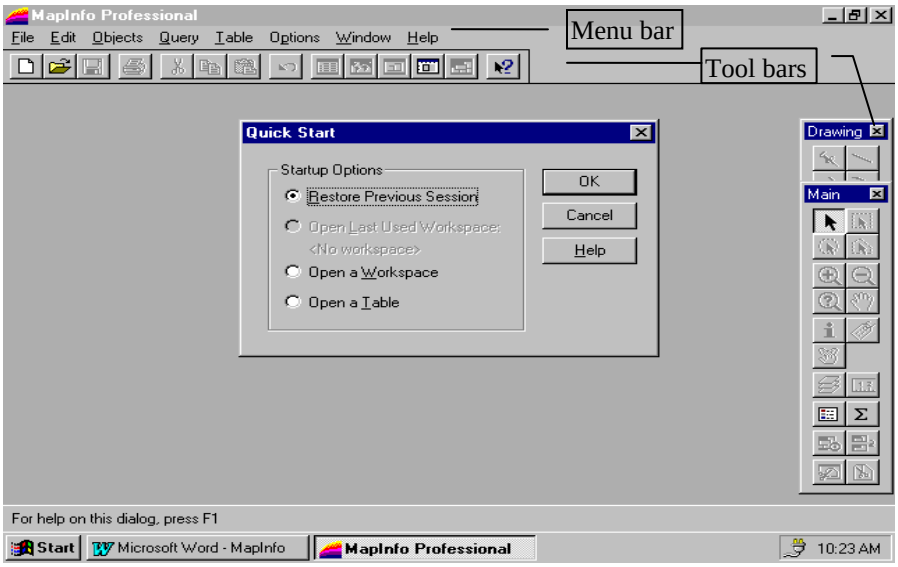


Digitized computer maps are readily available from different sources or can be created using the *digitizing feature* of MapInfo. In order to place data on the map, they need to contain (X,Y) coordinates. Without the (X,Y) coordinate MapInfo does not know where to position the data point on the map. The process of assigning coordinates to the data is called *geocoding*. With geocoding the (X,Y) coordinates are assigned to data by matching the geographical information in the data with geographical information in a table which already contains (X,Y) coordinates. For example, to display a database containing yearly maize production figures per district on a map, the data first must be geocoded by matching the names of the districts with those in the boundary map.

MapInfo supports raster images. Raster images are computerized pictures, for example a satellite image or an aerial photograph. These graphic images can serve as backgrounds for maps created or as a reference for displayed data. When a raster image is entered into MapInfo, its map coordinates must be specified so that MapInfo can display it properly. Once coordinates have been assigned, MapInfo creates a *'.tab file'* for the image. Then it can be opened like any other table in a Map window.

Basic mapping

The MapInfo screen displays a menu bar and several tool bars. Before the user selects each menu, he should be familiar with the functions each one performs.



The menu bar

The following can be accessed from the File menu:

- *New Table* allows one to create a new table. MapInfo tables have a graphic (mappable) component and a data (browsable) component. The New Table command allows setting up of these components.

- *Open Table* allows of the opening of a MapInfo table, dBase DBF file, delimited ASCII file, Lotus 1-2-3 spreadsheet, Microsoft Excel spreadsheet or raster image.
- *Open ODBC Table* enables one to download a table from a remote database. This table is known as a linked table. MapInfo supports access to the following databases: Microsoft ACCESS 2.0, DB2/2, INFORMIX 5, INGRES 6.4/04, ORACLE 7, GUPTA SQLBase, SQL SERVER, and SYBASE 10.x. However, other ODBC databases may also be accessable.
- *Open Workspace* allows the opening of a workspace that has been previously saved. A workspace stores a list of open tables, windows, and window positions.
- *Close Table* allows one to close tables, including query tables.
- *Close All* allows one to close all open tables and all layout windows.
- *Save Table* allows the saving of changes made to a table.
- *Save Copy As* allows one to create a new table from scratch or create a new table by saving an existing table under a new name.
- *Save Workspace* allows one to save information about the tables and windows used in the current session. This work setup can be returned to at will. This work setup is called a workspace.
- *Save Window As* allows one to capture the active window and save it as a bitmap (.BMP) or a Windows metafile (.WMF). Then the exported file can be used with other applications.
- *Revert Table* allows one to revert to a previous version of a table when temporary changes have be made.
- *Run MapBasic program* allows one to run a MapBasic program.
- *Page Setup* changes margins, paper source, paper size, and page orientation for the entire printout, the current section, or for selected sections of the printout.
- *Print* allows the printing of the contents of a, browser, redistrict, map, graph or layout window.

The following can be accessed from the Edit menu:

- *Undo* allows one to undo the last edit operation.
- *Cut* allows one to cut selected text and objects and move them onto the clipboard.
- *Copy* allows one to copy the selected text and/or graphic information and place it on the clipboard.
- *Paste* allows the copying of the contents of the clipboard into the table or window being edited (pasting text into a query table is not permitted).
- *Clear* allows one to delete selected text or objects.
- *Clear Map Objects Only* allows one to remove graphic objects from a table.
- Use *Reshape* to edit regions, polylines, lines, and points by moving, adding, and deleting nodes that define line segments. Selected nodes can also be copied and pasted selected nodes to create new points, lines, and polylines.
- *New Row* allows one to add a blank record at the bottom of the active browser.
- *Get Info* allows one to display the Object Attribute dialog for a selected object (editable or read only) in a Map or Layout instead of double clicking on the object. Use this dialog to specify geographic attributes for an object.

The following can be accessed from the Object menu:

- Use *Set Target* to prepare a selected object to accept subsequent editing commands (Combine, Erase, Erase Outside, Split and Overlay Nodes).
- Use *Clear Target* to clear as a target any object that was previously set for object editing.
- The *Combine* command lets one combine separate map objects into a single object. The Combine command also performs data aggregation, so that the new object's data columns contain sums or averages of the values from the original objects.
- *Split* is an editing command that allows one to cut up map objects into smaller parts using the currently selected object as the cutter.
- *Erase* is an editing command that allows one to remove a portion of a map object using the currently selected object as the eraser. The portion of the target object that is overlapped by the erasing object is removed.
- *Erase Outside* is an editing command that allows removal of a portion of a map object using the currently selected object as the eraser. The portion of the target object that is not overlapped by the erasing object is removed.
- The *Overlay Nodes* command adds nodes to the target objects at all points where the target objects intersect the currently selected objects.
- *Buffer* allows one to create a buffer polygon around a selected object or objects.
- *Smooth* allows the smoothening of a polyline, making it into a continuous curve.
- *Unsmooth* allows one to change a smoothed polyline back to its original state. While it is not necessary, in some cases it might be easier to reshape a smoothed polyline by returning it to an unsmoothed state. Unsmooth allows do this.
- *Convert to Regions* changes each of the selected objects into a region object.
- *Convert to Polylines* changes each of the selected objects into a polyline object.

The following can be accessed from the Query menu:

- *Select* allows one to query the database, select records and objects from a table according to certain criteria and create a results table that can be viewed as a map, a browser, or graph.
- *SQL* Select is a multi-purpose query tool. Use SQL Select to perform any or all of the following tasks:
    - Filter the data, so that only the rows and columns of interest are displayed.
    - Perform relational joins combining two or more tables into one results table.
    - Create derived columns (columns that calculate new values based on the contents of the existing columns).
    - Sort the data by numeric and/or alphabetic criteria.
    - Subtotal the data, so that only a listing of subtotals instead of the entire table is displayed.
- *Unselect All* allows one to unselect the currently selected objects in a map, layout or rows in a browser.
- *Find* allows one to locate individual objects or addresses. When an object is located, it is marked with a symbol.
- *Find Selection* allows one to automatically find and display a selection in all windows.
- *Calculate Statistics* allows the performance of statistical calculations for a column in a table or query/selection. These statistics can then be used in other applications.

The following can be accessed from the Table menu:

- Use *Update Column* to assign values to a column, add a new (temporary) column using data from another table, move values between columns and enter graphics information into columns for descriptive data.
- The *Geocode* command allows one to create a pin map, that is, assign point objects to rows in a table. Data in the record (i.e. street address, ZIP code, county, state) are used to match against a map (i.e. street map, ZIP code map, state map) to determine where the point for that record should go.
- *Create Points* allows one to create a pin map, that is, create point objects for a database that has X and Y coordinates. This function is used most often to create point objects for pointfiles imported from MapInfo for DOS. These points can then be displayed on a map.
- Use *Combine Using Column* to modify the geographical data and create one map object for each group.
- *Import* allows one to import vector (but not raster) graphics files.
- *Export* allows one to export tables to other formats. Graphics and tabular data can also be exported to MapInfo format (MIF) and AutoCAD DXF files. Only tabular data can be exported to delimited ASCII and dBASE DBF format.
- Choosing the table *Maintenance* leads to a submenu which contains options for:
    - The *Table Structure* command accesses the Modify or View Table Structure dialog. The Modify Table Structure dialog displays for editable tables and allows one to change the structure of a table (add, remove, rename, or reorder fields, add and remove indices). The View Table Structure dialog displays for read-only tables and is only for viewing the table structure.
    - *Delete Table* allows one remove a table from disk permanently. MapInfo tables consist of component files. All the component files of a table will be deleted.
    - *Rename Table* allows the renaming of a table and its component files.
    - *Pack* allows one to compress tables so that they use less disk space and to eliminate records that have been marked as deleted.
    - The *Make ODBC Table Mappable* command makes a table linked to a remote database mappable. Any MapInfo table may be displayed in a browser, but only a mappable table may have graphical objects attached. Only mappable tables may be displayed in Map windows.
    - The *Change ODBC Table Symbol* command allows one to change the symbol attributes for the point objects in a mappable ODBC table.
    - *Unlink ODBC Table* unlinks a table which was downloaded from a remote database and linked to a MapInfo table with the Open ODBC Table command.
    - *Refresh ODBC Table* enables one to refresh a MapInfo linked table with the most recent data residing on the remote database for that linked table.
- Choosing the table *Raster* leads to a submenu which contains options for:
    - Use *Adjust Image Styles* command to adjust the contrast or brightness of a raster image, or to display a color raster image in gray-scale mode.
    - Use *Modify Image Registration* to access the Register Raster Image dialog. This dialog allows one to prepare a raster image for use with MapInfo.
    - Use *Select Control Point* from Map to add control points to a raster image by clicking on a Map window.

The following can be accessed from the Option menu:

- Use *Line Style* to set the line type, thickness and color of line objects (lines, arcs and polylines). The type, thickness and color of objects being editing can also be changed.
- Use *Region Style* to access the Region Style dialog box to specify the color, pattern and outline of closed objects. This also sets the default style which is used when creating new objects.
- Use *Symbol Style* to specify symbol attributes (symbol type, size, color, rotation angle) for new or selected symbols.
- Use *Text Style* to choose a font and font settings for the text.
- Use *Toolbars* to access the Toolbar Options dialog. This dialog allows one to display or hide the Main Toolbar, the Drawing Toolbar, the Tools Toolbar, and the Standard Toolbar. The Toolbars can also be displayed as dockable Toolbars.
- *Show/Hide Legend Window* allows one to display or hide the legends associated with maps or graphs.
- *Show/Hide Statistics Window* allows one to show or hide the statistics window.
- *Show/Hide MapBasic Window* allows one to display or hide the MapBasic window.
- Use *Show/Hide Status Bar* to display or hide the status bar which is located at the bottom of the screen. The status bar shows messages to help one use MapInfo. The display includes map layer editing status, zoom status, and browser record status.
- Use the *Custom Colors command* to access a palette dialog containing a range of colors that can be used or customized. The customized color can also be saved and then automatically added to the palette in place of the original color. This color palette is accessible from the dialogs displayed when working with line, region, symbol, and text objects.
- Use *Preferences* to specify what group of settings, or preferences to view or change. MapInfos preferences consist of system, map window, startup, address matching, and directories. MapInfo retains these preferences from session to session.

The following can be accessed from the Windows menu:

- The *New Browser* window option allows one to view and work with textual data in table form.
- *New Map Window* allows one to display a table as a map.
- Use *New Graph Window* to display a table as a graph.
- *New Layout Window* allows one to arrange and annotate the contents of one or several windows for printing.
- *New Redistrict* creates a special table, called Districts, and displays the table in a Browser window. The Districts Browser, used in conjunction with a Map window, lets one perform redistricting. Map objects are assigned to a district by selecting the objects. As the objects are selected, MapInfo automatically calculates the net values for each district, and displays the values in the Districts Browser.
- Use *Redraw Window* to redraw the active window. This is useful when the <ESCAPE> key has been pressed to interrupt the window drawing. When a new table is added as the top layer of the map, only the affected portions of the map are redrawn. When switching between open windows, the map is preserved, so the entire window does not have to be redrawn.
- Use *Tile Windows* to arrange windows side by side so that all windows are visible.

- *Cascade Windows* allows one to overlap windows. Only the contents of the topmost window are visible and the titles of the other windows that are being worked with.
- *Arrange Icons* orders the icons of the minimized windows so they are more accessible.

The following can be accessed from the Map menu:

- Use the *Layer Control* command to access the Layer Control dialog. Use this dialog to:
  - Change display of map layers in the active window.
  - Determine which layers are displayed, editable, selectable, zoom layered.
  - Change the order of map layers.
  - Add or remove one or more layers from the active map.
  - Control labels.
  - Alter a thematic map.
- *Create Thematic Map* allows one to analyze data values associated with the map. The map objects can be shaded according to the data values (ranged, individual value), or thematic objects are created to display the data values. Single-variable thematic maps can be created (ranged, individual value, dot density, graduated symbols) as can multi-variable thematic maps such as pie and bar charts.
- Use *Modify Thematic Map* to modify a thematic map.
- *Change View* allows one to access the Change View dialog. Use this dialog to specify settings for Map window width (zoom), map scale, map resizing, and centering the map. One can also choose to display zoom, map scale, or cursor location in the Status Bar.
- *Previous View* allows one to return to a map or layouts in the immediately preceding view.
- *View Entire layer* allows one to zoom and display an entire layer or all layers in a map, get a map positioned in the Map window, or for reorientation.
- The *Clear Custom Labels* command allows one to clear all custom labels.
- *Save Cosmetic Objects* allows one to save the objects in the Cosmetic layer to a table.
- The *Clear Cosmetic Layer* command allows one to clear all objects (graphics and text) from the cosmetic layer.
- *Set Clip Region* redraws the map displaying only the clipped region.
- *Clip* a designated portion of a map for use in printing and presentations.
- *Digitizer Setup* allows one to configure MapInfo for digitizing.
- *Options* allows one to specify coordinate units, distance units, and area units for the map. New units appear in places such as the Ruler Tool window, on the status bar (if the map has been set up to show coordinates), and in dialog boxes that display area measurements, such as Region Object.

The following can be accessed from the Help menu:

- *Help Topics* enables one to display a dialog that contains a Contents tab, an Index tab, and a Find tab. The contents tab enables one to browse through topics by category. The Index tab enables one to see a list of index entries. The Find tab enables one to search for words or phrases that may be contained in a Help topic.
- *MapInfo Forum on the Microsoft Network* enables one to connect to the MapInfo on–line forum where one can learn about new products and services. The forum contains

information on MapInfo products, including downloadable demos. Additionally, there is a BBS message board.

- *MapInfo on the World Wide Web* enables you to access the MapInfo WWW home page.
- *About MapInfo* displays a dialog that tells which release of MapInfo is being used.

The toolbars

MapInfo contains several toolbars. The most important ones are the *Standard Toolbar*, *Main Toolbar* and *Drawing Toolbar*.

The standard toolbar

The standard toolbar contains tools for commonly performed menu functions from the file, edit and window menus, such as open and save a table, open new window, cut, copy and paste etc.

The main toolbar

- The *Select* button allows one to access the Select tool. Use the Select tool to select one or more objects or records for analysis. The Select tool can also be used to edit a map, layout or browser.
- The *Marquee Select* button accesses the Marquee Select tool. Use the Marquee Select tool to search for and choose objects within a given rectangle.
- The *Radius Select* button allows one to access the Radius Select tool. Use this tool to select all of the objects within a certain radius.
- The *Boundary Select* button accesses the Boundary Select tool. Use the Boundary Select Tool to search for and choose all the objects within a given region, such as a state or county boundary, a police patrol district, a sales territory, and so forth.
- The *Zoom-in* button allows one to access the Zoom-in tool. Use the Zoom-in tool to get a closer area view of a map or a layout.
- The *Zoom-out* button allows one to access the Zoom-out tool. Use the Zoom-out tool to get a wider area view of a map or a layout.
- The *Change View* button accesses the Change View dialog. Use this dialog to specify settings for Map window width, map scale, map resizing, and centering of the map.
- The *Grabber* button allows one to access the Grabber tool. Use the Grabber tool to reposition a map or layout within its window.
- The *Info* button allows one to access the Info tool. Use the Info tool to select a location on the map, including multiple overlapping objects, and display a list of all objects at that location. An object from the list can then be selected and the tabular data for that object viewed.
- The *Label* button accesses the Label tool. Use the Label tool to label objects with information from the related objects database.
- The *Drag and Drop* button enables one to drag an entire MapInfo map window and drop it into an OLE container application, such as Microsoft Word or Microsoft Excel. A map window can also be dragged within MapInfo. Dragging a map within MapInfo provides the same effects as Edit > Copy Map Window followed by Edit > Paste or Edit > Paste Special in an appropriate application.

- The *Layer Control* button accesses the Layer Control dialog. This dialog allows one to specify how the various tables in a Map window are layered and displayed.
- The *Ruler* button accesses the Ruler tool. Use the Ruler tool to determine the distance between two points.
- Use the *Legend* button to display the legend associated with a map or graph.
- Use the *Statistics* button to display the Statistics window. The Statistics window tallies the sum and average of all numeric fields for the currently chosen objects/records. The number of records chosen is also displayed. As the selection changes, the data are re-tallied, and the statistics window is updated automatically.
- Use the *Assign Selected Objects* button to permanently assign all selected map objects to the target district.
- Use the *Set Target District from Map* button to make the selected object's district the new target district.

The drawing toolbar

- The *Symbol* button accesses the Symbol tool. Use the Symbol tool to place point symbols ("push pins") on the map.
- The *Line* button allows one to access the Line tool. Use the Line tool to draw straight lines.
- The *Polyline* button accesses the Polyline tool. Use the Polyline tool to draw polylines (a connected sequence of lines that are not closed).
- The *Arc* button allows one to access the Arc tool. Use the Arc tool to draw an arc the size and shape of one quarter of an ellipse. Once an arc has been created, it can be reshaped to the desired size.
- The *Polygon* button accesses the Polygon tool. Use the Polygon tool to draw polygons one side at a time.
- The *Ellipse* button allows one to access the Ellipse tool. Use the Ellipse tool to create elliptical and round objects.
- The *Rectangle* button accesses the Rectangle tool. Use the Rectangle tool to draw rectangles and squares in an editable map or layout.
- The *Rounded Rectangle* button allows one to access the Rounded Rectangle tool. Use the Rounded Rectangle tool to draw rounded rectangles and squares.
- The *Text* button allows one to annotate maps and layouts.
- The *Frame* button accesses the Frame tool. Use the Frame tool to create frames in a layout. Each frame can display a map, graph, Browser, map legend, graph legend, Info window, statistics window, message window, text object, or it can be an empty frame.
- The *Reshape* button toggles in and out of Reshape mode. Use reshape to edit regions, polylines, lines, and points by moving, adding, and deleting nodes that define line segments. Selected nodes can also be copied and pasted to create new polylines.
- The *Add Node* button accesses the Add Node tool. Use the Add Node to add a node to regions, polylines, and arcs.
- Use the *Symbol Style* button to access the Symbol Style dialog box. The Symbol Style dialog box allows one to display symbols and to specify attributes for symbols. The attributes that can be specified are size, color, and symbol. The attributes of existing symbols can be changed and attributes for new point objects can be specified before they are created. The point objects must reside, or be created, in an editable layer.

- Use the *Line Style* button to access the Line Style dialog box. The Line Style dialog allows one to set the line type, thickness and color of line objects (lines, arcs and polylines). The type, thickness and color of line objects being edited can also be changed.
- Use the *Region Style* button to access the Region Style dialog. The Region Style dialog allows one to specify the color, pattern and border line style of closed objects. The color and pattern of objects being edited can also be changed.
- Use the *Text Style* button to access the Text Style dialog. The Text Style dialog allows one to choose a font and font settings for the text.

Starting MapInfo

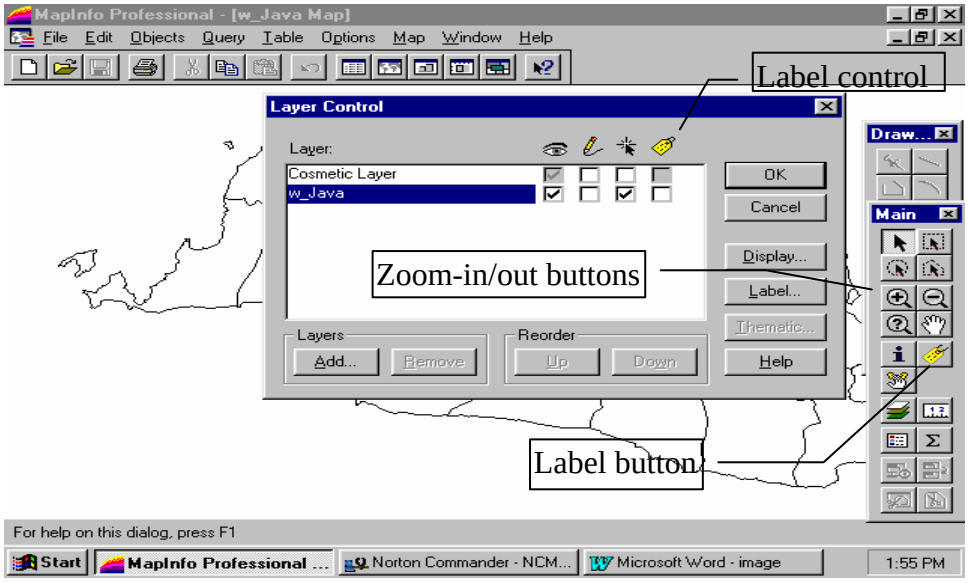
Everything in MapInfo starts by opening one or more table(s). First a Quick Start Dialog appears. From here one can return to a previous session or use the last workspace. Because this is the reader's first look at MapInfo, choose the *Open a Table* option to begin or cancel and select *Open table* under the *File menu*. The Open Table dialog is displayed. Choose the appropriate table to work with. Note that although a MapInfo table consists of two or more component files (.tab, .dat, .map, etc.), only the *.tab* file appears in the File Name box of the Open Table dialog. It is the only component file that has to be opened.

When a table is opened, all items can be displayed at once. To do this, go to the *Map menu*, select *View Entire Layer* and select the appropriate table. Click OK. A table can also be maximized and minimized.

Exercise I - Open the table W_Java.tab. Display all selected items at once. Maximize the map.

Labeling objects

By default map objects are not labeled. Objects can be automatically labeled by checking the box for the layer to be labeled in the layer control box which is situated under the menu (if the right-hand button on the mouse is clicked, a short-cut menu will appear). Upon return to the map window, the map will be labeled automatically. If objects are to be labeled individually, use the Label button from the main tool bar. When the label button is clicked, the cursor becomes a cross hair. Click wherever a label is required.



Exercise II - Label the districts on the map of West Java by individually and by automatic selection.

Zoom layering

The map can be viewed at different zoom levels. With the *Zoom-in* button on the main tool bar, one can get a closer aerial view of the map. With the *Zoom-out* button one gets a wider aerial view. The current zoom level is displayed in the lower right-hand corner. If a map is to be displayed only at a certain zoom level, use 'Zoom Layering' to control the display of a map layer, so that it only shows up at the predetermined zoom level. To set zoom layering, select a *Layer* in the *Layer Control Dialog*, and click the *Display* button. The Display Dialog appears. In the Zoom Layering section, check the box in front of **'***Display within Zoom Range'*. Type in a minimum and/or a maximum zoom distance for display of the layer.

Exercise III - Open the City.tab table. Label the district capitals. Unlabel the district layer map. Specify a maximum zoom level of 300 miles.

Entering data

Direct entry

Data can be entered directly in MapInfo or can be imported from other programs, such as a database or spreadsheet program. To enter the data directly in MapInfo, go to the *Table menu* and select *Maintenance*. In the following submenu select *Table Structure*. In the *Modify Table Structure* dialog that appears, fields can be now added and remove. In the Field Information section one can name the additional field, determine the type (character, integer, decimal, date, etc.) and width of column. Click *OK*.

To enter the data go to the *Window menu* and select *New Browser Window*. The new table with the added fields displays. One can start to entry the data into the data field. After

finishing updating, choose *Windows* again and select *New Map Window*. Each map object has now data attached to it. To view this data click the *Info button* [**i**] from the main tool bar. Next click on a map object. The Info Tool window displays the data. If the table contains many fields, resize the window or scroll down to see all the data. The data can also be visualized by using a graph window. Go to *Window menu* and select *New Graph Window*. Select the appropriate table and column. The data appear as a bar graph in a graph window. On the *menu bar*, the graph menu will be displayed. Choose this menu to customize the graph.

Exercise IV - The following data concern maize production for West Java at the district level for the year 1995. Enter the data into a newly created field.

District	'000 mt	District	'000 mt	District	'000 mt
Pandeglang	7,699	Tasikmalaya	27,552	Subang	3,670
Lebak	8,703	Ciamis	24,643	Purwakarta	10,186
Bogor	5,556	Kuningan	15,173	Karawang	1,013
Sukabumi	17,649	Cirebon	1,736	Bekasi	1,712
Cianjur	21,600	Majalengka	22,403	Tanggerang	3,037
Bandung	26,239	Sumedang	26,869	Serang	13,720
Garut	91,488	Indramayu	1,531		

Enter the data in MapInfo by modifying the district table. Name the field 'Pr_Maize95'. Choose the appropriate field type and column width. *Open Browser window* for the entry of data and open a *New Map window*. Use the [**i**] button to view the data attached to the districts.

Exercise V - Create a graph to visually compare the maize production data. Customize the graph using the graph menu (change graph title, graph type, etc.).

Import data from other programs

MapInfo allows one to use data that were created in other file formats. When data are brought into Maplnfo for the first time, the format must be specified. For example, if the data are in delimited ASCII format, choose delimited ASCII from the *'Files of Type'* drop-down list in the *Open Table* dialog.

Other choices include:

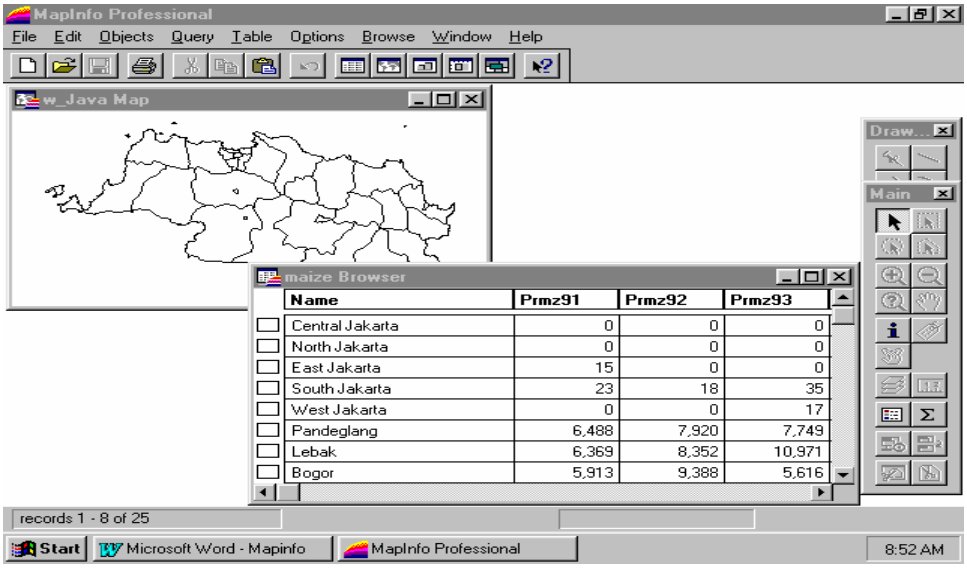
- dBASE DBF
- Lotus 1-2-3
- Microsoft Excel
- Raster Image

Once a particular file format is chosen, MapInfo will only list files that have the appropriate extension. For example, if dBASE DBF is selected from the Files of Type drop-down list, MapInfo will only list files that are in dBASE format. Once the file is opened Maplnfo creates a table structure for that data. When the table is opened in future work sessions, Maplnfo will treat these files as if they were in MapInfo's native format (i.e. *.tab*) If one attempts to open the file again with its original file format, MapInfo prompts with the message:

Table definition already exists. Overwrite it?

This message displays because Maplnfo has already created a table for that file. Press Cancel and open the associated *.tab* file.

Exercise VI - Open the database file maize.dbf (file type: dBASE DBF). This file contains data on maize production by district from the year 1991 - 1995. Click *OK* for the File Character Set *'Windows US & W. Europe ("ANSI")'*. Close the District Browser window and the Graph window. Now only the W_Java Map and Maize Browser window display on the screen.

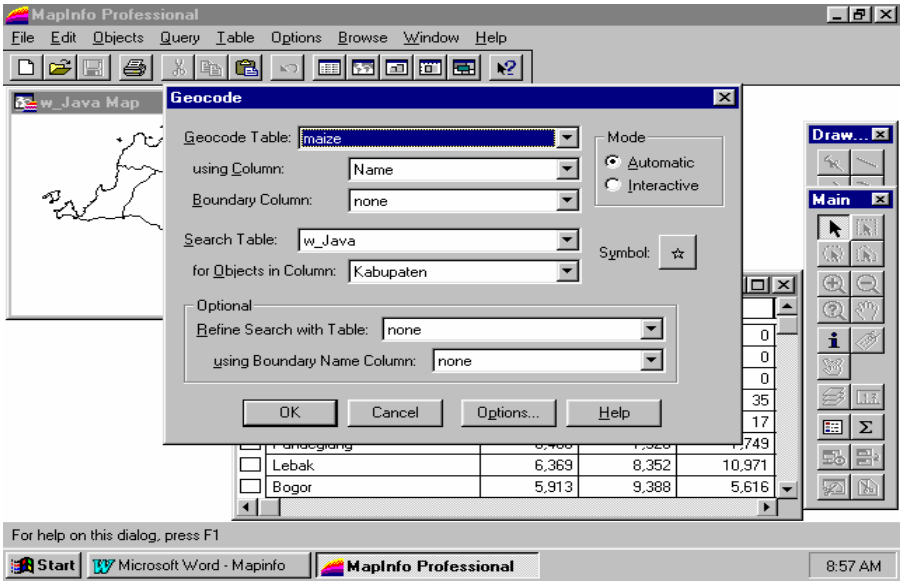


The imported data do not yet contain X and Y coordinates. Consequently, MapInfo does not know where to place the data on the map; for it to do so, the data must be geocoded.

Geocoding

To assign X and Y coordinates to the records in the table, choose *Geocode* from the *Table menu*. MapInfo displays the Geocode dialog. The following information needs to be entered.

- The name of the table to which X and Y coordinates are to be assigned (Geocode Table: …..)
- The column in the table that contains the location information that will be used for matching (using Column: …..)
- The name of the search table that already contains X and Y coordinates (Search Table: …..)
- The column in the search table with the geographical information to which the location information of the data table is matched.



In the dialog one can also specify the *mode* of searching: automatically or interactively. When geocoding a table automatically, MapInfo geocodes exact matches only and ignores all other records. It is the faster method, since MapInfo requires no user interaction once the geocoding process begins. When geocoding a table interactively, MapInfo pauses when it fails to match a record and asks the user to select from a list of close matches. The normal procedure is to first search automatically and then select the ungeocoded records interactively. When MapInfo has finished the geocoding process, the results are displayed in a summary dialog. Click *OK*.

Exercise VII - Geocode the table maize.tab. Search first in the automatic mode. Repeat the process in the interactive mode for records that are not geocoded. If you can't find a match, choose ignore.

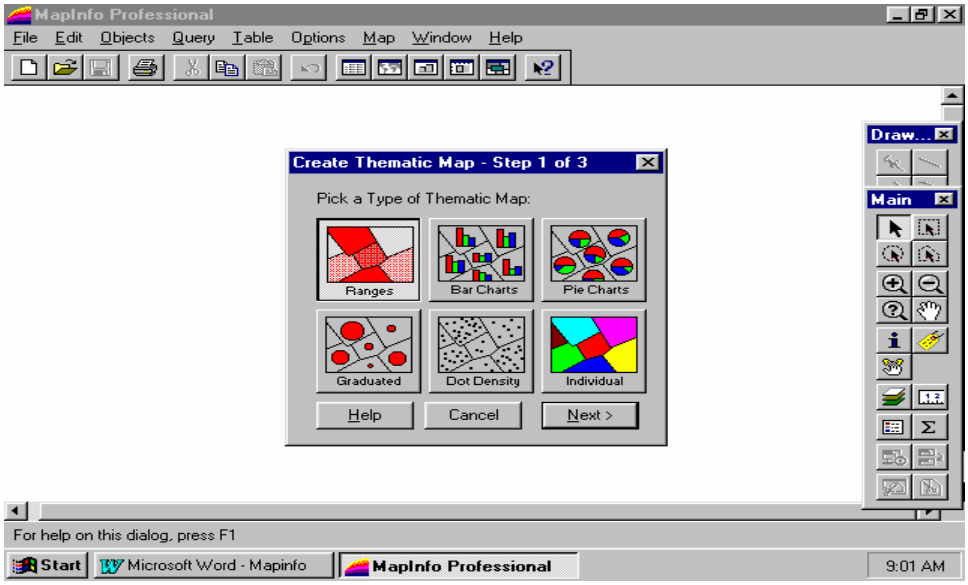
Adding multiple layers

After the geocoding process is done, the table can be added as a separate layer. In other words, it is stacked on top of the other map layers, just like transparencies on an overhead projector. The data will be represented by a symbol chosen in the geocode dialog. In this case, it is a ∃-symbol. To add multiple layers go to the *Map menu* and select *Layer Control*. In the *Layer Control* dialog click the *Add button*. The *Add Layer* dialog displays, showing a list of all tables currently open. Select the *table* to be added and click the *Add button*. Next click *OK*. The data table is now added as a separate layer on top of the other map layers.

Exercise VIII - Add the table containing the maize production data as a separate layer to the map. Use the [**i**] button to look at the data attached to the symbol. Click again on 'maize' in the Info Tool dialog. The yearly maize production in the district selected will be displayed.

Creating a thematic map

Using a ∃-symbol to represent the data in the table is not a very informative way of display. A much more informative way of presenting data is to give it a graphic form so one can actually see it on the map. Thematic mapping is the process of colouring in the map according to a particular theme, for example, a crop map or a soil map, etc. By colouring in the map it is easy to visualize and consequently to analyze the data. Colouring in not only refers to making use of colours, but also to different patterns, symbols or charts. With MapInfo one can create six different types of thematic maps: range of value, graduated symbols, dot density, individual values, and bar and pie chart:



- *Range of Values Map***.** In this kind of map, data are grouped into ranges and each record's object is assigned the colour, symbol or line for its corresponding range.
- *Graduated Symbol Map***.** Symbols are used to represent different values. The size of each symbol varies according to the data value.
- *Dot Density Map***.** In this map, data values are represented by dots. The total number of dots corresponds to the data value.
- *Individual Values Map***.** In these maps, each unique value gets its own colour or symbol.
- *Bar Charts Map***.** Using a bar chart allows one to examine more than one variable per record at a time. A bar chart is used to display the data values.
- *Pie Charts Map***.** Just like with the bar charts, the user can examine more than one variable per record at a time. Instead of comparing the heights of the bars, one can now compare the wedges in a pie.

Before starting to create a thematic map, it is important to know what variable(s) to display i.e. the *thematic variable* and where to obtain the data. Data can either come from the same table as the base table on which the map is based or from a different table. In case of the latter, it is necessary to create a temporary column in the base table (see step 2). When creating

a thematic map, the thematic symbols, colourings, etc. are added to the map as a separate layer. In the layer control dialog, the thematic layers are displayed as: *<thematic type> with/by <variable-list>*. That is, the type of thematic map is noted first, followed by the list of variables used.

A specific kind of map is the *Bivariate Thematic Map*. In such a map, one map object, such as a symbol represents two different pieces of data. For example, a star can represent one variable, while the fill-in colour or the size represents another. When creating a bivariate thematic map, one basically creates two thematic maps with one layer covering the other.

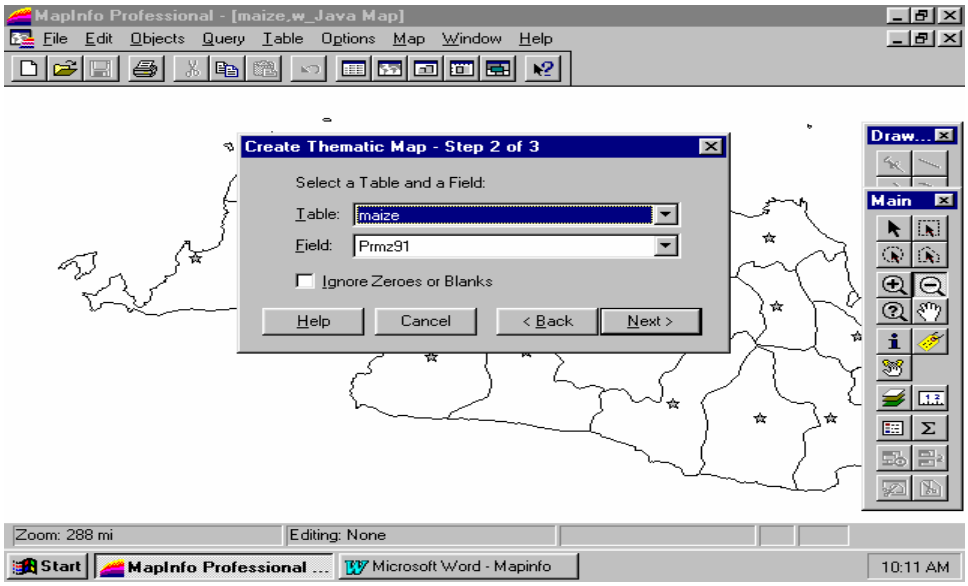
Thematic maps are created in three steps. In the first step select the type of thematic map. In step two involves the selecting of the thematic values and in the last step the thematic map is customized.

Step 1 - Type of thematic map

To open the Create Thematic Map dialog go to the *Map menu* and select *Create Thematic Map*. The first step dialog displays (see above). Choose one of the six types of thematic maps explained above. The Next button will bring one to step 2.

Step 2 - Selecting the thematic variable(s)

In the next dialog select the thematic variable(s) and the table where they are stored. If in step 1, a bar or pie chart was selected, now choose more than one thematic variable. In using the other types of thematic maps the user can only select one thematic variable. The step 2 dialog appears as follows.



At this stage it is important to realize on which table the thematic map is based and from which table the data are obtained. As mentioned above, data can either come from the same table on which base the thematic map is based or from a different table.

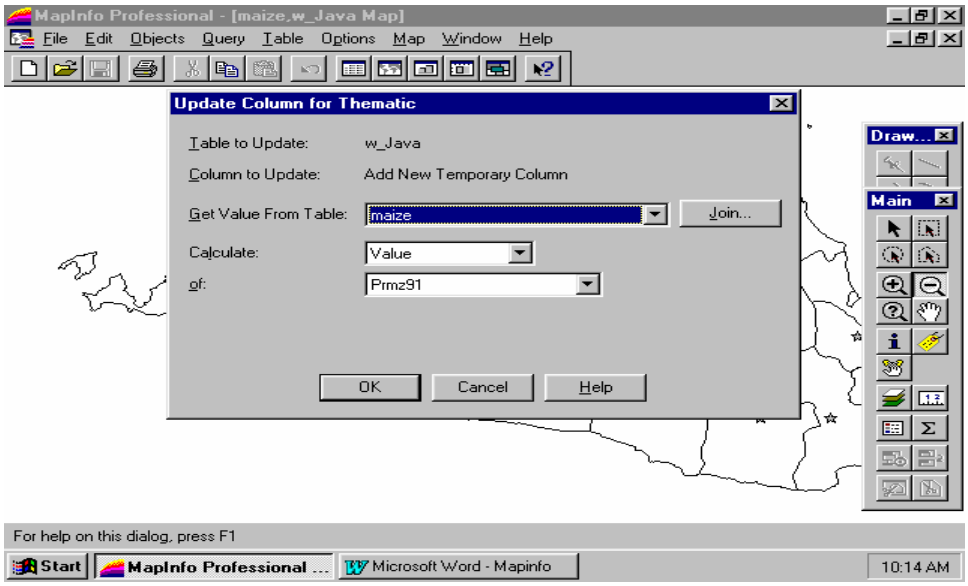
A) Using data from the same table

For example, to colour in the district map of West Java according to maize production in the year 1995, base the thematic map on the district table. In exercise IV the data needed were entered in the district table with field name Pr_Maize95. Simply proceed with choosing the district table from the table drop down list and select the field Pr_maize95, followed by pressing the Next button to proceed to step 3.

B) Using data from a different table

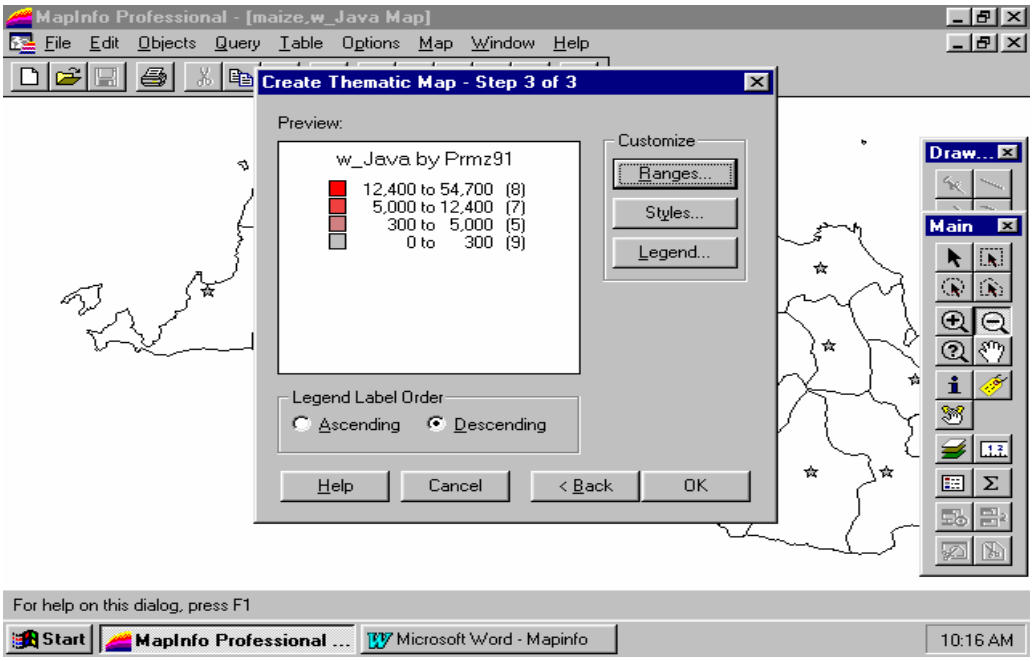
All the information to be used must be in the table on which the thematic map is based. The district table does not contain maize production figures for other years. They are stored in the table Maize.tab. Consequently, to create a thematic map for other years, update the district table with the data of that particular year by using *Update Column*. Update column is a temporary column in the base table storing the necessary data. Select the *District table* from the table drop-down list. In the Pick Field list box, choose *Join.* The Update Column for Thematic dialog displays.

In the *Get Value from Table* box select the maize.tab table. Data used for the Update Column can be a field directly taken from another table (value), or can be aggregated (sum, average, minimum, etc.). By default MapInfo selects the Value option in the *Calculate* box. In the *of* box choose the field i.e. the year data, for example, the maize production figures for the year 1991. Click *OK.*



Step 3 - Customizing the thematic map

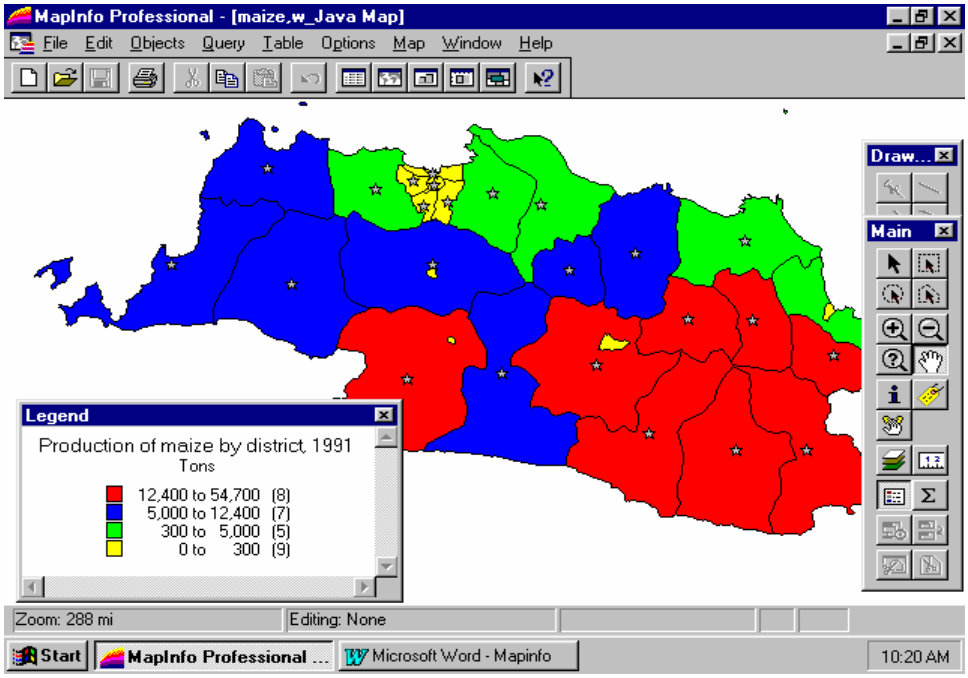
The last step enables customizing a thematic map, or creating a map based on the default settings. One can also preview the map legend before displaying the map, and change the legend's order.



One can change the legend label order. *Ascending* displays the data from lowest to highest value. *Descending* displays the data from highest to lowest value. In the customize box there are three buttons, Ranges, Styles and Legend. *Ranges* enables one to customize ranges on a ranged map. This option is only available for ranged maps. *Styles* enables one to customize style attributes such as colour and size. This option is available for ranged, pie, bar, and individual value maps. With *Legend* one can customize the legend, for instance changing the title and adding a subtitle. This option is available for all types of thematic maps.

Exercise IX - Create a thematic map showing the production of maize by district for the year 1991. Choose *Range* for the type of thematic map. Use the boundary district map as the base map. With *Update Column* get the data for the year 1991 from the Maize.tab table. Change the legend title to 'Production of Maize by District, 1991' and add as a subtitle '(Tons)'. Click the *Style* button if desired to change the colours and patterns.

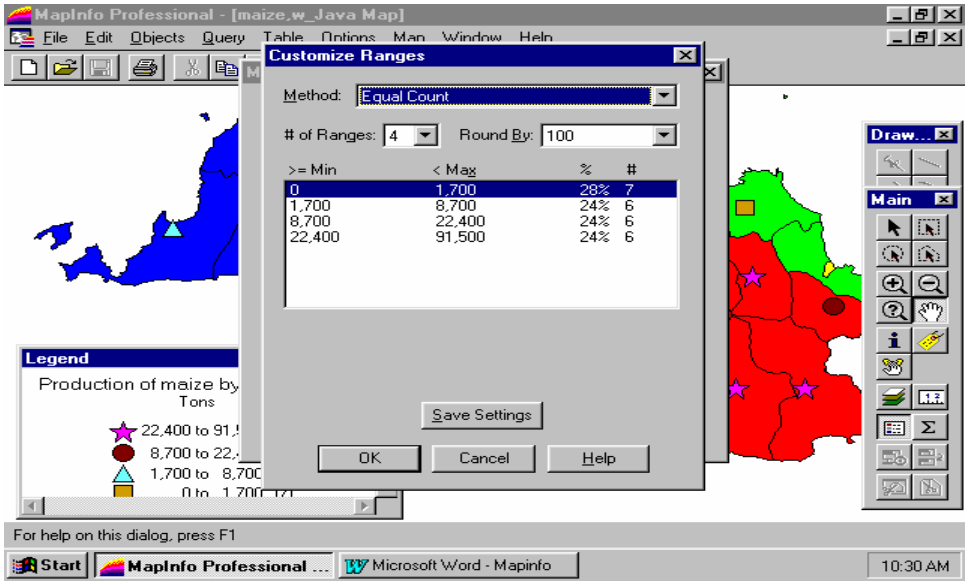
You have now created your first thematic map. The different patterns and colours refer to specific ranges.



Now to compare changes in the production of maize over time, for example, to compare the production of maize in the year 1991 with the production in 1995, either a bar / pie chart could be produced or another thematic map for the year 1995 could be created. As mentioned before, thematic maps form a separate layer of the map. So, to make a comparison one could put the thematic map for 1995 on top of the map for 1991. However, to do so one would have to choose another base map for the year 1995. If the same base map i.e. the district boundary map is used, the result is a similar map like the one for the year 1991. If this thematic map is put on top of the one for the year 1991, the latter will not be visible. Therefore, one must choose another table to base the map on, in this case the Maize.tab table. The data in this map are currently represented by the ∃-symbol. To create the thematic map for the year 1995, follow the same three steps explained above. The data for the year 1995 are already available in this table. Consequently, one does not have to use the *Update Column* and can directly choose the field in the step 2 dialog.

Exercise X - Create a thematic map showing the production of maize by district for the year 1995. Choose *Range* for the type of thematic map. Use the *maize.tab* table as the table on which to base the map. Change the legend title to 'Production of Maize by District, 1995' and add as a subtitle '(Tons)'. Click the **Style** button to change the symbol, so that each range is represented by a different symbol.

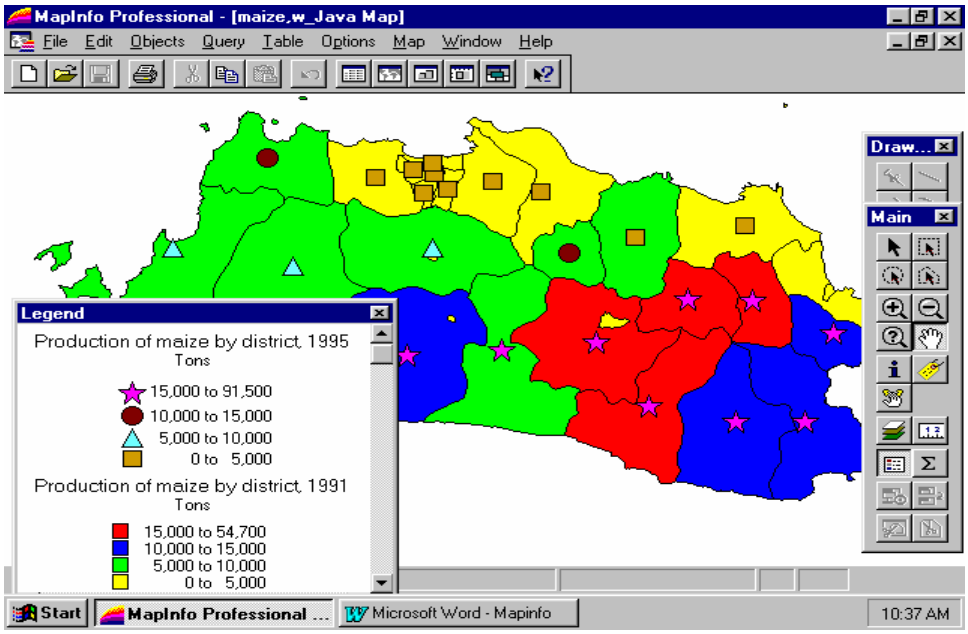
It is clear that the ranges for the year 1991 and 1995 are different. The ranges are by default based on 'equal count'. This means that MapInfo tries to divide the records in such a way that an equal number (as far as possible) fall within a particular range. In order to more easily compare the changes over the years, it would be ideal to use the same ranges for both years. To change the range go to the *Map menu* and choose *Modify Thematic Map*. From the drop-down list select the thematic layer to modify. Next click on the *Ranges button*.



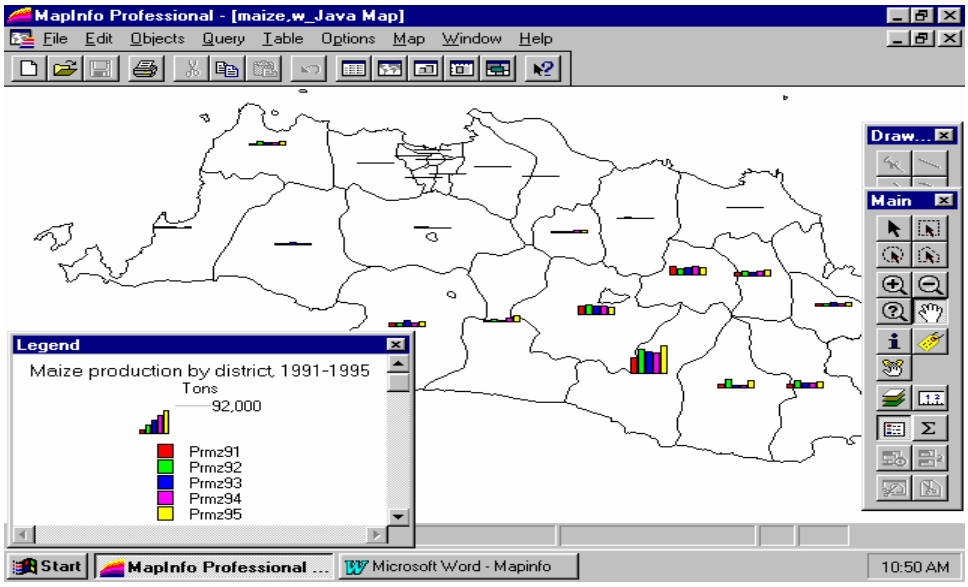
In the Method box select *Custom* from the drop-down list. Now customize the ranges. MapInfo does not accept an open-ended range. Simply choose a high enough upper limit and afterwards customize the legend.

Exercise XI - Change the ranges for both thematic maps. Use the following classes: 0 - 5,000, 5,000 - 10,000, 10,000 - 15,000, and 15,000 upwards.

In the resulting maps it is easy to see the shift that has taken place in the production of maize during the five years under consideration.



To compare the yearly difference in the level of maize production by district, it might be informative to use a bar chart with the height of the bars representing the level of maize production. An example of such a map is presented below.



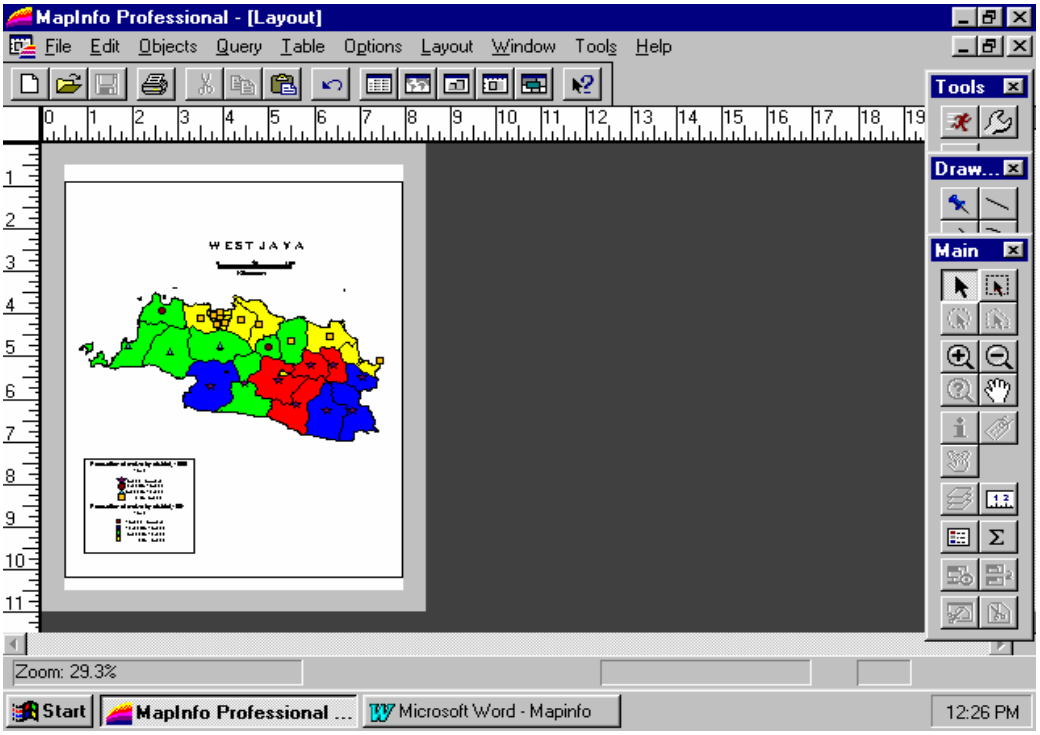
Exercise XII - Try making this bar chart map.

Saving and printing

To save the current work setup make use of *Workspaces*. Workspaces are used to keep a variety of tables and windows available so that one doesn't have to open and display them each time to work with them. However, *saving a workspace will not save changes made to tables or queries!* These should be explicitly saved by using Save as or Save commands in the File Menu. To save a workspace, go to the *File menu* and then select *Save Workspace*. MapInfo displays a dialog in which to name the workspace and set the directory to which it should be saved.

Exercise XIII - Save the thematic maps as a workspace. Choose the correct directory and name it 'my_map.wor'.

To prepare the map for printing, make use of a *Layout Window*. MapInfo's layout window is a page layout feature that allows one to combine a Map, Browser, and Graph Window on one page and prepare them for output. Any currently open window can be added to the layout and moved and resized. Also Titles and labels can be added.

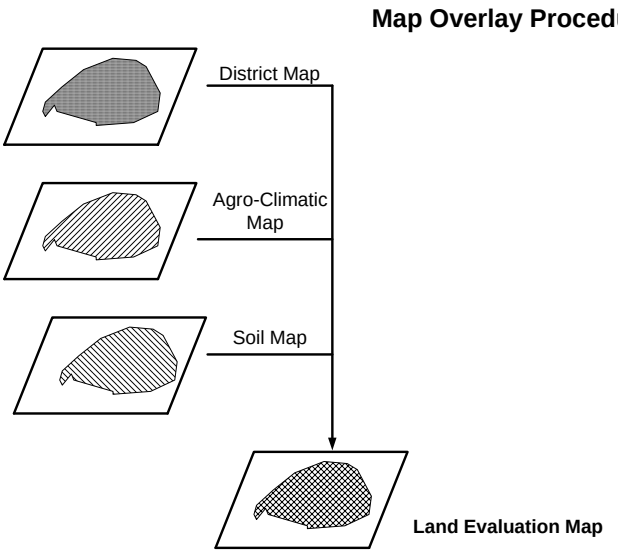


To open a layout window go to the *Window menu* and choose *New Layout Window*. A New Layout Window appears in which to specify which window to display in the layout. After the layout window has been opened, MapInfo places the Layout menu on the menu bar. The drop-down list gives different options to view the layout. When fully satisfied with your layout, go to the *File menu* and select *Print*.

Exercise XIV - Make a layout for the thematic map, showing maize production data for the years 1991 and 1995. Print the results.

Map overlay and area calculations

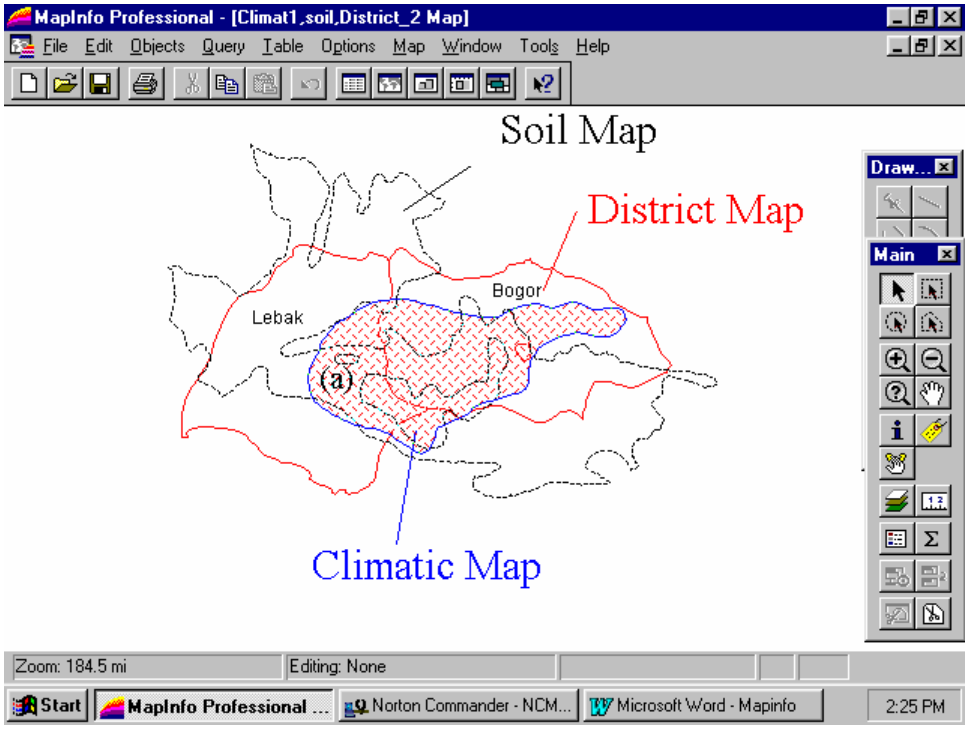
When doing land-use studies it is often necessary to classify land according to its unique characteristics, such as soil type and climatic conditions. Using a geographic information system, an overlay of different maps can be made to obtain a land evaluation map. A land evaluation map consists of land evaluation units (LEU) which all have the same characteristics. The figure below clarifies the process.



Here a district map has been combined with a soil and a climatic map. Each LEU on the land evaluation map is now characterized by the same type of soil and the same climatic conditions. This information can be important to establish an optimal use of the land. For example, certain crops grow better on a specific type of soil under specific climatic conditions. In this way a regional optimal development plan for the land-use can be formulated. To do so, information on the area of LEUs is needed. In this section, the procedure for area calculation in MapInfo is explained.

When combining a district boundary map with a soil and a climatic map, an enormous number of LEUs result. Therefore, climate and soil types are normally grouped together in more general categories. To simplify even more, the next exercises will include only two districts, Bogor and Lebak. So, instead of looking at the provincial map of West Java, only consider part of it. Except for the two districts mentioned all other districts have been deleted and this reduced map saved under the name District_2.tab. The same has been done for the soil and climatic map.

Exercise XV - Open the tables District_2, Soil and Climate.



The three tables are displayed in MapInfo as three layers on top of each other, consisting of a district boundary map, a soil map and a climatic map. Similar types of soil can be found within the dotted line and a different type outside. The shaded area of the climatic map represents a particular climatic condition. Outside that area other climatic conditions prevail.

To determine the total area of the land evaluation unit (A), an area in the district of Lebak, select and extract that area (the target) from the map. For this, use '*Splitting Object*'. 'Splitting object' allows one to divide a map object, like an area, into smaller objects by using another object as a cutter.

To be able to split a map object, the layer where that object can be found has to be *editable.* Editable means that one can make changes to it. Only one layer at a time can be editable. To make a layer editable go to the *Map menu* and choose *Layer Control*. In the Layer Control dialog, tick the check box of the layer to be edited in the check box column marked with a pencil. To split a map object proceed as follows:

- Select the *object(s)* to split in an editable layer. This is the target object.
- Go to the *Objects menu* and select *Set Target*. The object selected will now be marked.
- Select a *map object* from another layer to serve as a cutter object along which lines the target object will be cut.
- Go to the *Objects menu* and select *Split*. The Data Disaggregation dialog displays. Choose the appropriate disaggregation method for each field or in case of no data, tick the check box *No Data*. Click *OK*.
- Repeat the steps to split the object again.

After these steps, the target object will be split into smaller map objects. After splitting the target object, MapInfo deletes the original target object from the table and replaces it with the new objects created.

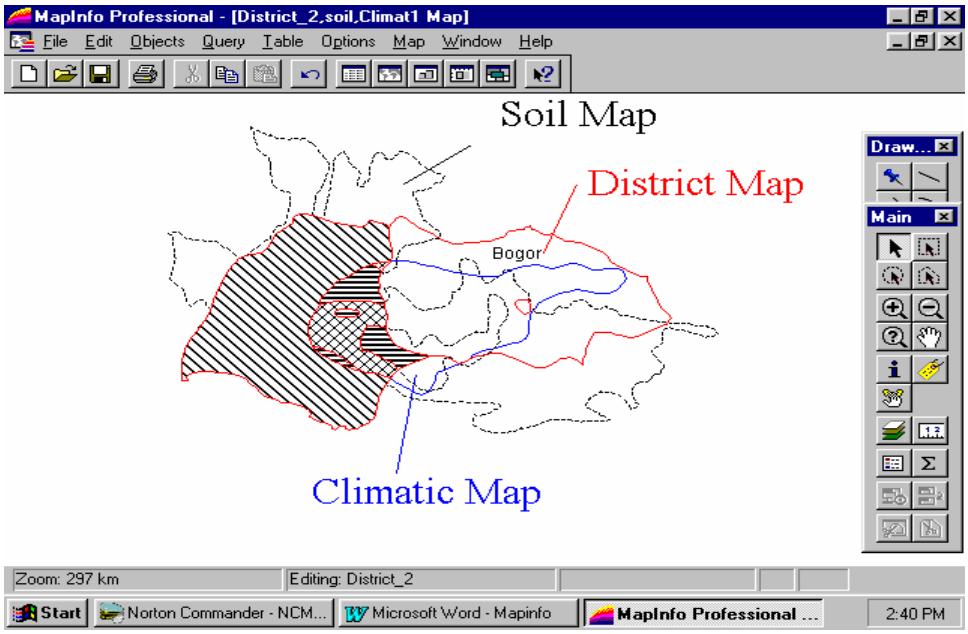
Exercise XVI - Spilt the district of Lebak into different LEUs. Make use of the following tips:

- Make layer District_2 editable.
- Select with the mouse the district Lebak.
- Make this district the target object.
- As a cutter object select the object on the climatic map.

The district is now split into two halves. Each half is characterized by a specific climate.

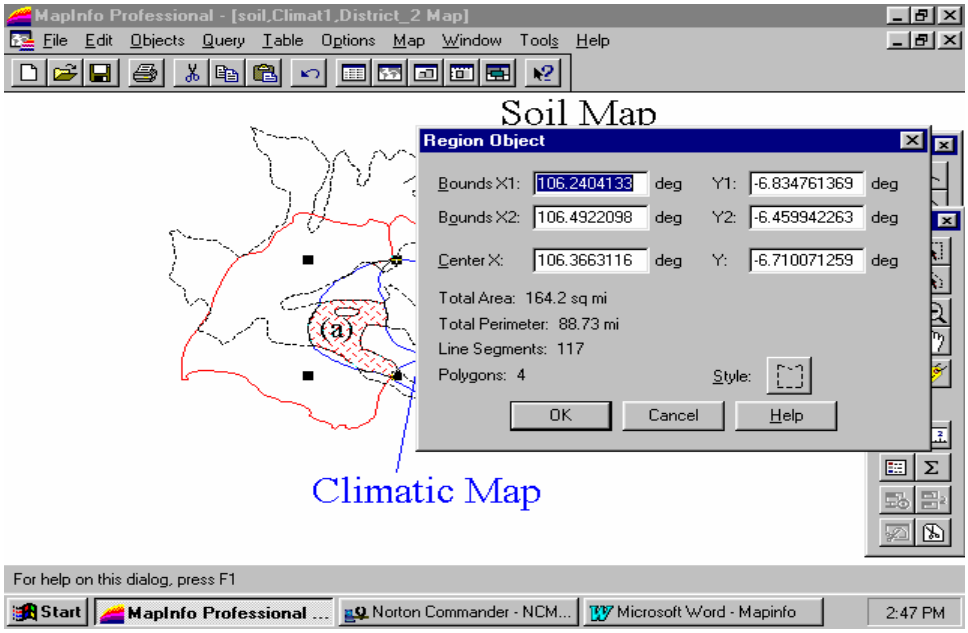
- Select the middle part of the district and make this the new target object.
- Split this target object again. This time using the object on the soil map as a cutter.
- Repeat the last step for the outer sections.

The district of Lebak is now split into three areas, each with its own unique characteristics.



To establish the total area of a LEU, simply double click on it with the mouse and a Region Object dialog will appear. The Region Object dialog displays information on the total area of the region, its perimeter, as well as the number of line segments and number of polygons making up the region.

The land evaluation unit (A), for example, has a total area of 164.2 square miles (425.4 square kilometers).



Other possibilities

MapInfo offers many other interesting features that can not be covered here. In this section, a few of these uncovered areas are mentioned.

Queries

MapInfo has two query tools: *Select and SQL Select*. With Select a table of information is queried and a subset of it may be selected. It can be used to highlight objects in a Map window or a Browser that meet a certain criteria. For example, in which district is the production of maize more than 100 mt? MapInfo stores the results of these kinds of questions in a *query table*. The records in the query table generated by the select command are the same records are the ones in the base table. except that now only a subset, which contains no new information, is displayed. With SQL Select one can create query tables containing information that was only implicit in the base table(s). One can for example create a query table with information on the population density while in the base table only the total population was given.

Drawing and editing maps

Drawing and editing tools are accessible from the Drawing Toolbar. With these tools one can draw and modify objects on a map. These tools can be used to customize the colours, fill patterns, line types, symbols, and text on the map. Furthermore powerful geographic analysis can be performed with them. For example, one can draw circles, polygons, and other bounded objects and can then search for records within these objects.

Redistricting

Redistricting is the process of grouping map objects into districts in order to perform aggregate calculations on the data. Map objects with a common filed are grouped together into districts to provide total values for the districts

Raster images

In MapInfo, raster images are used as display layers only. They are particularly useful as a backdrop for the MapInfo vector map layers. Raster images, however, cannot have any data attached to them.

Projections

Maps are created on a flat piece of paper or computer screens. The world, however, is not flat. The problem is: how to flatten the curved surface of the world when drawing a map? For this projection is used. A projection is a system that defines how to flatten objects. MapInfo allows the display of maps in different projections.

Digitizing

If digitized maps are not available for the area of interest, the map has to be created. This is done by using MapInfo's digitizing feature. By using a *digitizing table* one simply traces the outlines of a paper map with a *puck* and records the tracings as a vector image. This can then be displayed in MapInfo as a map layer.

Bibliography

MapInfo Corporation. MapInfo Professional User's Guide. 1992-1995. Troy, New York.
MapInfo Corporation. MapInfo Reference. 1992-1994. Troy, New York.

Sets: An Approach to Decision Making and Conceptualization

J.W.T Bottema and Mohammad A.T. Chowdhury[*]

Introduction

Sets and the algebra of sets are enjoying increasing popularity in business and the sciences, especially in applied statistics. One reason for this popularity is that an understanding of the basic concepts of sets and set algebra provides a form of language through which the business specialist or the planner can communicate important concepts and ideas to his associates. Specialists in electrical engineering, for example, use set algebra in the design of circuits. At the opposite extreme, specialists in written communications use sets in the analysis of statements and preparation of reports. Students of business administration use sets in the study of probability, statistics, programming, optimization, etc. Policy-makers use sets in spatial planning. This section provides an introduction to sets and set algebra along with selected applications. This is entirely based on Childress (1974), but many changes have been made to focus on database management and agriculture.

Sets defined

A *set* is defined as a collection or aggregate of objects. This can be classified into two categories such as structured (ordered pair) and non-structured. To illustrate this definition, the districts in a given state are a set of administrative units. Similarly, the provinces in a given country are a set of geo-political regions; the subscribers of a national newspaper are a set of individuals; participants in a given course are a set of students; the faculty members of a university are a set of professors. Moreover, the houses in an area are a set of settlements; likewise the farms in a given geography are a set of agricultural holdings.

From these examples, the student can see that the concept of a set is relatively straightforward. There are, however, certain requirements for the collection or aggregate of objects that constitute the set. These requirements are:

i. The collection or aggregate of objects must be well defined, i.e., we must be able to determine unequivocally whether or not any object belongs to the set.
ii. The objects of a set must be distinct; i.e., we must be able to distinguish between the objects and no object may appear twice.
iii. The order of the objects within the set must be immaterial, i.e., a, b, c is the same set as

c, b, a. However, in the case of structured sets ordering is of great importance.

On the basis of these requirements for a set, suppose we are asked to verify that the students in a technical college are a set of learners. To determine if the students are a set, we ask three questions. First, can we determine if a student is registered for the college? Second, is it

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

possible to distinguish between the students? Third, is the order of students in the college immaterial, i.e., are the college students the same set whether arranged alphabetically by their last name or by their identification number? If the answer to all three questions is yes, we conclude that the registered students in a given college are a set.

Example: Verify that the farms in a region satisfy the requirements for a set. The farms are farm 1,2,3 and so on, these are well-defined, distinct, and the order of the farms is immaterial. Therefore, the farms are a set.

Example: The set of districts is defined as the districts {1, 2, 3, 4, 5, 6, 7, 8, 9}. Determine which of the following districts are members of this set: districts {3, 14, 137, IV}.

District 3 is the only member of the set.

The above examples of sets are non-structured in nature. It should be cautioned, however, that in the case of probability the sequence or order of occurrence is important.

Example: A coin is tossed three times. Denoting a head by H and a tail by T, determine the set that represents all possible outcomes of the three tosses.

The possible outcomes are {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}, where HHH represents a head on the first toss, a head on the second toss, and a head on the third toss. HHT represents a head on the first toss, a head on the second toss, and a tail on the third toss, etc.

There are eight elements in this set and, corresponding to the requirements given for a set, the order of the elements is immaterial. Is each of the elements, however distinct? Does the element HHH differ from the element HTH? If, for instance, a value is added that a head would occur in the first toss of the coin and a tail would occur on the second toss, then the cumulative results will be affected by the very way the head and tail are ordered. Further discussion on ordered pairs is elaborated in the section on Cartesian Products.

Sets are the core business of information, whether in business, planning, or in economic appraisals or spatial analysis. It should be noted that the very existence of a set which satisfies all requirements could in fact be a hypothesis. Cartographic themes are sets, albeit somewhat complicated, rice growers are sets, and so on.

Specifying sets and membership in sets

It is customary to designate a set by a capital letter. For instance, the set of districts defined in the above example could be designated as *D*.

The objects that belong to the set are termed the *elements* of the set or *members* of the set. They can be designated by two methods: (1) the roster method, or (2) the descriptive method. The roster method involves listing within brackets all members of the set. The descriptive method (also called the defining property method) involves describing the membership in a manner such that one can determine if an object belongs in the set.

To illustrate the specification of set and membership, consider again the set of districts. The set of districts is designated by the capital letter *D*. The elements in the set (or alternatively the members of the set) are shown either by listing all the elements in the set within brackets or by describing within brackets the membership. If the roster method were used, the set of districts would appear as:

$$D = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

This is read as "*D* is equal to that set of elements 1, 2, 3, 4, 5, 6, 6, 7, 8, 9." If the descriptive method of specifying the set is used, the set would be:

$$D = \{x|x = 1, 2, 3, \ldots, 9\}$$

This is read as "*D* is equal to that set of elements *x* such that *x* equals 1, 2, 3, . . . , 9." In interpreting the symbolism used in set notation, it is useful to think of the left bracket as shorthand for "that set of elements" and the vertical line as shorthand for "such that." Commas are used to separate the elements, and the raised periods mean "continuing in the established pattern." The 9 in the above set is interpreted as the final number of the set. The right bracket designates set completion.

Example: The positive integers or "natural numbers" are the numbers 1, 2, 3, 4, 5, ... Show the set of natural numbers. Assume that the set of natural numbers is represented by *N*. We cannot, of course, list all of the members of the set. We therefore use the descriptive method of specifying set membership and write:

$$N = \{x|x = 1, 2, 3, 4, 5, \ldots\}$$

Example: Develop the set notation for the English alphabet. We can use either the roster method or the descriptive method of specifying set membership. Representing the set of letters in the English alphabet by *A*, the set is:

$$A = \{a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z\}$$

The descriptive method would conserve some space:

$$A = \{a, b, c, \ldots, y, z\}$$

Either method is acceptable. When one is using the descriptive method, however, it is important to remember that the description must be sufficient for one to determine the membership of the set. The elements must be well defined, distinct, and order must be immaterial.

Example: The possible convention sites for the Asia-Pacific Economic Cooperation (APEC) are Bogor, Bangkok, San Francisco and Vancouver. The set of convention sites is:

S= {Bogor, Bangkok, San Francisco, Vancouver}

Example: The rice growers in a monsoon climate are a set. Likewise, the tea producers in the hill-tracts of district D are a set. Although botanically diverse, secondary crops can be considered as a set because of their important role in subsistence farming.

Set membership

The Greek letter $\in$ (epsilon) is customarily used to indicate that an object belongs to a set. If *A* again represents the set of letters in the English alphabet, then $a \in A$ means that *a* is an element of the alphabet. The symbol $\notin$ (epsilon with a slashed line) represents nonmembership. We could thus write $\alpha \notin A$, meaning that alpha is not a member of the English alphabet. Similarly, referring to the convention site set *S*, we can write Los Angeles $\in S$ and San Diego $\notin S$.

Finite and infinite sets

A set is termed *finite or infinite*, depending upon the number of elements in the set. The set *A* defined above is finite since it has 26 members, the letters in the English alphabet. The set *D* is also finite, since it has only the nine digits. The set *N* of positive integers or natural numbers is infinite, since the process of counting continues infinitely.

Example: Rational numbers are defined as that set of number *a/b*, where *a* represents all integers, both positive and negative including 0. Develop the set *R* of rational numbers and specify whether the set is finite or infinite.

Letting *R* represent the set of rational numbers, we have:

$$R = \{a/b \mid a = \text{all integers including 0}, b = \text{all integers excluding 0}\}$$

The set of rational numbers is infinite. Examples of rational numbers include any number that can be expressed as the ratio of two positive or negative whole numbers.

Example: The individuals who are members of the American Economic Association comprise a set. Assuming that a list of the membership is available, we could write the set as:

$$S = \{\text{all members of the American Economic Association}\}$$

This set is finite, although quite large. It is a set because the elements are well-defined, distinct, and of inconsequential order.

Set equality and subsets

Two sets *P* and Q are said to be equal, written $P = Q$, if every element in *P* is in Q and every element in Q is in *P*. Set equality thus requires all elements of the first set to be in a second set and all elements in the second set to be in the first set. As an example, consider the set:

$P = \{0, 1, 2, 3, 4, 5\}$ and $Q = \{2, 0, 1, 3, 5, 4\}$

These sets are equal, since every element in *P* is in Q and every element in Q is in *P*.

The student will often have occasion to consider only certain elements of a set. These elements form a *subset* of the original set. As an example, assume that F represents the farmers in the humid tropics of Asia and the Pacific, and R represent the rice farmers in the flat deltaic terrain of South Asia. It is immediately clear that there is a partial overlap between these two categories. If we restrict the two sets to the lower Gangetic Plain of the Indian sub-continent, R is a subset of F.

Subsets can be defined as follows. A set *R* is a subset of another set *F* if every element in *R* is in *F*. For example, if $F = \{1, 2, 3, 4\}$ and $R = \{1, 2\}$, then every element in *R* is in *F*, and *R* is a subset of *F*. The symbol for subset is $\subseteq$. *R* is a subset of *F* is written $R \subseteq F$.

Example: Let *L* represent the letters in the English alphabet and *C* represent the letters in the word 'cultivation'. Verify that *C* is a subset of *L*.

Since *L*={a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z} and C={c, u, l, t, i, v, a, t, i, o, n}, we see that:

$C \subseteq L$.

Example: Let *S* represent the students who are enrolled in an introductory agricultural economics course, and F represent the female students. Verify that *F* is a subset of *S*.
Since *S*= {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} and *F* = {2, 4, 6, 8, 10}, we see that $F \subseteq S$.

Example: Let *D* represent the set of digits and *I* represent the set of all integers including 0. Since *D* ={0, 1, 2, 3, 4, 5, 6, 7, 8, 9} and I = {. . . , -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, . . .} the finite set *D* is a subset of the infinite set *I*,i.e., $D \subseteq I$.

In the preceding example, *D* was defined as the set of all digits including 0. We previously defined *N* as the set of all positive integers excluding 0. Assume that one is interested in determining if *D* is a subset of *N*. By inspection of the sets, *D* includes the elements 0, whereas *N* does not. Consequently, *D* is not a subset of *N*. This is written as $D \not\subseteq N$. The notation for subset $\subseteq$ and not subset $\not\subseteq$ parallels the notation for member $\in$ and not member $\notin$.

Proper subset

The term *subset* is often differentiated from that of *proper subset*. A proper subset is designated by the symbol $\subset$. A proper subset *P* is a subset of another set *U*, written $P \subset U$, if all elements in *P* are in *U* but all elements in *U* are not in *P*. This simply means that for *P* to be a proper subset of *U*, then *U* must have all elements that are in *P* plus at least one element that is not in *P*. As an example, if

$$S = \{0, 1, 2, 3, 4\}$$

and

$$R = \{ 0, 1, 2\}$$

then *R* is a proper subset of *S*, i.e., $R \subset S$.

Example: Verify that the set *C*, the letters in cultivation, is a proper subset of *L*, the letters in the English alphabet. Since every letter in *C* is in *L*, but every letter in *L* is not in *C*, we conclude that $C \subseteq L$.

Universal set

In discussing sets and subsets, the terms *universal* set and *null* set is often encountered. The term “universal set” is applied to the set that contains all the elements the analyst will want to consider. In contrast, the *null* set is defined as a set that has no elements or members. If, for example, we are interested in categorizing the farmers in Asia, the universal set would be all farmers in Asia regardless of other characteristics. The various categories of farmers (rice, wheat etc.) would then be the subsets of the universal set. To be more specific, let us assume that an analyst in the Ministry of Agriculture is required to conduct a sample survey of target farmers in a set of districts, where D = {1, 2, 3). It is clear that this new set would be a representative sample, and is in fact a subset of the universal set U that includes all the districts under investigation. Similarly, if the analyst were interested in a certain combination of letters, the universal set would be defined as L, the letters of the English alphabet. It would then be possible to specify various subsets of the universal set L, such as C, the letters in the word cultivation.

To give an example of the *null* set, let us assume that three farmers are provisionally selected by the Export Promotion Bureau of a country for possible prize distribution. This is a special group of farmers who have demonstrated outstanding performance in a nation-wide agricultural exhibition. If we define the universal set as consisting of all farmers in the country, then it would follow that the group of three farmers is a subset of the national universal set.

Suppose that this subset of farmers is now required to take a general skill test to be eligible for the possible prizes. If we refer to the farmers as F1, F2 and F3, the results of the skill test could be {F1, F2, F3}, {F1, F2}, {F1, F3, {F2, F3}, {F1},{F2}, {F3}, { }. Note that there are eight possible outcomes and these are shown as subsets. One of the subset is the null set { }.

The null set, also referred to as the *empty* set, is designated by the Greek letter ∅ (phi). This indicates that set, subsets and the null set are included as subsets of the universal set. This concept is further elaborated by the following examples with specific reference to management and database.

From a management perspective, one can begin with the tasks of the state agencies as a given, and approach the overall picture by making lists of state organizations, departments, agencies and so on, to build a directory of tasks from the bottom up. All the activities required to perform and manage the defined tasks can be considered as a universal set. Specific activities that are related to the management of particular state organizations, departments or agencies may be seen as subsets of the universal set.

From a database context, we can start with the Central Bureau of Statistics (CBS) in country X that holds the statistical data at the national level. All the information available at the CBS domain may be treated as a universal set. On the other hand, the data that are organized at some specific geographic or sectoral levels may be considered as subsets of the universal set. It is also possible to examine subsets in a variety of ways.

Assume a situation where a researcher in an agricultural development agency would like to find out the production of certain crops e.g. wheat and other biophysical data at the district level. He can get the required information from the CBS database which will be the domain of his universal set. Now, if the researcher were interested in areas that produce wheat under low rainfall condition, then the new group of districts would be thought of as a subset of the universal set U.

To further illustrate the example, let us consider that the researcher wants to group those districts that have reported high yields in wheat production. Districts representing higher wheat productivity would constitute another subset of the universal set U. It has been observed that high productivity is closely associated with the use of chemical fertilizer. Now, the researcher may be willing to construct an additional subset of districts (of universal set U) showing the use of fertilizer at a certain rate.

Counting subsets

The number of possible subsets of the universal set can be calculated through the use of a straightforward formula. The number of possible subsets is given by the formula:

$$N = 2^n \tag{1.1}$$

where n represents the number of elements in the universal set and N is the number of possible subsets. If the universal set contains three members, there are eight possible subsets. Similarly, for a universal set with four members there are $N = 2^4 = 16$ possible subsets.

Set algebra

Set *algebra* consists of certain operations on sets whereby the sets are combined to produce other sets. It provides a technique through which the analyst, manager or planner can

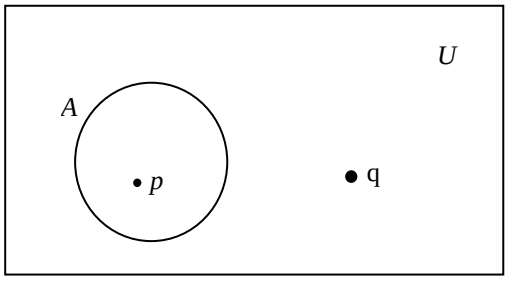
make decisions by analyzing different elements of the universal set U. Successive formations of sets are an important consideration in spatial planning. Because, spatial units can be organized on the basis of either a single criterion or multiple criteria. In the event of multivariate regionalization, it is necessary to define those areas that meet more than one condition. The delineation of regions (areas of similar attributes or homogenous in character) is, therefore, a very important aspect in geographic decision-making. This is where algebra of sets comes in. By using different combinations of set operations, i.e. complementation, intersection and union, we can construct the boundary of new regions for efficient allocation of scarce resources. The following are some of the examples of possible set operations:

These operations are most easily illustrated through use of the Venn diagram.

Venn diagram

The Venn diagram, named after the English logician John Venn (1834-83), consists of a rectangle that conceptually represents the universal set. Subsets of the universal set are represented by circles drawn within the rectangle. In Figure 1 the universal set *U* is represented by the rectangle and the subset *A* by the circle. The Venn Diagram shows that *p* is a member of *A* and that Q is not a member of *A*. Both *p* and Q are members of the universal set.

Figure 1 The universal set.

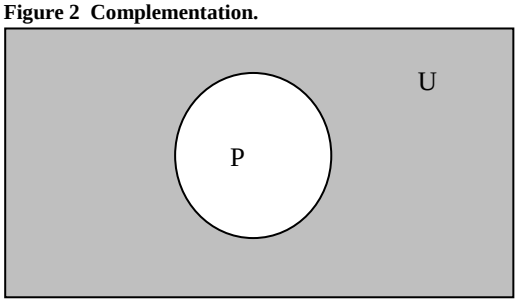


The Venn seems completely trivial, but it is not. The use of a two-dimensional picture to reflect sets points straight to spatial analysis. The following operations - complementation, intersection, union, and combined operations - are all part of the spatial analytical toolkit, as well as logical steps in set algebra.

Complementation

The first set operation we consider is that of *complementation*. Let *P* be any subset of a universal set *U*. The complement of *P*, denoted by *P ′* (read "*P* complement"), is the subset of elements of *U* that are not members of *P*. The complement of P is indicated by the shaded portion of the Venn diagram in Figure 2.

Figure 2 Complementation.



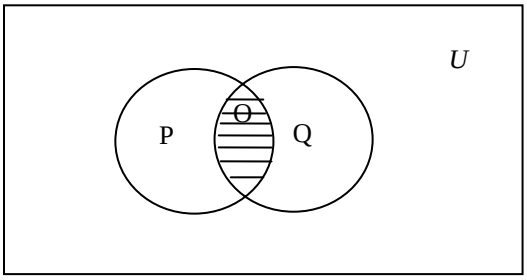
Example: For the universal set D = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} with the subset P = {0, 1, 3, 5, 7, 9}, determine P'.

The complement of *P* contains all elements in *D* that are not members of *P*. Thus, P' = {2, 4, 6, 8}.

Intersection

The second set operation is intersection. Again, let P and Q be any subsets of a universal set U. The intersection of P and Q, denoted by P ∩ Q (read "P intersect Q"), is the subset of elements of U that are members of both P and Q. P ∩ Q is shown by the shaded area of Figure 3.

Figure 3 Intersection.



Example: For the universal set *D* = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} with subsets *P* = {0, 1, 3, 5, 7, 9} and Q = {0, 2, 3, 5, 9}, determine the intersection of *P* and Q.

The intersection of *P* and Q is the subset that contains the elements in *D* that are simultaneously in *P* and Q. Thus,

$$P \cap Q = \{0, 3, 5, 9\}$$

With a little thought the reader can also recognize that

$$P \cap U = P$$

and

$$P \cap P' = \emptyset$$

Sets such as $P \cap P' = \emptyset$ which have no common members are termed *disjoint* or *mutually exclusive*.

Example: In spatial terms, let us consider region A as a subset of districts that produce rice or wheat, and region B is another subset of districts that grow maize or wheat. From these two subsets, we can specify a new subset for those districts that produce wheat.

For the universal set U={1, 2, 3, 4, 5, 6, 7, 8, 9} with subsets A ={1, 3, 5, 7, 9} and B={2, 3, 4, 5, 6, 8}, determine the intersection of A and B.

The intersection of A and B is the subset of those districts that contain the elements in U that are simultaneously in A and B. Thus,

$$A \cap B = \{3, 5\}.$$

The reader can also recognize that:

$$A \cap U = A$$

and $A \cap A' = \varnothing$

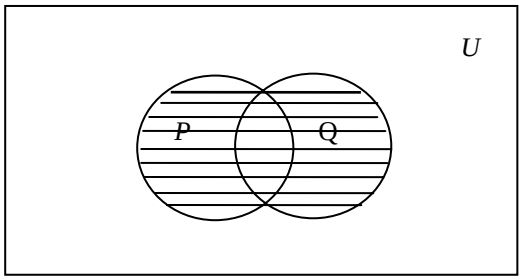
Sets such as $A \cap A' = \varnothing$ which have no common members are termed disjoint or mutually exclusive.

Union

The third set operation is *union*. If we again let *P* and Q be any subsets of a universal set *U*, then the union of *P* and Q, denoted by $P \cup Q$ (read "*P* union Q"), is the set of elements of *U* that are members of either *P* or Q. $P \cup Q$ is shown by the shaded portion of the Venn diagram in Figure 4.

Example: For the universal set $D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ with subsets $P = \{0, 1, 3, 5, 7, 9\}$ and $Q = \{0, 2, 3, 5, 9\}$, determine the union of *P* and Q.

Figure 4 Union.



The union of *P* and Q is the subset that contains the elements in *D* that are in Q. Thus,

$$P \cup Q = \{0, 1, 2, 3, 5, 7, 9\}$$

From the definitions given, it also follows that:

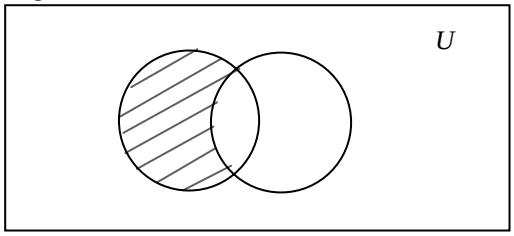
$$P \cup U = U$$
$$P \cup P' = U$$

Other set operations

Two additional set operations are sometimes included in the algebra of sets. The two operations are *difference* and *exclusive union*. Since subsets formed by either of these set operations can also be formed by use of complementation, intersection, and union, these operations are often excluded in discussing set algebra.

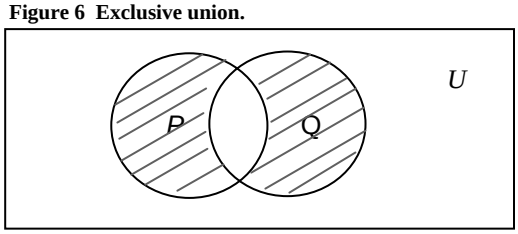
Let *P* and Q be any subsets of the universal set *U*. The difference of *P* and Q, denoted by $P - Q$ (read "*P* minus Q"), is the subset that consists of those elements that are members of *P* but are not members of Q. This subset is shown in the Venn diagram in Figure 5. The difference of $P - Q$ can also be expressed as $P \cap Q'$.

Figure 5 Difference.



The exclusive union of *P* and Q, denoted by $P \underline{\cup} Q$ (read P exclusive union Q"), is the set of elements of U that are members of P or of Q but not of both. The subset is shown in the Venn diagram (Figure 6). The subset can also be expressed as $(P \cap Q) ' \cap (P \cup Q)$ or $(Q' \cap P) \cup (P' \cap Q)$.

Figure 6 Exclusive union.



Combining set operations

Part of the utility of the algebra of sets is due to the ability to combine two or more sets into new sets through the use of the set operations.

The use of set algebra to form sets can be illustrated by Figure 7. This figure consists of a Venn diagram, representing the universal set, and two subsets *P* and Q. Areas in the Venn diagram are shown by the letters *a*, *b*, *c*, and *d*. The set that consists of areas *a*, *b*, *c*, and *d* is the universal set *U*, while the set *P* consists of *a* and *b*, the set Q consists of *b* and *c*, etc. Using this notation, we can use the algebra of sets to construct all possible subsets. Students should carefully study Figure 7 and the construction of subsets using the set operations describing the areas *a*, *b*, and *c* presented in Table.

Figure 7 The use of set algebra to form sets.

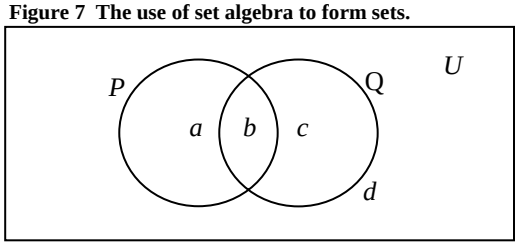


Table 1 Construction of subsets.

Area	Set
a, *b*, *c*, *d*	*U*
a, b	*P*
b, *c*	Q
a, *d*	Q
c, *d*	*P*
b	$P \cap Q$
a, *b*, *c*	$P \cup Q$
d	$(P \cup Q)'$
a	$P \cap Q'$ or *P*—Q
c	$P' \cap Q$ or Q—*P*
a, *c*	$(P \cap Q') \cup (P' \cap Q)$ or $(P \cap Q)' \cap (P \cup Q)$, or $(P \underline{\cup} Q)$, or (*P*—Q) ∪ (Q—*P*)
a, *c*, *d*	$(P \cap Q)'$

Problem statement - regional agricultural planning: The Agricultural Development Corporation (ADC) of country X wants to expand the irrigation facilities in regions *A*, *B* and *C* through the installation of deep tube wells. The major problem is that the ground water tables in these regions are not uniform. Information shows that water tables in the northern districts are relatively low compared to the southern districts. Further, the ADC is concerned about the current land-use patterns. The major objective of the organization is to bring a set of districts under the new irrigation scheme where multiple cropping systems can be easily introduced. To delineate the boundary of the planning/functional region, the planner would first construct the union of the districts denoted by regions *A*, *B* and *C* (subsets of the universal set *U*) on the basis of ground water availability. To pinpoint the areas that are good for multiple cropping, the planner would request the intersection of the subsets (regions).

Example: For the universal set *U* = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10.....100} with subsets *A* ={1, 2, 3}, *B*={4, 5, 6} and *C*={7, 8, 9}, determine the union of *A*, *B* and *C*. The union of *A*, *B* and *C* is the subset that contains the elements in *U* that are in *A*, *B* and *C*. Thus,

$$A \cup (B \cup C) = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

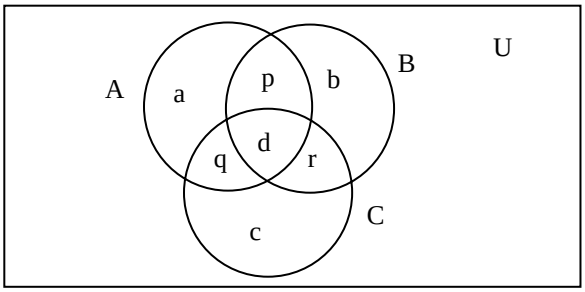
Now, assume that the planner wants to construct the intersection of regions A, B and C. It is the idea of the planner to pull out those districts that have potential for multiple cropping as judged by the soil condition and water levels. If = {1, 3, 5, 7}where M denotes districts with multiple cropping areas the intersection of *A*, *B* and *C* would be as follows: the intersection of A, B and C is the subset M that contains the elements in U, the properties of which are simultaneously in A, B and C. Thus,

$$A \cap (B \cap C) = \{1,3,5,7\}$$

It is important to note that in constructing a new structured subset like this one (through the intersection operational method) the elements in U can not be found simultaneously in cooperating subsets (in this case districts in regions A, B and C). However, a common underlying regularity or communality should be observed in the element of U created by the intersection. In this example, districts reporting good soils and ground water accessibility can not be simultaneously in regions A, B and C as elements of U. However, the discriminating properties of the elements in U could be observed simultaneously in A, B and C. This is a major difference between structured and non-structured subsets.

The use of set algebra to form structured sets can be illustrated by Figure 8. This figure consists of Venn diagram, representing the universal set U, and three subsets A, B and C. Areas in the Venn are shown by letters a, b, c, d, p, q and r. The set that consists of these areas is the universal set U, while the set d consists of A, B and C which is equivalent to the set M (representing multiple cropping areas).

Figure 8 The use of set algebra to form structured sets.



Problem statement - agricultural policy planning: Consider the Planning Commission of country Y. The Central Planning Agency is in a phase of identifying the policy alternatives between two important food crops, rice and maize. The main objective of the commission is to promote maize cultivation in rice producing districts that receive low average rainfall on an annual basis and have no access to the agricultural input centers (e.g. use of chemical fertilizer at subsidized rate).

Although rice is a popular staple diet, largely grown and consumed in the country, it's stock was never sufficient to meet the growing demands of the population. The current food issue dictates the needs of other crops that can be produced on a cost-effective basis, even in the adverse climatic conditions. Cost-benefit analysis suggests that the cost of production of maize per unit area is much lower than that of rice. This is simply because the cash expenses used for procuring modern or traditional inputs from outside sources are extremely limited. As a low cost crop, maize occupies an important position in the subsistence economy. Given the above conditions, our operations of sets will be as follows.

Example: Let U represent the universal set of districts in country Y, where U = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10.....100}, R refers to the subset of rice growing districts, where R ={1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, P represents the subset of districts receiving low average annual precipitation, where P= {1, 2, 3, 8, 9, 10 }, and F refers to the subset of districts having no access to centers for commercial fertilizer, where F={1, 8, 9, 10}, determine the intersections of subsets R, F and P.

The intersection of R, F and P is the subset that contains the elements in U that are simultaneously in R, F and P. Thus,

$$R \cup (F \cup P) = \{1, 8, 9, 10\}$$

Agro-climatic regionalization: Suppose the National Agricultural Research Institute of country A wants to produce an agro-climatic map of region R. The prime objective of the map is to provide a basic means of assessing the climatic suitability of geographical areas for various agricultural alternatives. It recognizes that the major aspects of climate that affect plant growth are precipitation and temperature.

Example: Let us consider a set of districts denoted by region R, where R = {1, 2, 3, 4, 5, 6, 7, 8, 9}. Three datasets that are available to help the researcher to produce the agro-climatic map are as follows:

i) A subset of districts X showing high mean annual rainfall
where X= {3, 4, 5, 6, 7,8, 9}
ii) A subset of districts Y representing relatively high altitude
where Y= {5, 6, 7, 8, 9}
iii) A subset of districts Z reporting moderate mean annual temperature
where Z = {1, 2, 3, 4, 5, 6, 7}

From the above datasets the researcher can construct a number of intersections. If the researcher wants to generate a map of the districts that show high mean annual rainfall and high altitude then he would construct the intersection of subsets X and Y. Thus,

$$X \cap Y = \{5, 6, 7, 8, 9\}$$

If he were further interested in creating a new group of districts that meet the conditions of high average annual rainfall, high altitude and moderate range of annual temperatures, then he would delineate the boundary of the new subset as follows:

$$X \cap (Y \cap Z) = \{ 5, 6, 7\}$$

Law of the algebra of sets

The algebra of sets is composed of certain laws. In some cases these laws are similar to the algebra of numbers. We shall state these laws in the form of postulates, thus relieving the reader of the burden of studying mathematical proofs of the laws. The less obvious postulates are illustrated with Venn diagrams.

Postulate 1: $P \cup Q = Q \cup P$
Postulate 2: $P \cap Q = Q \cap P$

Postulates 1 and 2 are termed the *commutative law*. These postulates state that the order in which we combine the union or intersection of two sets is immaterial. This corresponds to the commutative law in ordinary algebra, which states that $p + q = q + p$ and that $p \cdot q = q \cdot p$.

Postulate 3: $P \cup (Q \cup R) = (P \cup Q) \cup R$
Postulate 4: $P \cap (Q \cap R) = (P \cap Q) \cap R$

Postulate 3 and 4 are termed the *associative law*. These postulates state that the selection of two of three sets for grouping in a union or intersection is immaterial. Thus, the order in which the sets are combined is immaterial. These postulates correspond to the associative law in

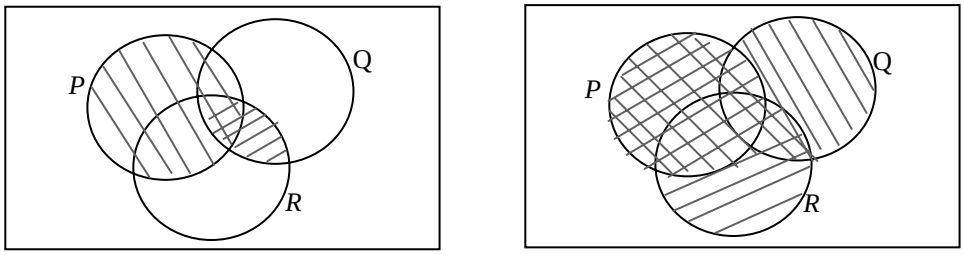
ordinary algebra, which enables us to state that $p + (q + r) = (p + q) + r$ and that $p \cdot (q \cdot r) = (p \cdot q) \cdot r$.

Postulate 5: $P \cup (Q \cap R) = (P \cup Q) \cap (P \cup R)$
Postulate 6: $P \cap (Q \cup R) = (P \cap Q) \cup (P \cap R)$

Postulates 5 and 6 are called the *distributive law*. Postulate 5 has no analogous postulate in ordinary algebra. Postulate 6 corresponds to the distributive law in ordinary algebra that enables us to state that that $p + (q + r) = p \cdot q = p \cdot r$.

Since Postulates 5 and 6 are not as obvious as Postulates 1 through 4, we shall verify one of them through Venn diagrams. To verify Postulate 5, we must show that the area represented by $P \cup (Q \cap R)$ is the same as that represented by $(P \cup Q) \cap (P \cup R)$. The area representing the set $P \cup (Q \cap R)$ is shown in Figure 9(a). In Figure 9(a) the intersection of Q and R is shown by the shading from lower left to upper right, and P is shown by shading from upper left to lower right. The union of P with $Q \cap R$, represented by both the shaded and crosshatched area in the figure, gives $P \cup (Q \cap R)$.

Figure 9 Verification of Postulate 5 by Venn diagrams.

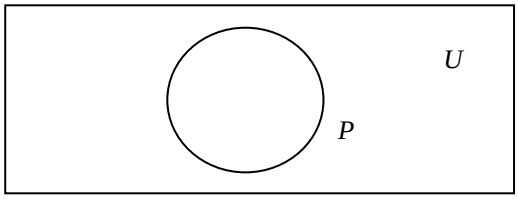


The area presenting the set $(P \cup Q) \cap (P \cup R)$ is shown in Figure 9(b). In Figure 9(b) the union of P and Q is shown by shading from upper left to lower right, and the union of P and R is shown by shading from lower left to upper right. The intersection of the two shaded areas gives the crosshatched area described as the set $(P \cup Q) \cap (P \cup R)$. Since the shaded and crosshatched area in Figure 9(a) equals the crosshatched area in Figure 9(b), we conclude that Postulate 5 is true. Postulate 6 is verified in the same manner as Postulate 5.

Postulate 7: $P \cap P = P$
Postulate 8: $P \cup P = P$
Postulate 9: $P \cup \varnothing = P$

Postulates 10 through 12 are based upon the null set, the universal set, and the complement. These postulates are obvious and can also easily be verified by the Venn diagram shown in Figure 10.

Figure 10 Verification of Postulates 10 to 12 by Venn diagram.

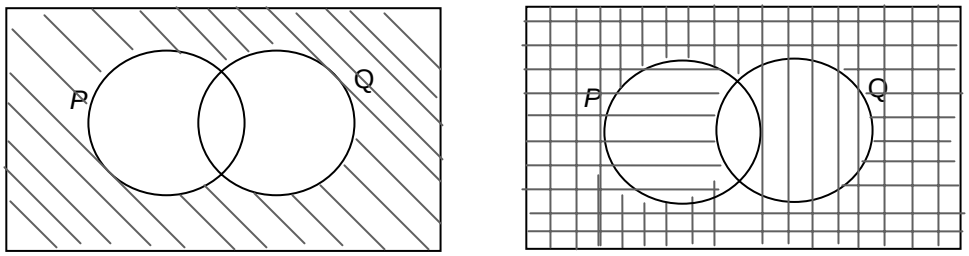


Postulate 13: $(P \cup Q)' = P' \cap Q'$
Postulate 14: $(P \cap Q)' = P' \cup Q'$

Postulates 13 and 14 are termed De Morgan's law. Postulate 13 states that the complement of the union of two sets is equal to the intersection of the complement of each set. Postulate 14 states that the complement of the intersection of two sets is equal to the union of the complement of each set. We shall verify Postulate 13 through the use of Venn diagrams. The student is asked to verify Postulate 14.

Figure 11(a) shows the set $(P \cup Q)'$ as the area shaded from upper left to lower right. In Figure 11(b), P' is shown by shading from upper left to lower right and Q' is shown by shading from lower left to upper right. The intersection of P' with Q' is shown by the crosshatched area. Since the shaded area in Figure 11(a) equals the crosshatched area in Figure 11(b), it follows that $(P \cup Q)' = P' \cap Q'$.

Figure 11 Verification of Postulate 13 by Venn diagrams.



(a) (b)

The laws of set algebra are used analogously with the laws of ordinary algebra. Just as the laws of ordinary algebra can be used to simplify algebraic expressions, the laws of set algebra can be used to simplify sets. This is illustrated by the following examples.

Example: Simplify the set $(A \cup B) \cup (A \cap B)$.

1. Let $P = (A \cup B)$:	$P \cup (A \cap B)$
2. Postulate 5:	$(P \cup A) \cap (P \cup B)$
3. Substitute for P:	$((A \cup B) \cup A) \cap ((A \cup B) \cup B)$
4. Postulate 1:	$(A \cup A \cup B) \cap (A \cup B \cup B)$
5. Postulate 7:	$(A \cup B) \cap (A \cup B)$
6. Postulate 7:	$A \cup B$

In this example, it is important for the student to use parentheses to enclose $P = (A \cup B)$. This is further illustrated by the following example.

Example: Previously, we stated that $(P \cap Q)' \cap (P \cup Q)$ was equal to $(Q' \cap P) \cup (P' \cap Q)$. Using the laws of set algebra, verify this relationship.

Given:	$(P \cap Q)' \cap (P \cup Q)$

1. Postulate 14:	$(P' \cup Q') \cap (P \cup Q)$
2. Let $R = (P' \cup Q')$	$R \cap (P \cup Q)$
3. Postulate 6:	$(R \cap P) \cup (R \cap Q)$
4. $(P' \cup Q') = R$	$((P' \cup Q') \cap P) \cup ((P' \cup Q') \cap Q)$
5. Postulate 6:	$(P \cap P') \cup (P \cap Q') \cup (P' \cap Q) \cup (P \cap Q')$
6. Postulate 12:	$P \cap Q') \cup (P' \cap Q)$

Thus

$$(P \cap Q)' \cap (P \cup Q) = (P \cap Q') \cup (P' \cap Q)$$

Example: Simplify the expression $P \cup (P' \cap Q)$.

Given:	$P \cup (P' \cap Q)$
1. Postulate 5:	$(P \cup P') \cap (P \cup Q)$
2. Postulate 11:	$U \cap (P \cup Q)$
3. Postulate 10;	$P \cup Q$

The expression simplifies to $P \cup Q$.

Example: Show that the set $(P \cup Q') \cap (P \cap Q)'$ equals Q'.

Given:	$(P \cup Q') \cap (P \cap Q)'$
1. Postulate 14:	$(P \cup Q') \cap (P' \cup Q')$
2. Postulate 5:	$Q' \cup (P \cap P')$
3. Postulate 12:	Q'

The primary difficult in understanding this example is in converting $(P \cup Q') \cap (P' \cup Q')$ to $Q' \cup (P \cap P')$. This step involves rewriting (Postulate 1) the expression as $(Q' \cup P) \cap (Q' \cup P')$. Postulate 5, the distributive law, is applied to write the expression as $Q' \cup (P \cap P')$. Since $P \cap P'$ is equal to $\varnothing$, the expression reduces to Q'.

Cartesian product

Suppose that one is asked to list the possible outcomes of two tosses of a coin. Since either a head or a tail occurs on a single toss, the possible outcomes are described by the set O = {(H, H), (H, T), (T, H), (T, T)}. There are four elements in this set and, corresponding to the requirements given for a set, the order of the elements is immaterial. Is each of the elements, however, distinct? To answer this question we must ask if the elements, (H, T) differs from the element (T, H). The answer is that these elements do differ, since the order of the occurrence is important. If, for instance, an individual bet $1 that a head would occur in the first toss of the coin and $1 that a tail would occur on the second toss, he would be $2 richer if element (H, T) occurred and $2 poorer if (T, H) occurred.

The elements (H, H), (H, T), (T, H) and (T, T) are examples of *ordered pairs*. One of the two components of each ordered pair is designated as the first element of the pair, and the

other, which need not be different from the first, is designated as the second element. If the first element is designated as *a* and the second element is designated as *b*, we have the ordered pair (*a*, *b*). This ordered pair differs from the ordered pair (*b*, *a*), and both ordered pairs differ from the set {*a*, *b*}.

Ordered pairs are formed by the *Cartesian product* of two sets. If *A* and *B* are two sets, the Cartesian product of the sets, designated by $A \times B$, is the set containing all possible ordered pairs (*a*, *b*) such that $a \in A$ and $b \in B$. If the set *A* contains the elements a_1 , a_2 , a_3 and the set *B* contains the elements b_1 and b_2 , the Cartesian product $A \times B$ is the set $A \times B = \{(a_1 , b_1), (a_1 , b_2), (a_2 , b_1), (a_2 , b_2), (a_3 , b_1), (a_3 , b_2)\}$. All possible ordered pairs are included in the set.

The concept of the Cartesian product is quite useful in many decision problems. The student of probability and statistics will often be asked to consider the possible outcomes of an experiment. As an example, consider again the problem of determining all possible outcomes of two tosses of a coin. If we define the outcome of the first toss as the set O_1 = {H, T} and the outcome of the second toss as the set O_2 = {H, T}, then the Cartesian product $O_1 \times O_2$ gives all possible outcomes of the two tosses. As we have seen, these outcomes are $O_1 \times O_2$ = {(H, H), (H, T), (T, H), (T, T)}.

The Cartesian product of two sets can be determined quite easily with the aid of a box diagram. Figure 12 shows the box diagram for the Cartesian product of O_1 and O_2. The method of constructing the diagram involves listing the elements of O_1 to the left of the box and O_2 above the box. The blanks in the box are then filled in with the ordered pairs. The Cartesian product $O_1 \times O_2$ consists of the elements in the box, i.e., (H, H), (H, T), (T, H), (T, T).

Figure 12 The Cartesian product of two sets.

		O_2	
		H	T
O_1	H	H, H	T, T
	T	T, H	T,

The Cartesian product can be expanded to combine more than two sets. This means that the concepts discussed for ordered pairs can, for example, be applied to *ordered triplets*. To illustrate, assume that we are asked to list all possible outcomes of three tosses of a coin. Denoting O_i = {H, T}, where O_i represents the possible outcomes on the *i*th toss (i.e., $i = 1$ represents the first toss, $i = 2$ represents the second toss, etc.), the possible outcomes would be given by the Cartesian product $O_1 \times O_2 \times O_3$. This Cartesian product is determined by finding $O_1 \times O_2$ and then $(O_1 \times O_2) \times O_3$. From Figure 12, we know that $O_1 \times O_2$ = {(H, H), (H, T), (T, H), (T, T)}. $(O_1 \times O_2) \times O_3$ is shown in Figure 13 The Cartesian product of $(O_1 \times O_2) \times O_3$ is {(H, H, H), (H, T, H), (T, H, H), (T, T, H), (H, H, T), (H, T, T), (T, H, T), (T, T, T)}.

Figure 13 The Cartesian product of $(O_1 \times O_2) \times O_3$.

		$O_1 \times O_2$			
		H, H	H, T	T, H	T, T
O_3	H	H, H, H	H, H, T	H, T, H	H, T, T
	T	T, H, H	T, H, T	T, T, H	T, T, T

The box diagrams of Figures 12 and 13 provide a straightforward method of determining of the Cartesian products of sets. With some practice, the student can apply this concept without difficulty. Before illustrating the concept with examples, however, let us carry the Cartesian product one additional step. Assume that we are asked to list all possible outcomes of four tosses of a coin. There are 16 such outcomes. Most individuals would be extremely hard pressed to think of all sixteen. If we use the concept of the Cartesian product, the task becomes routine. The possible outcomes are given by the set $O_1 \times O_2 \times O_3 \times O_4$. The box diagram can be expressed in terms of $(O_1 \times O_2) \times (O_3 \times O_4)$ or by any other grouping of the parentheses. Figure 14 shows the Cartesian product of $(O_1 \times O_2) \times (O_3 \times O_4)$.

Figure 14 The Cartesian product of $(O_1 \times O_2) \times (O_3 \times O_4)$.

$O_3 \times O_4$

$O_1 \times O_2$	H, H	H, T	T, H	T, T
H, H	H, H, H, H	H, H, H, T	H, H, T, H	H, H, T, T
H, T	H, T, H, H	H, T, H, T	H, T, T, H	H, T, T, T
T, H	T, H, H, H	T, H, H, T	T, H, T, H	T, H, T, T
T, T	T, T, H, H	T, T, H, T	T, T, T, H	T, T, T, T

The elements in the box diagram such as (H, H, H, T) contain four members. Mathematicians call such an element an ordered "4-tuple." Using this term, an ordered pair could be referred to as an ordered 2-tuple, an ordered triplet as an ordered 3-tuple, etc. In general, then, an element containing *n* members is referred to as an ordered *n*-tuple.

Example: A retailer specializes in three products: color televisions, black and white televisions, and stereos. He offers a service contract with the sale of each of the products, which the customer may or not elect to purchase. Determine the possible combination of sales options.

Let the products be represented by the set $P = \{C, B, S\}$ and the sales contract by $R = \{E, E'\}$. The Cartesian product of $P \times R$ gives the combination of sales options. Figure 15 shows the elements of $P \times R$.

Figure 15 The Cartesian product of $P \times R$.

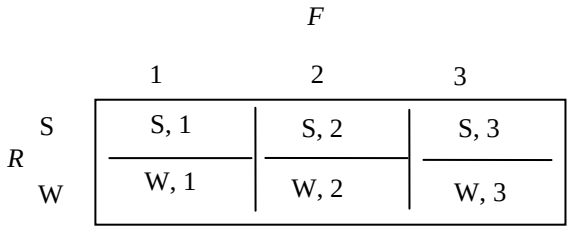
P

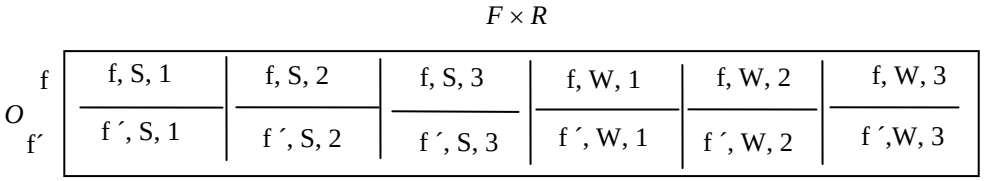
R	C	B	S
E	E, C	E, B	E, S
E′	E′, C	E′, B	E′, S

Example: A builder has three basic floor plans: single story, two story, and trilevel. Each of these plans can have either a shake roof or a wood shingle roof. In addition, the plans are available with or without fireplaces. Determine the number of combinations of plans and show these plans in a box diagram.

Let the basic floor plans be represented by the set $F = \{1, 2, 3\}$, the roofing material by the set $R = \{S, W\}$, and the fireplace option by the set $O = \{f, f'\}$. The combination of plans is represented by the set $\{F \times R \times O\}$. The box diagram for the set is shown in Figure 16.

Figure 16 The Cartesian product of $F \times R \times O$.



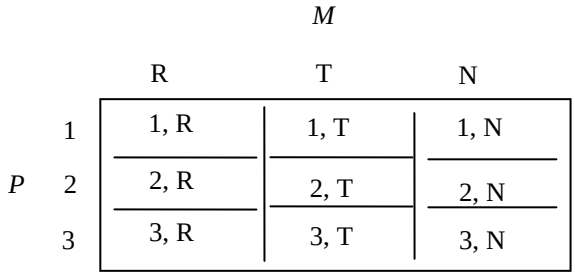


There are twelve combinations of plans.

Example: An advertising agency is placing ads for three products. The media available for advertising are radio, television, and newspaper. The ads will be written by either the agency or the sponsor. Develop the possible combinations with the aid of a box diagram.

Let the products be represented by the set $P = \{1, 2, 3\}$, the media by the set $M = \{R, T, N\}$, and the source of the advertisement by the set $S = \{A, A'\}$. The box diagram for $P \times M \times S$ is constructed as shown in Figure 17.

Figure 17 The Cartesian products of $P \times M \times S$.



$M \times P$

S	1, R	1, T	1, N	2,R	2, T	2, N	3, R	3, T	3, N
A	A, 1, R	A, 1, T	A, 1, N	A, 2, R	A, 2, T	A, 2, N	A, 3, R	A, 3, T	A, 3, N
A'	A´, 1, R	A´, 1, T	A´, 1, N	A´, 2, R	A´, 2 , T	A´, 2, N	A´, 3, R	A´, 3, T	A´, 3, N

The total number of members formed by the Cartesian product of two sets is given by the product of the number of elements in each set. Thus if set A contains five elements and set B contains four elements, then the Cartesian product $A \times B$ contains 5(4) = 20 elements. The rule applies to the Cartesian product of more than two sets. For the three sets A, B, and C, containing N_1, N_2 and N_3 elements respectively, the Cartesian product of the sets $A \times B \times C$ contains $N_1 \cdot N_2 \cdot N_3$ elements.

Example: Set A has 10 elements, set B has 6 elements, set C has 12 elements, and set D has 3 elements. Determine the number of elements in the Cartesian product of $A \times B \times C \times D$.

The number of elements is given by product of the number of elements in each set. Thus the Cartesian product contains $10 \cdot 6 \cdot 12 \cdot 3$, or 2,160 elements.

Reference

Childress, R.L 1974. Mathematics for Managerial Decisions. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

Introduction to Spreadsheets

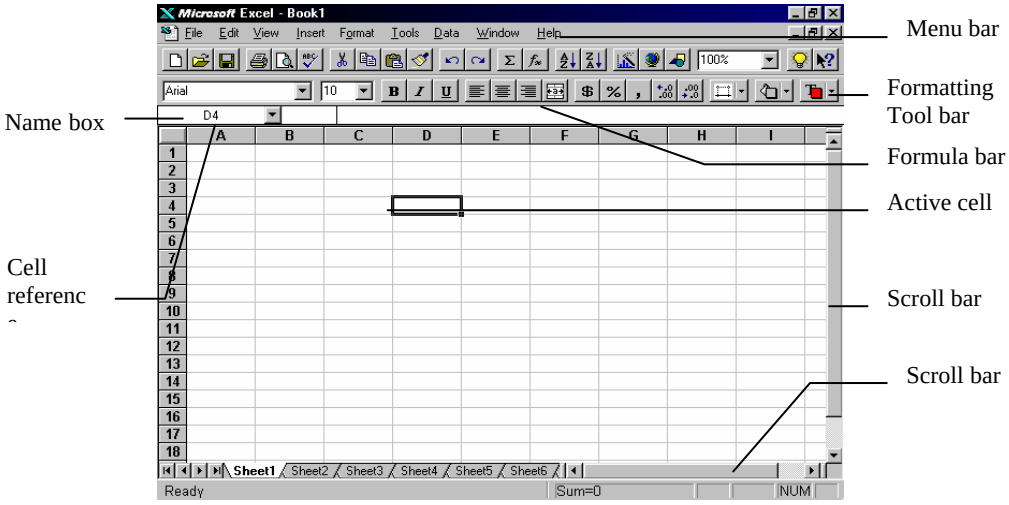
Gary Timoshenko[*]

Introduction

Spreadsheets are easy to use and powerful computer applications. They provide flexibility and convenience when doing numerical work and help simplify the process of calculating groups of numbers. Microsoft Excel is one of the most commonly used spreadsheets along with Lotus 123. Almost everything you learn about Microsoft Excel can be applied to other spreadsheets. This document is only intended as a basic introduction to spreadsheets. It is intended to help the first time user of spreadsheets so he can go on to other parts of the course such as linear programming and statistics where spreadsheets are used. The uses and functions of spreadsheets are so varied that it is impossible to teach everything.

A worksheet or spreadsheet is a collection of cells organized into columns and rows. Each cell may contain data in the form of text, numbers or formulae. The cell of your spreadsheet that is highlighted is called the active cell. This is the cell that is ready to receive data or a command. When you change the active cell, the name box located in the upper left-hand side of the spreadsheet shows a new cell reference. You can change the active cell in a worksheet by using the mouse or the keyboard. You can also scroll to different locations using the scroll bars but this does not change the active cell. To change the active cell after scrolling, simply move the mouse to a cell and click.

The cell reference can also be changed by clicking F5, by choosing the Goto command from the Edit menu, or by typing a cell name (eg. B7) in the name box. The Excel screen looks like this:



[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

Entering data

The data can be entered as labels, numbers or a formula.

Labels. When an alphabetical character or a symbol (‘~!#%^&*(){};:`”<>,?) is entered as the first character in a cell, the cell contains a label. A label means the cell cannot be used in calculations. By default each cell is 9 characters wide; however it is possible to view a label that is longer than the cell width if the cell to the right is blank. A label is entered in the cell after you press the enter key, press an arrow key, click another cell or click the enter box on the formula bar  . Label contents are automatically left justified.

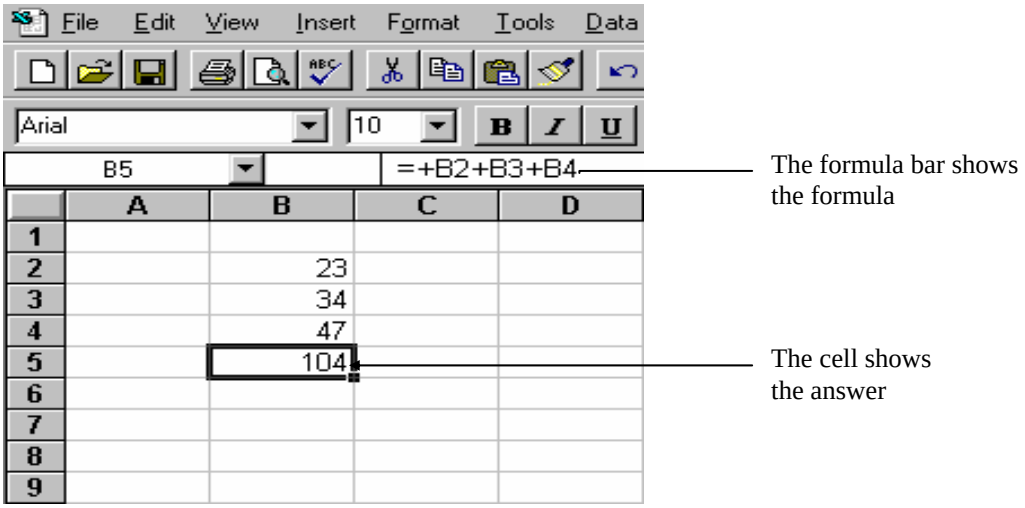
There are a few ways to correct the contents of a label. Before the label is entered you may use the backspace key. To delete the entire entry press the Escape key or click the cancel box on the formula box. After the text is entered you can correct by typing over the existing text or, by hitting F2, you can edit the entry.

Values. When a number or certain symbol (+-.=$) is entered as the first character in a cell, the cell contains a value. Values can be used in calculations. If a value is larger than the cell, Excel displays the number in scientific notation or number signs (######) appear in the cell. The cell width must be reset in order to view the number. We will cover this later.

Sometimes we want to enter numbers which we do not want to use in calculations. Examples of these are telephone numbers and identification numbers. These should be entered as numeric labels. To indicate that these numbers should be treated as labels, not values, it is necessary to begin the entry with a label prefix, an apostrophe (’) .

Formulae. A formula is an instruction to calculate a number. A formula is entered in the cell where the answer should appear. As you type the formula, it appears in the cell and in the formula bar. After the formula is entered, the answer is displayed in the cell and the formula is displayed in the formula bar.

Cell references and mathematical operators are used to develop formulae. A formula must start with an equal sign (=). For example =C3+C4+C5 adds the values in these three cell locations. Any changes made to the value of these cells causes the answer to change automatically.



The standard mathematical operators used in formulae are: + addition; - subtraction; * multiplication; / division; and ^ exponentiation. The mathematical order of operations is standard. Parentheses have the highest priority, followed by exponents, then multiplication and division and finally addition and subtraction.

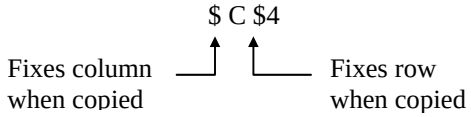
Saving a worksheet

To save a worksheet, click File then Save. Type the name of your worksheet. Excel automatically adds the extension .XLS. You must save your worksheet before you exit Excel or your work will be lost. If you attempt to exit Excel before saving your worksheet, you will be asked if you want to save the changes.

Copying data

Copying labels and values is done much as in other windows programs using either the toolbar buttons or the Edit menu. Copying formulae is different. When a formula is copied, the cell references change relative to their new location. For example, if the formula =C2+C3 in cell C4 is copied to cell D4, the formula in cell D4 will be =D2+D3.

In some cases, a value in a formula must remain constant when copied to another location. This is referred to as an absolute reference. To identify a cell as an absolute reference, a dollar sign ($) must precede the row and the column references for that cell. A dollar sign before the column reference fixes the column when the formula is copied and a dollar sign before the row reference fixes the row.



Exercise 1

In the exercises we will be constructing a spreadsheet to compare the profitability of planting two different crops. To begin enter the following starting in cell A1:

Column A	Column B	Column C
Crop	Maize	Soybean
Hectares	20	20
Yield (kg/ha)	3000	1400
Price ($/kg)	0.13	0.3
Seed Rate (kg/h)	25	45
Seed Cost ($/kg)	0.6	0.8
Fertilizer Rate (kg/h)	140	120
Fertilizer Cost ($/kg)	0.4	0.4
Machinery/Livestock ($/h)	4	7
Transport ($/h)	2	2.4
Labour Costs ($/h)	20	28

Above are the cost assumptions for our farm. We now want to calculate the expected return for maize. First calculate revenue = Yield*Price*Hectares. Then calculate the individual costs using formulae:

Seeds	(Seed Rate*Seed Costs*Hectares)
Fertilizer	(Fertilizer Rate*Fertilizer Cost*Hectares)
Machinery/Livestock	(Machinery*Hectares)
Transport	(Transport*Hectares)
Labour Costs	(Labour Costs*Hectares)
Total costs	(Sum all the costs)
Profit	Revenue-Total Costs

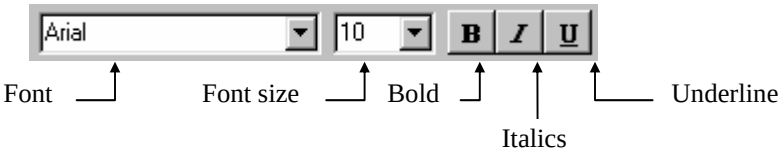
To start, in cell A13 type “Revenue” and in B13 type “=B3*B4*B2”. In cell A14 type “Seeds” and in cell B14 type “=B5*B6*B2”. Then figure out the rest of the formulae.

Once you have entered the formulae for the Maize column, copy the formulae into the Soybean column. Remember to click and hold the right mouse button to highlight the area you want to copy, then use the Edit menu or the Copy and Paste buttons. The screen should now look like this:

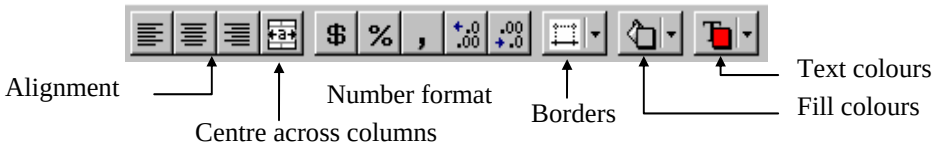
	A	B	C	D
1	Crops	Maize	Soybean	
2	Hectares	20	20	
3	Yield (kg/h	3000	1400	
4	Price ($/kg	0.13	0.3	
5				
6	Seed Rate	25	45	
7	Seed Cost	0.6	0.8	
8	Fertilizer R	140	120	
9	Fertilizer C	0.4	0.4	
10	Machinery	4	7	
11	Transport	2	2.4	
12	Labour Co	20	28	
13	Revenue	7800	8400	
14	Seed Cost	300	720	
15	Fertilizer	1120	960	
16	Machinery	80	140	
17	Transport	40	48	
18	Labour	400	560	
19	Total Cost	1940	2428	
20	Profit	5860	5972	
21				

Formatting data

Excel allows you to format the data to make the screen view and printout more attractive. Starting at the left of the formatting tool bar we can first change the font and the font size. When the font size is changed the row height automatically adjusts if a taller row is needed. The column width does not adjust automatically. Next we can **Bold**, *italicize*, or underline the cell entries by highlighting the cells we want to change and clicking the appropriate buttons.



The easiest way change the alignment of cells is by using the alignment buttons on the Formatting toolbar. After you have entered a label, simply click on the buttons to change the alignment. We can also center across columns. Next we can change the numbering format to percent, dollars, or comma delineated and we can add or take away decimal places. Finally we can add borders, colours or change the colours of the text. Each of these functions can also be found in the Format, Cells menu.



We can also format the row height and the column width. We can do this two ways. One way is using the menu. Select at least one cell from each column you want to adjust. Then Format, Column, Width, and enter the desired width in the Column Width edit box. Click OK or press Enter to set the column width. The second way is to move the mouse pointer to the column header area and position the pointer at the right edge of the column you want to adjust. Press and hold down the left mouse button, then drag the column to the desired width. Row height can be adjusted in much the same way. The Format, Row, Height menu can be used or the mouse pointer can be used to drag rows to the desired height.

Functions

A function is a built-in formula that performs a special calculation automatically. For example the SUM function can be used to add all the values in a specified range of cells. The formula =SUM(A4:A6) would add the values in A4, A5, A6.

Some other common functions are Average(), which gives the average value of a range of cells, Count() which counts the non-zero cells in a range, Max() and Min() which give the maximum and minimum values for a specified range, Round() which rounds a number to a specified number of digits and If() which denotes a logical argument. The Function wizard,

 located on the toolbar, lets you select functions from a list and prompts you for the required arguments.

Inserting and deleting columns and rows

Under the Insert menu click Rows and a row will be inserted above the current row, or click column and a column will be inserted to the left of the current column.

Exercise 2

In this exercise we will attempt to improve the presentation of our spreadsheet.

- First increase the column widths so we can see all the data. This can be done by selecting Column, Width under the Format menu or using the mouse pointer in the column header area.
- Insert 3 rows above the first row and enter the title "Expected Profits from Selected Crops" in the first row. To insert a row, choose Row from the Insert menu, and a row will be inserted above the active cell.
- Make the title bold with a font size of 14 by highlighting the title then using the font size option on the formatting tool bar.
- Delete "Crops" in cell A4. Simply make A4 the active cell then press the Delete key.
- Insert a row between "Price" and "Seed Rate".
- Insert two rows between "Labour Costs" and "Revenue" and insert one row between "Revenue" and "Seed Costs" and between "Total Costs" and "Profit".
- Change the number format for Profits row to Currency Style. Highlight the entries for Maize profits and Soybean profits the click the dollar sign  on the formatting toolbar.
- Add a border between rows 24 and 25 in column B and C to indicate we are adding the costs. Highlight cells B24 and B25 then click the Borders button on the formatting tool bar.  If the border is not set as a single line below the cell, Click the arrow beside the borders button to choose a new border.

Your spreadsheet should now look like this:

	A	B	C	D
1	**Expected Profits from Selected Crops**			
2				
3				
4		**Maize**	**Soybean**	
5	Hectares	20	20	
6	Yield (kg/ha)	3000	1400	
7	Price ($/kg)	0.13	0.3	
8				
9	Seed Rate (kg/h)	25	45	
10	Seed Cost ($/kg)	0.6	0.8	
11	Fertilizer Rate(kg/h)	140	120	
12	Fertilizer Cost ($/kg)	0.4	0.4	
13	Machinery/Livestock ($/	4	7	
14	Transport ($/h)	2	2.4	
15	Labour Costs ($/h)	20	28	
16				
17				
18	Revenue	7800	8400	
19				
20	Seed Costs	300	720	
21	Fertilizer	1120	960	
22	Machinery/Livestock	80	140	
23	Transport	40	48	
24	Labour	400	560	
25	Total Costs	1940	2428	
26				
27	Profit	$ 5,860.00	$ 5,972.00	
28				

Filling in a series

You can use the Fill series command on the Edit menu to quickly enter sequential values. Simply fill in the first two cells of the row or column you want to fill with the first two entries of your sequence. Then highlight the range you want to fill and click Edit, Fill, Series and OK.

Another way to fill a range with a series is to drag the fill handle of a selection containing the first or first and second series values over the range in which you want the series to be completed.

Paste special

Sometimes we want to copy only the values created by formulae. We cannot do this with a regular Copy and Paste because the cell references change. We must use the Copy and Paste Special commands found in the Edit menu. After selecting Paste Special we see the Paste Special dialogue box. Select Values to extract labels, values and the results of formulae.

Exercise 3

This exercise will not use the spreadsheet we have been creating in the earlier exercises. In this exercise we will look at rice production over a number of years. The data are as follows:

1990	876
1991	964
1992	798
1993	1008
1994	987
1995	1087
1996	1176
1997	1345

- First click on Sheet 2 at the bottom of the page so that the exercise takes place on a clean sheet.
- To enter the years, in cell A1 type 1990 and press enter. Highlight 1990 and seven cells beneath it. Click Edit, Fill, Series, Columns, Okay. The years should automatically be entered.
- Now enter the rice production for each year.
- Using Formulas try and find the sum of rice production, the maximum rice production, and the average rice production: =SUM(B1:B8); =MAX(B1:B8); =AVERAGE(B1:B8)
- Now try and copy average rice production to another cell. Copying the normal way results in a different answer. Use Copy and Edit, Paste Special, Value, to copy the Average as a value.

Graphing

Excel contains a powerful, easy to use graphing facility. You can create charts embedded in the worksheet alongside the data you are working on or as separate chart sheets. All charts are linked to the data they plot. When you change the data, the chart also changes.

To create a chart you must first select the data to plot. The selection should be rectangular and it should not contain blank rows. Leaving a blank cell in the upper left hand corner of the selection tells Excel the data below and to the right of the blank cell contains labels for the values to plot.

After we select the data to plot, there are two ways to create a chart. The first way is to use the Chart Wizard Button  . The mouse pointer then becomes a small chart with a cross above it. Click and hold the right mouse button where you want one corner of the chart to be then drag the mouse till the box is the size desired and release the right mouse button. You will now be able to select new data for your chart, select the type of chart and choose some chart options. After this the chart appears. Using Insert, Chart you go through most of the same steps. With this option you can also select whether the chart appears On This Sheet for an embedded chart or As New Sheet.

Chart attributes can be edited by using the left mouse button to bring up options or by double clicking on the parts of the chart to be changed to bring up options.

Exercise 4

Make a graph comparing the different costs of growing maize and soybean by:

- highlight the costs section o the worksheet (A20:C24)
- click the Graph Wizard button.
- click on a clear section of the sheet where you want your graph to be
- drag the pointer to select the size you want your graph
- keep the same data, click Next
- select a Column Chart, click Next, then select column chart style number 3, click Next
- select Data Series in rows
- click Next, Finish.

Now try to make different styles of graphs. Also see if you can improve the appearance of the graph by changing titles and legends. Move the mouse pointer to any part of the graph and double click the right mouse button to change that feature. Click the left mouse button for a menu of attributes to change.

IF statement

An IF statement is a logical function which sets up a conditional statement to test data. Whether the statement is true or false determines the results of the statement. The format for an IF statement is:

=IF(condition,X,Y)

If the condition is true the function results in X, while if the condition is false, the function results in Y. As an example, a teacher may wish to determine whether a student passed

or failed. A passing grade is 50. An IF statement can be used to determine whether the final grade is greater than or equal to 50. If the condition is true then PASS is entered in the function location. If the condition is false, then FAIL is entered in to the function location. The statement would look like this:

IF(E8>=50, “PASS”,”FAIL”)

IF statements can be used in conjunction with AND, OR or NOT statements to evaluate complex conditions. Also IF statements can be nested to evaluate a number of conditions.

Lookup functions

The lookup functions (VLOOKUP and HLOOKUP) select a value from a table and enter it into a location on the worksheet. For example, VLOOKUP may be used to lookup costs for different crops. The lookup function is entered in the location on the worksheet that requires data from a table. There are two ways to lookup data depending on the way the data are arranged: vertically or horizontally:

VLOOKUP (vertical lookup) looks up data in a particular column in a table;

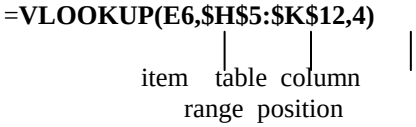
HLOOKUP (horizontal lookup) looks up data in a particular row in a table.

The VLOOKUP function uses the following format and contains three arguments (parts), defined below:

=VLOOKUP(item, table-range, column-position):

- Item is text, a value, or a cell reference of the item you are looking for (search item) and should be the first column of the VLOOKUP table.
- Table-range is the range reference or range name of the lookup table in which the search is to be made.
- Column-position is the column number in the table from which the matching value should be returned. The far left column has a position number of one. The second column has position number of two, etc.

If you need to look up more than one item and copy the lookup formula, the formula should use the cell reference (not the value) as the search item. In addition, the range should be absolute so the table range remains constant:



Exercise 5

Starting in Cell J1, enter the following data:

Crop	Soybean	Cassava	Sweet Potato
Hectares	20	20	20
Yield (kg/ha)	1400	16000	11000
Price ($/kg)	0.3	.03	.05
Seed Rate (kg/h)	45	0	0
Seed Cost ($/kg)	0.8	0	0.0
Fertilizer Rate (kg/h)	120	130	120
Fertilizer Cost ($/kg)	0.4	0.4	0.4
Machinery/Livestock ($/h)	7	18	12
Transport ($/h)	2.4	4.5	4.0
Labour Costs ($/h)	28	36	45

The first and second columns can be copied from the earlier work.

We will write a formula so depending on what crop is entered in cell C4, the data for that crop is automatically entered. To do this we will use nested IF statements. In cell C5 enter:

@IF(C4=”Soybean”,K2,(@IF(C4=”Cassava”,L2,(@IF(C4=”Sweet Potato”,M2,0))))).

We want to copy this statement to cells C6 to C15. If we copy it, however, C4 in the formula changes, and we want it to remain constant. To make C4 an absolute cell reference, edit the formula in cell C5. To edit, make C5 the active cell then hit F2. Now change C4 to $C$4 in the formula:

@IF($C$4=”Soybean”,K2,(@IF($C$4=”Cassava”,L2,(@IF($C$4=”Sweet Potato”,M2,0)))))

Now copy the formula to cells C6 to C15. See what happens when you change cell C4 to “Cassava” or “Sweet Potato”.

Spreadsheets as databases

Sometimes people use spreadsheets as databases. On a small scale this is not a problem as almost everything that a database program does can be duplicated on a spreadsheet. Problems arise when the small amount of data begins to grow. With a large amount of data and more complex uses of it, the limitations of spreadsheets become more and more constricting.

Spreadsheets were made for data manipulation and calculations. Databases are used for organizing, displaying and storing data. Database software has special functions for linking data from different subjects, for displaying data in user-friendly forms, for selecting and sorting data and avoiding unnecessary repetition of data. Using a spreadsheet as a database is like using a spreadsheet as your word processor. It is possible, but it is not efficient.

In agricultural research, we use large databases which combine data on crop production, climate, prices soils, land use, demographics and more. Organizing this data can only be done efficiently using database software package.

Elementary Statistical Methods for Agricultural Research

Siemon Hollema[*]

General concepts

Objective

In agricultural research statistical techniques are used for planning and performing studies and for interpreting their results, approximating unknown quantities, predicting future occurrences, interpreting research literature and classifying objects in different categories. This widespread use of statistics makes it necessary for the agricultural researcher to study and understand basic statistical concepts. This manual will discuss some of the most common statistical techniques used in agricultural research. Much of this manual is based on a small handbook, Elementary Statistical Methods for Foresters, written by Frank Freese and published in 1967 by the US Department of Agriculture Forest Service. It is a clearly written and very understandable handbook for the amateur statistician. Here, it has been made more applicable to socio-economic agricultural research. Hopefully, it will prove helpful as background material in courses on statistical techniques for agricultural research.

To simplify the ground covered by statistical methods, we can say that the basic objectives boil down to:

- the estimation of population parameters i.e. values that characterize a particular population.
- the testing of hypotheses about these parameters.

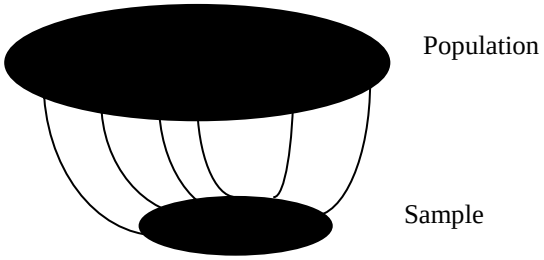
A common example of the first is estimating the coefficient a and b in the linear relationship:

$$Y = a + bX$$

between the variable Y and X. To accomplish this objective, one must first define the population involved and specify the parameters to be estimated. If every member of the population can be investigated and the characteristics of each member determined, then the most accurate picture can be obtained. Unfortunately, this is not very common in agricultural research. Mostly, one has to do work with sample data taken from a population. The next figure makes this clear.

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

Figure 1 Sample data.



The different methods for obtaining sampling data are described in the section on sampling. The population can be described on the basis of sample data. The value of a sample parameter is likely to differ somewhat from the population value. The unique contribution of statistics to research is that it provides ways of evaluating how far off the estimate may be. This is normally done by computing *confidence intervals*. Confidence intervals give the limits that have a known probability of including the true population value of the parameter. Thus, the mean yield of rice might be estimated from a sample as 2.0 ton/ha with 95% confidence limits of 1.8 and 2.2 ton/ha. These limits tell us that, unless a one-in-twenty chance has occurred in sampling, the true yield lies somewhere between 1.8 and 2.2 ton/ha.

The second basic objective of statistics is to test some hypothesis about the population parameters. An example is a test of the hypothesis that the regression coefficient b in the linear model:

$$Y = a + bX$$

has a value of zero. Another example is the test of the hypothesis that the difference between the means of two populations is zero, for example, the mean yield of a new rice variety compared to the mean yield of a traditional rice variety.

These and other most common tests of hypothesis together with the method of estimating some parameters will be the subject of this manual.

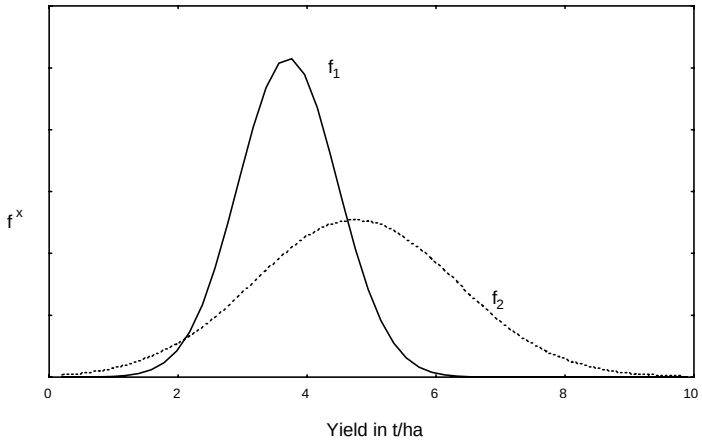
Probability and distributions

As statistics concerns probability, we will start with the role of probability in statistics. Most statistical tests are of the following nature: a hypothesis is formulated and an experiment is conducted or a sample is selected to test it. The next step is to compute the probability of the experiment or sample results occurring by chance if the hypothesis is true. If this probability is less than some predetermined value (perhaps 0.05 or 0.01), the hypothesis is rejected. For example, imagine a game in which you have to throw a six with a dice to win. The chance of throwing a six is 0.1667 provided the dice is honest. Your playmate picks up the dice and throws five sixes in a row. If the chance of throwing a six is 0.1667 then it can be proven that the chance to throw a six five times in a row is as low as 0.00013 (or 1 in 7,776). This is where statistics ends. You can make your own conclusion about the honesty of the dice or playmate. If you conclude that you are cheated, then in statistical terms you are rejecting the hypothesis that the probability of throwing a six is 0.1667. Note that nothing has been proved. Not even that the hypothesis is false. We only infer the latter due to the low probability associated with the experiment or sample results.

Obviously, our inference might be wrong if we are given inaccurate probabilities. Reliable computations of these probabilities require knowledge of how the variable we are dealing with is distributed. Or in other words, what is the probability of chance occurrence of different values of the variable. The most common distribution is the normal distribution. Other

distributions are the Poisson distribution, the binomial distribution, chi-square distribution and the gamma distribution. Figure 2 illustrates normal probability distributions of two grain varieties.

Figure 2 Probability density functions f_1 and f_2.



Observe that although the second variety has a higher mean yield it is also liable to almost total destruction of the yield. Thus, for example if we know that the yield of a grain variety follows a normal distribution then we can compute the probability of a harvest failure. However, if we assume that grain yield follows a normal distribution while in reality is follows a gamma distribution, the computed probability for harvest failure might be completely wrong.

Most statistical tests assume that the variable follows a normal distribution. If we are not sure that this is the case, the particular test should not be applied. A *goodness-of-fit* test can be applied to sample data to see if they actually come from the assumed distribution (see section on chi-square tests).

Even when we have reliable probabilities, statistical tests can still lead to the wrong conclusions. Sometimes a hypothesis that is true will be rejected. If, for example, we always test at the 0.05 level, we will make this mistake once every twenty times. If this degree if risk is unacceptable we can test at the 0.01 or 0.001 level. You may ask yourself, why not always test at the 0.00001 level? We would then only be wrong once every 100,000 times. A researcher, however, can make more than one kind of error. Instead of rejecting a hypothesis that is true (a *type I error*) he can also accept a hypothesis that is false (a t*ype II error*). Unfortunately, reducing the risk for one kind of error increases the risk of the other kind. A critical step in designing experiments is to attain an acceptable level of probability for each type of error. This is usually accomplished by specifying the probability of a type I error and then making the experiment large enough to attain an acceptable level of probability for an error of type II.

Descriptive statistics

Descriptive statistics embody a group of statistical methods used to describe, summarize and present a set of data. It is the first step in data analysis. Using the methods of descriptive

statistics, great insight into the data can be acquired. This is necessary before starting to test hypotheses or build models. Hypotheses for example might change on the basis of these first results. Furthermore, many statistical procedures are based on specific assumptions about the underlying distribution of the data values. Therefore, the statistical techniques we are planning to use have to be tested for their appropriateness. Lastly, survey data are never error free. Mistakes have to be identified and eliminated as much as possible. Also gaps in the distribution of data values, extreme values, strange patterns and unexpected variability may occur. Reasons for their occurrence need to be found.

Data commonly are summarized by using frequency tables and bar charts. For further description of the data set, use is made of a variety of summary measures or statistics.

Data representation: frequency tables

Frequency distributions can be helpful for condensing a large number of values into a more manageable form. A frequency table is simply produced by counting the number of observations within a specific category. For example, the percentage of rice protein in the grains of a F_3 plant derived from a cross between two varieties was measured. There are 50 observations given in Table 1.

Table 1 Percentage of rice protein.

4.6	4.8	6.7	8.8	3.5	7.2	7.3	6.5	9.0	6.0
6.9	6.3	5.2	4.9	5.6	5.5	7.0	6.7	6.8	6.7
5.3	9.5	3.4	3.8	5.0	5.6	6.2	7.3	7.5	8.0
6.4	7.2	5.7	5.9	4.2	3.8	9.3	6.7	6.2	5.8
9.4	3.9	4.3	5.0	5.5	5.8	6.8	6.0	7.4	8.4

The frequency distribution is given in Table 2. We distinguish 8 classes or categories.

Table 2 Frequency table.

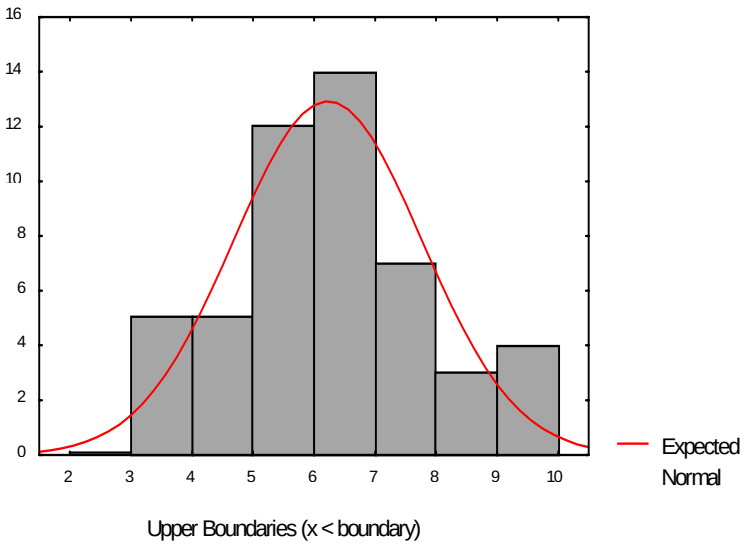
Category	Count	Cumul. Count	Percent of Valid	Cumul. % of Valid
2.00 <= x < 3.00	0	0	0.000	0.00
3.00 <= x < 4.00	5	5	10.000	10.00
4.00 <= x < 5.00	5	10	10.000	20.00
5.00 <= x < 6.00	12	22	24.000	44.00
6.00 <= x < 7.00	14	36	28.000	72.00
7.00 <= x < 8.00	7	43	14.000	86.00
8.00 <= x < 9.00	3	46	6.000	92.00
9.00 <= x < 10.0	4	50	8.000	100.00

Each row represents a specific class. The number of observations in a specific class (the frequency) is shown in the second column. Note that the cumulative count adds up to 50, the total number of observations. The percentage is shown in the third column and the cumulative percentage in the last column.

Histogram

While the numbers in a frequency table can be studied and compared, it is often useful to present the data in a form that can be interpreted visually. This can be done by using a bar chart. The bars are generally labelled by their class while the height of the bars is equal to the class frequency. The bars are centred on the class midpoints. Such bar charts are called histograms. Figure 3 shows the histogram for the data above.

Figure 3 Histogram with normal curve superimposed.



Histograms should not be used to display values when there is no underlying order to the values, *i.e.*, when measured on a nominal scale (see next section). Other useful charts are a *stem-and-leaf* plot or the *boxplot*.

Scale of measurement

The choice of the statistic depends on characteristics of both the data set and the statistic. An important characteristic is the scale of measurement of each variable studied. Traditionally the scale of measurement is divided into nominal, ordinal, interval or ratio. The nominal scale is a categorical scale of measurement where each value represents a specific category that the variable's values fall into (each category is "different" from the others but cannot be quantitatively compared to the others). For example, a number might be assigned to sex by, male≡1, female≡2. Thus, numeric values attached to categories act merely as identifiers. The ordinal scale of measurement represents the ranks of a variable's values. Values measured on an ordinal scale contain information about their relationship to other values only in terms of whether they are "greater than" or "less than" other values but not in terms of "how much greater" or "how much smaller." The interval scale of measurement allows you to quantify and compare the sizes of differences between items that are measured. This can be illustrated by using temperatures. The difference between 20^{o}C and 21^{o}C is the same as the difference between 5^{o}C and 6^{o}C. However, you can not say that it is "twice as hot" on a day when the

temperature is 30°C as when it is 15°C. The interval scale does not have a determined zero point. The ratio scale of measurement contains an absolute zero point. It therefore allows you to not only quantify and compare the sizes of differences between values, but also to interpret both values in terms of absolute measures of quantity or amount (e.g., time; 3 hours is not only 2 hours more than 1 hour, but it is also 3 times more than 1 hour).

The scale of the measurement has to be determined before the appropriate statistical technique can be applied.

Measures of central tendency

The mean, median, and mode are used to describe the location of the distribution. In general, these measures identify a point in the central part of a distribution and are therefore called measures of central tendency.

The mode is the most frequently occurring value(s). The mode for the data in Table 1 is 6.7. A distribution can have more than one mode. A distribution is often characterized by its number of modes i.e. peaks. A unimodal distribution has one peak. A distribution with two peaks is referred to as bimodal, while distributions with three or more peaks are called multimodal. The mode can be measured at any scale although in practice it is seldom used with measurements on the ordinal, interval and ratio scale as it ignores much of the available information. It is useful when dealing with qualitative data.

The median is the value for which one-half (50%) of the observations lie above that value and one-half lie below that value. It divides the distribution exactly in half. The median of the distribution of protein content is 6.2. When there is an even number of observations, no unique center value exists. In this case, the mean of the two middle cases is usually taken as the median value. The median should not be used for nominal data since it uses ranking information.

The mean or arithmetic average is the most commonly estimated population parameter. Given a simple random sample, it can be estimated by dividing the sum of the values of all observations by the number of observations. Thus,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where, n is the number of observations and X_i the value of the i^{th} observation. For the data on rice protein the mean is 6.226. Measurements for calculating the mean should be on interval or ratio scale.

As you see the three measurements for central tendency are not necessarily the same. The mean is greatly influenced by values at either end of the distribution, while the median is not. For symmetric distributions the mean, median and mode are usually close in value. Table 3 summarizes the various types of measurement scales and the appropriate measures of central tendency.

Table 3 Measurement scales and the appropriate measure of central tendency.

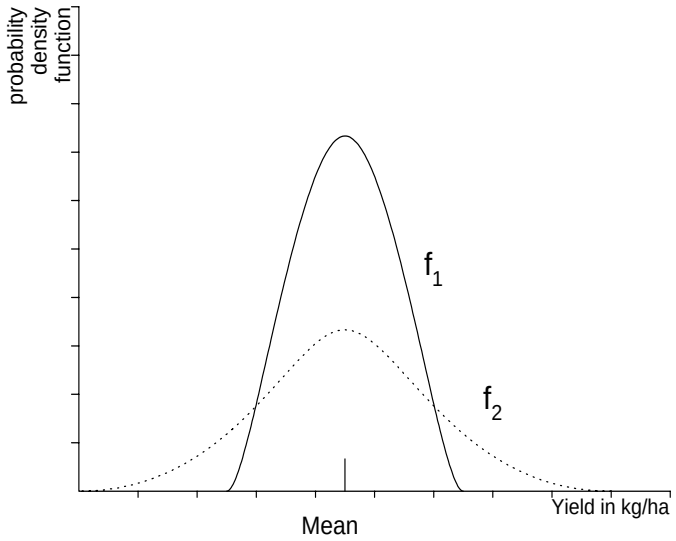
	Appropriate Measure of Central Tendency		
Measurement Scale	Mode*	Median	Mean
Nominal	yes	no	no
Ordinal	no	yes	no
Interval	no	yes	yes
Ratio	no	yes	yes

*Technically the mode may be used on an ordinal, interval and ratio scale. In practice, however, it is not very common.

Measures of dispersion

Distributions can have the same measure of central tendency and yet be completely dissimilar. Figure 4, for example, shows the probability distribution of the yield of two varieties of maize. The mean for both distributions is the same. However, variety f_2 is more risky in the sense that the likelihood of crop failure is greater. A quick and useful measurement of dissimilarity, or dispersion is the r*ange*. It is defined as the difference between the maximum and minimum observed values. The data given for the rice protein content have a range of 6.1. Since the range is calculated using only the maximum and minimum values it is sensitive to extremes. It does not take into account the distribution of values between the maximum and minimum observations. The range is most useful as a measure of variation with ordinal data.

Figure 4 Probability distributions of two varieties of maize.



A measure of variation that does take into account all observations is the *variance*. The variance represents the difference between the observations and the mean. Variance for the *population* and *sample* data is calculated differently. Computer statistical programs and calculators usually assume sample data are distributed normally. The sample variance is represented by S^2 and calculated as follows:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

If all the observations are the same, that is if there is no variation, the variance is equal to 0. The more spread out the observations, the higher the variance. In Figure 4, the variance of f_2 is higher than f_1. The *standard deviation* (S) is the square root of the variance. It is easier to interpret than the variance as the latter is expressed in squared units while the standard deviation is expressed in the same units of measurements as the observations. Some properties of the variance and the standard deviation are presented in Table 4.

Table 4 Properties of variance and standard deviation.

$S_{bX}^2 = b^2 S_X^2$	$S_{bX} = \|b\| S_X$
$S_{X+c}^2 = S_X^2$	$S_{X+c} = S_X$
$S_c^2 = 0$	$S_c = 0$

Note: b and c are constants.

As mentioned above the variance or standard deviation is sometimes used as a measure of 'risk'. This is, however, a very relative term as populations with large means often show more variation than populations with small means. For example, a $1,000 investment may have returns in the range of -$100 to +$150 with a mean of $50 and a standard deviation of $60. A second investment of a larger amount, say $100,000 may have returns in the range of -$5,000 to +$7,500 with a mean of $5,000 and a standard deviation of $1,500. To say that the second investment has a higher risk because it has a higher standard deviation does not take into account the difference in the magnitude of the investment. A measure that facilitates comparison of variability among different sized means is the *coefficient of variation* (*C*). The coefficient of variation is estimated by:

$$C = \frac{s}{\overline{X}}$$

The coefficient of variation for the first investment is 0.83 (or 83%), while the second investment has a coefficient of variation of 0.30 (or 30%).

Measures of shape

Most variables have a distribution which shows a concentration of observations near the middle. As the distance from the central concentration increases, the frequency of observations decreases. Such distributions are often "bell-shaped". By far the most important bell-shaped distribution in statistics is the *normal distribution*. The normal distribution is symmetric and, consequently, the mean, median and mode are the same. In Figure 3 a normal distribution is superimposed on the histogram. It indicates what the distribution of protein content would be if the variable had a normal distribution with the same mean and variance. The normal distribution serves as a reference point in measuring the *skewness* and *kurtosis* of a distribution.

Skewness measures the deviation of the distribution from symmetry. A normal distribution is perfectly symmetrical and has a skewness of 0. If there are more observations with larger values, the distribution is *positively skewed*. If more observations occur towards lower values, the distribution is *negatively skewed*. The distribution of protein content is slightly positively skewed with an index of 0.25.

Kurtosis measures the "peakedness" of a distribution. If the kurtosis is clearly different from 0, then the distribution is either flatter or more peaked than normal; the kurtosis of the normal distribution is 0. If observations cluster more than those in a normal distribution (i.e. the distribution is more peaked) the distribution is called *leptokurtic*. If the observations cluster less (i.e. the distribution is flatter) the distribution is termed *platykurtic*. The index for kurtosis for the distribution in Figure 3 is -0.27.

Sampling

Sampling is the process of investigating a limited number of units selected from a population with the aim of estimating some characteristic of a population. On the basis of sample results, generalizations are drawn about the population without measuring all population units. In order for generalizations to be made, samples must be taken in a *random* manner. There are several random sampling methods. The most basic sampling method is a simple random sample. Other random sample techniques include systematic sampling, stratified sampling and cluster sampling (for discrete variables).

Simple random sampling

A simple random sample is defined as a sample of *n* units from a population of *N* units selected in such a way that every possible combination of *n* units has an equal chance of being selected. Sampling can take place with or without replacement. In agricultural research, sampling without replacement is most common, i.e. once a unit has been included in the sample it can not be selected again.

For example; from a population of a hundred pigs (*N*=100) a sample of 20 units (*n*=20) were selected at random and their gain in weight was measured after a 20 day period (Snedecor 1956). The measurements were:

32	31	11	30	19
24	53	44	19	30
39	34	33	12	21
40	39	17	22	33

The sum of all 20 random units = 583

From this sample the population mean is estimated as:

$$\overline{X} = \frac{\sum X}{n} = \frac{583}{20} = 29.15$$

However, if we took another sample of 20 units from the same population, it might have a mean of 28, or another with a mean of 31, and so forth. Actually, an important theorem, *the central limit theorem*, states that the distribution of the sample means (and other statistics) approaches normal if the sample size increases regardless of the shape of the original population. This can be of great convenience in practical applications where the distribution of the sampled population is unknown.

Standard error

Clearly, it would be desirable to know the variation likely to be encountered among the means of samples from this population. A measure of variation among sample means is the *standard error of the mean*. The computation of the standard error of the mean depends on the manner in which the sample was selected. If sampling is done without replacement, the formula for the estimated standard error of the mean is:

$$s_{\overline{x}} = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

Thus, if we had taken a sample of 20 pigs without replacement, the estimated mean gain in weight (29.15) would have a standard error of:

$$s_{\bar{X}} = \sqrt{\frac{120.45}{20}\left(1-\frac{20}{100}\right)} = 2.195$$

The term $\left(1-\frac{n}{N}\right)$ is called the *finite population correction* (fpc). If the sample is very small (say less than 5%) or if sampling is with replacement, the fpc may be omitted and the standard error of the mean would then simply be:

$$s_{\bar{X}} = \sqrt{\frac{s^2}{n}} = 2.454$$

The standard error of the mean is used to compute confidence intervals for a population mean.

Confidence intervals

Sample estimates are subject to variation. Each sample taken from the same population is likely to produce different sample estimates. How much they vary depends on the variability of the population (σ^2) and on the size of the sample (n) and of the population (N). Establishing a confidence interval is the statistical way to express the reliability of a sample estimate. It gives the limits that have a known probability of containing the true population value of the parameter. For estimates made from a normally distributed population, the confidence interval is given by:

(Estimate) ± (t) (Standard error)

We already have all the information we need to set confidence limits on the mean gain in weight except for the value of *t*. This *t* value can be obtained from the table of the *t* distribution (Appendix 1). In this table the column headed df (degrees of freedom), refers to the size of the sample. For the mean of a simple random sample we would select a *t* value with n-1 degrees of freedom. The columns labeled 'probability' refer to the kind of odds we demand. If we want to say that the true population mean falls within certain limits unless a one-in-twenty chance occurred, we select the *t* value from the column headed .05. If we want to say that the true mean falls within certain limits unless a one-in-hundred chance occurred, we select the column .01.

In the example the sample of n=20 had a mean of 29.15 and a standard error of 2.454 (ignoring the fpc). For a 95% confidence interval on the mean, we would use a *t* value from the .05 column and the row for 19 degrees of freedom. As $t_{.05}$ = 2.093, the confidence interval is given by:

$$\bar{X} \pm (t)(s_{\bar{X}}) = 29.15 \pm (2.093)(2.454) = 24.01 \text{ to } 34.29$$

wich says that unless a one-in-twenty chance has occurred in sampling, the true population mean is somewhere between 24.01 and 34.29. It does not tell you where the sample mean of future samples from this population might fall. Nor does it say where the mean may be if mistakes have been made in the measurements. For a 99% confidence interval the limits would be 22.13 and 36.17. These limits are wider and therefore more likely to contain the true population mean.

Sample size

Taking a sample costs time and money. So do errors. It is important to remember this when planning a survey. The question is, therefore, what should be the size of the sample? The

answer depends on the required precision and the inherent variability of the population being sampled. In statistics sampling precision is normally expressed in terms of a confidence interval on the mean. In planning a survey we could therefore state that the computed confidence interval should be lower or equal to a specified value E unless a one-in-twenty chance has occurred in sampling. That is, we want:

$$t\, s_{\bar{X}} = E$$

or, since $s_{\bar{X}} = \frac{s}{\sqrt{n}}$, we want $t\left(\frac{s}{\sqrt{n}}\right) = E$

Solving this for n gives the desired sample size.

$$n = \frac{t^2 s^2}{E^2}$$

To apply this equation we need three pieces of information. Firstly, we need a specification of the largest confidence limit to be tolerated (E), secondly, a value for the students t, and thirdly, an estimate of the population variance (s^2). The latter might constitute a problem. One solution is to make the sample survey in two stages. In the first stage n_1 observations are made and from these an estimate of the variance is computed. This value is plugged into the equation for the sample size, where the t value has n_1-1 degrees of freedom and is selected from Appendix 1. The computed n is the total size of the sample needed. As the previous sample already contained n_1 observations we have to observe an additional (n - n_1) units. If pre-sampling is not possible we have to make an estimate of the variance the basis of our knowledge of the range to be encountered. If this estimate can be considered reasonably reliable, then for 95% confidence the size of the sample needed to estimate the mean within $\pm E$ units is approximately:

$$n = \frac{4 s^2}{E^2}$$

Less reliable variance estimates may be doubled as a kind of safety factor.

When sampling is without replacement, the sample size estimates above should be adjusted if the sample consists of 5% or more of the population. The adjusted value of n is:

$$n_a = \frac{n}{1 + \frac{n}{N}}$$

Systematic sampling

Suppose we want to draw a sample of 10 units from a numbered population of 100. We might then select a number at random between 1 and 10, say 3, and pick every 10th unit thereafter; i.e. the units numbered 3, 13, 23, and so on. A sample of this kind is called a *systematic sample*, since the choice of its first member determines the whole sample. The main advantage of a systematic sample over a simple random sample is that it distributes the sample more evenly over the population. For this reason it often gives more accurate results then simple random sampling. It has however two drawbacks. If the sampled population contains a periodic type of variation, and if the interval between successive units happens to coincide with this, we may obtain badly biased sample results. For example, a systematic sample of plants in a field might have the selected plants at the same position along every row. The second disadvantage is

that there is no reliable method of estimating the standard error of the sample mean. There are various formulas for $s_{\bar{X}}$ but each formula is only valid for a particular type of population.

Stratified random sampling

In stratified random sampling the population is divided into subpopulations, called *strata*, of known size. A simple random sample is then drawn independently in each subpopulation. The advantage of stratified sampling is that the estimate of the population mean will be more precise than given by a simple random sample of the same size if there is more variation between subpopulations than within them. Another advantage of this method is that it provides separate estimates for each subpopulation.

For example; a survey on percapita income ($/ha) was held in a village. The households were divided into three strata based on the farm size. A simple random sample was done in each stratum and the means, standard deviations and standard errors of the means were computed.

Type	Stratum No.	Stratum size (N_h)	Sample size (n_h)	Stratum mean	Within stratum standard deviation	Standard error of the mean
Large farmers	1	24	4	260	49.67	22.67
Small farmers	2	30	3	112	35.55	19.48
Landless workers	3	41	4	82	9.63	4.57
	Sum	95				

The standard error of the mean for stratum *h* is computed by the formula given for the simple random sample:

$$s_{\bar{X}_h} = \sqrt{\frac{s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)}$$

With these data the population mean is estimated by:

$$\overline{X}_{st} = \frac{\sum N_h \overline{X}_h}{N}$$

where, $N = \sum N_h$

For this example the estimated population mean is thus:

$$\overline{X}_{st} = \frac{24 \cdot 260 + 30 \cdot 112 + 41 \cdot 82}{95} = 136.44$$

The formula for calculating the standard error for the stratified mean is cumbersome but not difficult.

$$s_{\bar{X}_{st}} = \sqrt{\frac{1}{N^2}\left[\sum N_h^2 s_{\bar{X}_h}^2\right]}$$

If the sample size is fairly large, the 95% confidence limits on the stratified mean are approximately given by:

$$\overline{X}_{st} \pm 2\, s_{\bar{X}_{st}}$$

There is however no simple way of computing confidence intervals in case of a small sample as is the case in our example.

Sample size

In stratified sampling we not only have to decide upon the size of the total sample but also choose the sample sizes within the individual strata. The most common procedure is to allocate the sample in proportion to the size of the stratum. Thus if a stratum has half of the units of the total population, we would take half of the samples in that stratum. In proportional allocation, the number of units that are to be selected in stratum *h* is given by:

$$n_h = \left(\frac{N_h}{N} \right) n$$

In the household population discussed above, the proportional allocation of the 11 sample households would be 3 large farmers, 4 small farmers and 4 landless workers.

Some other possibilities are equal allocation and optimum allocation. In equal allocation an equal number of units are selected in each stratum. In optimum allocation an attempt is made to obtain the smallest standard error of the stratified mean possible for a sample of *n* units. This is done by sampling more heavily in the strata having a larger variation. The equation for optimum allocation is:

$$n_h = \left(\frac{N_h s_h}{\sum N_h s_h} \right) n$$

A problem with optimum allocation is that it requires estimates of the within-stratum standard deviations, information that may be difficult to obtain. A refinement of optimum allocation is to take into account sampling cost differences between strata. In this case the sample is allocated in such a way as to get the most information per dollar. If c_h represents the cost per sampling unit in stratum *h* the equation for optimum allocation is:

$$n_h = \left(\frac{\frac{N_h s_h}{\sqrt{c_h}}}{\frac{\sum N_h s_h}{\sqrt{c_h}}} \right) n$$

After we have decided on the method of allocation we can estimate the size of the sample (n) necessary for a specified error (*E*) at a given level of confidence. Proportional allocation is the most common and the size of the sample needed for the population mean to be within ±*E* units of the stratified mean using a 95% confidence level can be approximated by:

$$n = \frac{N\left(\sum N_h s_h^2\right)}{\frac{N^2 E^2}{4} + \sum N_h s_h^2}$$

For example, if we want the mean estimated per capita income to be within ± 15 $/ha unless a one-in-twenty chance occurs in sampling, we define E =15. Given the within stratum standard deviations based on the previous sample, *n* would be 15.8 or 16. These 16 sample units are allocated to the strata using the formula:

$$n_h = \left(\frac{N_h}{N} \right) n$$

giving n_1=4, n_2=5, and n_3=7.

Cluster sampling

The sampling methods discussed so far all apply to data measured on a continuous or nearly continuous scale. However, these methods might not be applicable if units observed are classified as dead or alive, man or female, infected or not infected etc. Data of this kind may follow a *binomial distribution*. When sampling discrete variables the aim is to estimate the proportion or the number of units in a population containing a certain characteristic (attribute) rather than a single unit having or not having that characteristic. For example, suppose that a simple random sample of 100 was taken from a farming population, and 86 of them expressed approval of a certain agricultural programme. The estimated proportion in favour for the agriculture programme is then:

$$\bar{p} = \frac{86}{100} = 0.86 \text{ or } 86\%$$

The confidence interval can easily be obtained from Appendix 1. Look in the number observed (*f*) column for 14 (86 exceeds 50 and therefore read 100 – *f*), and then move crosswise to the column for the sample size 100. The 95% confidence interval is given as 78 and 92 (the confidence limits 8 and 22 subtracted from 100). For samples of a size 250 or 1000 look in the column fraction observed (*f/n*) rather than the number actually observed.

Appendix 1 can also be used to estimate the size of the sample necessary to estimate a population proportion with some specified precision. Suppose, for example, that we wanted to estimate the proportion in favour for the agricultural programme within ± 5% at a 95% confidence level. The first step is to guess roughly what the proportion will be. If a good guess is not possible then set $\bar{p} = 0.50$, ensuring the maximum sample size. Next, pick any of the sample sizes given in the table (10, 15, 20, 30, 50, 100, 250 or 1000) and look at the confidence interval for the specified value of $\bar{p}$. Inspection of these intervals will tell you whether or not the specified precision will be met with a sample of this size or if a larger or smaller sample would be more appropriate. Thus, if we guess $\bar{p} = 0.85$, then in the sample of n =100 we would expect to observe 85 to be in favour of the agricultural programme. The table says that the 95% confidence interval would be 0.76 and 0.91. Since these limits are not within 5%, a larger sample would be necessary. For a sample of 250 the limits are 0.80 and 0.90. Since both are within 5% of $\bar{p}$, a sample of 250 would be adequate.

In case of discrete variables simple random sampling, however, might be difficult or impractical. For example, in estimating how many cassava plants are infected by brown leaf spot we could, instead of selecting individual plants at random, select rows at random and observe all the plants in the selected row. This is an example of *cluster sampling*. The rows selected are called clusters. If the clusters are large enough (>100 individual items per cluster) and nearly equal in size, then the statistical methods described earlier can often be applied. Thus, suppose that the percentage of cassava plants infected is estimated by selecting 10 clusters of 200 plants. The observed proportion of plants infected in each cluster is:

Cluster (*n*)	1	2	3	4	5	6	7	8	9	10	Sum
Percentage infected (*p*)	43.2	35.7	51.2	23.2	68.3	21.0	45.3	38.0	61.4	49.7	437.0

The mean percentage infected is estimated by:

$$\bar{p} = \frac{\sum p}{n} = \frac{437.0}{10} = 43.7\ \%$$

The standard deviation of *p* is:

$$s_p = \sqrt{\frac{\sum p^2 - \frac{\left(\sum p\right)^2}{n}}{n-1}} = 15.076$$

and, ignoring the fpc, the standard error for $\bar{p}$ is:

$$s_{\bar{p}} = \sqrt{\frac{s_p^2}{n}} = 4.767$$

The 95% confidence interval can then easily be calculated by:
$\bar{p} \pm (t_{.05})(s_{\bar{p}})$, *t* having (*n*-1) = 9 degrees of freedom, giving a confidence interval of 32.9 to 54.5.

When calculating confidence intervals this way, we assume that the individual percentages follow something like a normal distribution with homogenous variance (i.e. the percentages have the same variance regardless of their size). In the case of small clusters (say, less than 100 individuals per cluster) or if some of the percentages are more than 80 or less than 20, this assumption may not be valid. Consequently, the computed confidence interval may be unreliable. In such cases it preferable to use the transformation:

$$y = arc\ sine \sqrt{percent}$$

and analyze the transformed variable. The transformation is easily done using Appendix 1, Table 3. If we transform the percentages in the previous example we would get:

Percentage infected	43.2	35.7	51.2	23.2	68.3	21.0	45.3	38.0	61.4	49.7	Sum
$arc\ sine \sqrt{percent}$	41.1	36.7	45.7	28.8	55.7	27.3	42.3	38.1	51.6	44.8	412.1

Then calculating the mean, standard variation and standard error of the transformed variable *y* instead of the percentage *p* will yield the following 95% confidence interval for $\bar{y}$:

$$\bar{y} \pm (t_{.05})(s_{\bar{y}}) = 41.21 \pm (2.262)(2.847)$$
$$= 34.8 \textit{ to } 47.7$$

Searching for these values in Appendix 1 Table 3 gives us the corresponding percentages of 32.6 to 54.7, which do not differ much from our previous results.

Chi-square tests

The chi-square test has many uses. It is most commonly used to test hypotheses concerning the frequency distribution of populations. Here we consider four tests that are common in agricultural research.

Test for independence

One of the most interesting applications of the chi-square is the test for independence. It determines whether two or more variables of classifications are related. For example, an agricultural economist wishes to know if the adoption of newly introduce high yielding rice varieties is affected by the tenure status of the farmers (Gomez and Gomez 1984). In doing so,

the farmers can be classified according to their tenure status and whether they adopt the rice varieties or not. The resulting data form a 3 x 2 contingency table.

	Number of farmers		
Tenure Status	Adopter	Non-adopter	Row total (R)
Owner operator	102	26	128
Share-rent farmer	42	10	52
Fixed-rent farmer	4	3	7
Column total (C)	148	39	
Grand total (G)			187

The question is whether the farmer's adoption of the new rice varieties is independent of tenure status. This can be tested by chi-square. The procedure of applying the chi-square test of independence in an *r x c* contingency table (i.e., *r* rows and *c* columns) is as follows. First the *expected value* of each cell is computed using the equation:

$$E_{ij} = \frac{R_i \cdot C_j}{G}$$

where, E_{ij} is the expected value of the $(i,j)^{th}$ cell, R_i is the total of the i^{th} row, C_j is the total of the j^{th} column, and G is the grand total. The resulting expected values under the hypothesis of independence in our example are:

Tenure Status	Adopter	Non-adopter
Owner operator	101.3	26.7
Share-rent farmer	41.2	10.8
Fixed-rent farmer	5.5	1.5

The chi-square value can then be computed as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where, O_{ij} denotes the observed value of the $(i,j)^{th}$ cell and E_{ij} the computed expected value. For our example, the χ^2 value is:

$$\chi^2 = \frac{(102 - 101.3)^2}{101.3} + \frac{(42 - 41.2)^2}{41.2} + \cdots + \frac{(3 - 1.5)^2}{1.5} = 2.01$$

This result is compared to the tabular χ^2 value from Appendix 1, Table 4, with $(r\text{-}1)(c\text{-}1)$ degrees of freedom. We reject the hypothesis of independence if the computed χ^2 value is larger than the corresponding tabular χ^2 value at the prescribed level of significance. In our example, the tabular χ^2 values with $(r\text{-}1)(c\text{-}1) = (2)(1) = 2$ degrees of freedom are 5.99 at the 5% level of significance and 9.21 at the 1% level. Because the computed χ^2 value is smaller that the tabular χ^2 values, the hypothesis of independence between the adoption of newly introduced rice varieties and the tenure status of the farmer can not be rejected.

In this example we distinguished two classifications: tenure status and adoption status. Such data are also referred to as a *two-way* classification. The test of independence can be extended to more than two classifications (*three-way* classification, and so on). However, in doing so, it may be difficult to formulate meaningful hypotheses.

Test for a hypothesized count

The test for a hypothesized count is a method to test if a set of data conforms to a hypothesized frequency distribution. Assume, for example, that the agricultural economist hypothesized that two-thirds of the farmers adopted the new high yielding rice varieties and one-third did not. Thus, the expected counts in each class are:

Type	Adopter	Non-adopter	Total
Observed (O_j)	148	39	187
Expected (E_j)	124.7	62.3	187

As the observed counts differ from those expected, we might wonder if the hypothesis is false. Or can differences as large as this occur strictly by chance? This can be tested using the chi-square test for a hypothesized count. The χ^2 value depends on the number of classes. If there are more than two classes, the chi-square is:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

where, k is the number of classes, O_j the observed count for the j^{th} class, and E_j the expected count in the j^{th} class if the hypothesis is true.

For two classes the chi-square is:

$$\chi^2 = \frac{(|O_1 - E_1| - 0.5)^2}{E_1} + \frac{(|O_2 - E_2| - 0.5)^2}{E_2}$$

where, $||$ refers to absolute value. In our example there are two classes. Consequently, the chi-square is computed as:

$$\chi^2 = \frac{(|148 - 124.7| - 0.5)^2}{124.7} + \frac{(|39 - 62.3| - 0.5)^2}{62.3} = 12.51$$

Because the computed χ^2 value is greater than the corresponding tabular χ^2 value (1 degree of freedom) at the 1% level of significance, the hypothesis that two-thirds of the farmers are adopters and one-third are non-adopters is rejected.

If the adoption rate had depended on the tenure state, separate tests for each group of farmers would have had to be performed. For example, a test for the hypothesis that 75% of the owner farmers are adopters or the hypothesis that only 50% of the share-rent farmers are adopters.

Test for homogeneity of variance

Many statistical techniques are valid only if the variance is homogenous. For example the *t*-test discussed in the next section assumes that the variance is the same for each group, and so does the *analysis of variance*. Also, the fitting of an unweighted regression as described later in the section on regression analysis assumes that the dependent variable has the same degree of variability for all levels of the independent variable. The chi-square test for homogeneity of variance, commonly known as *Bartlett's* test, offers a means of evaluating the assumption of equal (or homogeneous) variances among different groups. Suppose, for example, that data are collected on grain yield in three different locations. The sample variances are 2.34, 1.72, and 2.80, based on samples of 80, 60, and 65, respectively. The researcher wishes to know if there are differences in yield between the three locations. However, before applying a test for mean

comparison, he first has to establish whether the population variance is the same for each group. The parameters needed for Bartlett's test are:

Group	Sampling variance (s^2)	Degrees of freedom (n-1)	Corrected sum of squares SS (n-1)s^2	Log s^2	(n-1)(log s^2)	$\frac{1}{n-1}$
1	2.339829	79	184.8465	0.369184	29.16555	0.012658
2	1.717916	59	101.3571	0.235002	13.86512	0.016949
3	2.804066	64	179.4602	0.447788	28.65845	0.015625
K=3	Sum	202	465.6638		71.68911	0.045232

From this we compute the pooled within-group variance:

$$\bar{s}^2 = \frac{\sum SS_i}{\sum (n_i - 1)} = \frac{465.66}{202} = 2.305266$$

Then the χ^2 value for the test of homogeneity is computed as:

$$\chi^2 = \frac{2.3026\left[\sum (n_i - 1)\left(\log \bar{s}^2\right) - \sum (n_i - 1)\left(\log s_i^2\right)\right]}{1 + \frac{1}{3(k-1)}\left(\sum \frac{1}{(n_i - 1)} - \frac{1}{\sum (n_i - 1)}\right)}$$

with (k-1) degrees of freedom, where k is the number of groups.

Thus, in our example the χ^2 value is:

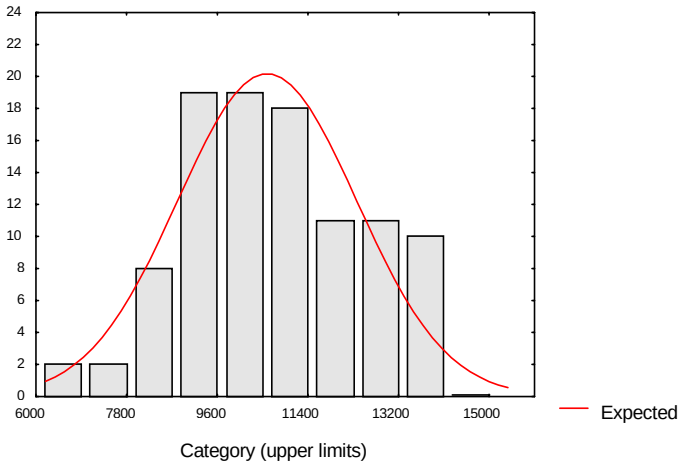
$$\chi^2 = \frac{2.3026\left[(202)(\log 2.305266) - 71.68911\right]}{1 + \frac{1}{3(3-1)}\left(0.045232 - \frac{1}{202}\right)} = 3.6151$$

This value is now compared to the tabular χ^2 value from Appendix Table 4 for the desired probability level. The tabular χ^2 values with 2 degrees of freedom are 5.99 at the 5% level and 6.63 at the 1% level. As the computed χ^2 value does not exceed these values, the hypothesis of homogeneous variance is not rejected.

Test for goodness-of-fit

The test for goodness-of-fit determines whether a set of observed data actually comes from the assumed distribution. For example, data on rice yield may be suspected of following a normal distribution and the spatial distribution of weeds in a field may be suspected to have a *Poisson* distribution. In a study sugar yield from sugar beet (kg/ha) was measured. Sugar yield was expected to follow a normal distribution. However, visual examination of the data shows a considerable deviation from the superimposed normal distribution (Figure 5).

Figure 5 Sugar yield from sugar beet.



The researcher wishes therefore to verify whether, despite the deviations, it is still possible that the sample comes from a normal distributed population. For this purpose, the chi-square test for goodness-of-fit can be applied. The first step is to find the sample mean and standard deviation. These values are $\overline{X} = 10590$ and s = 1,780. The sample size was 100. The number of observations was divided into 10 classes with a class range of 900. We now would like to know the expected frequencies in each class based on the hypothesized normal probability distribution. This is done by computing for each class two standardized Z values, one for the lower class limit (Z_l) and another for the higher class limit (Z_h):

$$Z_l = \frac{L_l - \overline{X}}{s}$$

$$Z_h = \frac{L_h - \overline{X}}{s}$$

where, L_l and L_h are the lower and upper class limits of each class. The lower class limit of the first class is -∞ and the upper limit of the last class is +∞. Thus, for example, Z_l and Z_h for the second class are computed as:

$$Z_l = \frac{6900 - 10590}{1780} = -2.073$$

$$Z_h = \frac{7800 - 10590}{1780} = -1.567$$

The standardized values for all 10 classes are shown in the third and fourth columns of Table 5.

Table 5 Goodness-of-fit test.

Class (upper boundary)	Observed Frequency (f_i)	Standardized Z values Z_l	Standardized Z values Z_h	Probability (P_i)	Expected Frequency $F_i=(n)(P_i)$	$\frac{(f_i-F_i)^2}{F_i}$
≤ 6900	2	- ∞	-2.073	.0191	1.91	0.004241
7800	2	-2.073	-1.567	.0395	3.95	0.962658
8700	8	-1.567	-1.062	.0855	8.55	0.03538
9600	19	-1.062	-0.556	.1450	14.50	1.396552
10500	19	-0.556	-0.051	.1906	19.06	0.000189
11400	18	-0.051	0.455	.1957	19.57	0.125953
12300	11	0.455	0.961	.1563	15.63	1.371523
13200	11	0.961	1.466	.0970	9.70	0.174227
14100	10	1.466	1.972	.0470	4.70	5.976596
infinity	0	1.972	+ ∞	.0243	2.43	2.43
Sum	100				100	χ^2=12.47732

Next, the probabilities associated with each class interval are determined as:

$$P = P(Z_l \langle X \langle Z_h)$$

where the term $P(Z_l \langle X \langle Z_h)$ refers to the probability that X lies between Z_l and Z_h. The probability associated with each class can be determined from Appendix 1, Table 5 by reading the area under the standardized normal curve between Z_l and Z_h. The accuracy of the Z values given in the table is only to 2 decimals. If the computed Z values have more than two decimals interpolation is necessary. For example, in determining the area from -∞ to 1.466, we first obtain from Appendix 1, Table 5 the area under the curve up to 1.46 (=0.9279) and 1.47 (=0.9292), and then through linear interpolation compute the area up to 1.466 as:

$$0.9279 + \frac{(0.9292 - 0.9279)(1.466 - 1.46)}{(1.47 - 1.46)} = 0.9287$$

The area between two Z values carrying the same sign is equal to the area from -∞ to Z_h minus the area from -∞ to Z_l. If the signs are negative the Z values swap place when making them positive. For example, the area between the two Z values of the second class is computed as:

$$P(-2.073 \langle X \langle -1.567) = P(1.567 \langle X \langle 2.073)$$
$$= 0.9809 - 0.9414 = 0.0395$$

If the Z values carry different signs, the negative Z value is treated as positive and the associated probability is then subtracted from 1. For example, the class with the upper boundary of 11400 has a lower and upper Z value of –0.051 and 0.455, respectively. Then the area between the two Z values is computed as:

$$P(-0.051 \langle X \langle 0.455) = 0.6754 - (1 - 0.5203) = 0.1957$$

The computed probabilities for all classes are shown in the fifth column of Table 5. In the next step the expected frequency for each class is calculated as the product of the total number of observations (100) and the probability associated with each class. Thus,

$$F_i = (n)(P_i)$$

The expected frequencies are presented in the sixth column of the table. Finally, the χ^2 value for the test for goodness-of-fit is computed as:

$$\chi^2 = \frac{\sum_{i=1}^{k}(f_i - F_i)^2}{F_i}$$

with (k-3) degrees of freedom, where k is the number of classes. If the computed χ^2 value exceeds the tabular χ^2 value from Appendix 1 Table 4 at the prescribed level of significance, the hypothesis that the sample comes from a normally distributed population is rejected. The computed χ^2 value is 12.48. Compare this with the tabular χ^2 value of 14.07 with 7 degrees of freedom at the 5% level and we can conclude that the hypothesis of a normal distribution can not be rejected.

A somewhat more powerful procedure for testing about the distributional form for continuous random variables is the *Kolmogorov-Smirnov* test. This test is based on comparing the cumulative distribution function of the hypothesized distribution with the sample cumulative distribution function. The chi-square test in contrast compares sample data with the probability density function.

t-Test

The *t*-test is the most commonly used method to evaluate the difference between means of two groups. Often the groups will represent types of treatment that we wish to compare. For example, we are interested in differences in yield produced by fertilizers or differences in gains in weight produced by feeds. The question is not whether the two sample means ($\overline{X}$) are equal but whether the two population means are equal. Under certain assumptions, e.g. the population is normal distributed and the population variances are the same, the *t*-test can be applied. This test has a variety of applications. Here we will confine our attention to test the hypothesis that there is no difference between treatment means. The computational routine depends on how the observations have been selected or arranged, i.e. paired (dependent) or unpaired (independent).

Paired

With paired observations, pairs of similar individuals are selected and one member of every pair is chosen at random to receive the first treatment while the other member receives the second. Pairing is done when we can distinguish two individuals which behave alike (apart from random variation) when treated the same. If individuals do not behave alike, it can not be known if the differences in response are due to difference in treatment or to other causes. The more the individuals differ, the more pairs will be required to balance out the random differences, leaving clear the effect of the treatments. Two observations on the same individual would always be paired. For example, data would be paired in comparing the mean crop yields in successive seasons or in different locations. For illustration of a *t*-test with paired data, consider the following example. The effect of spraying was evaluated on 14 farms by measuring the maize yield in bushels per acre from both sprayed and unsprayed strips in each field (Snedecor 1956). The results were as follows:

Sprayed	64.3	78.1	93.0	80.7	89.0	79.9	90.6	102.4	70.7	106.1	107.4	74.0	72.6	69.5
Unsprayed	70.0	74.4	86.6	79.2	84.7	75.1	87.3	98.8	70.2	101.1	83.4	65.2	68.1	68.4
Difference	-5.7	3.7	6.4	1.5	4.3	4.8	3.3	3.6	0.5	5.0	24.0	8.8	4.5	1.1

We wish to test the hypothesis that there is no difference between the means of sprayed maize yield and unsprayed maize yield. In doing so, first the difference in response of each pair of observations is calculated (d). The t value with (n –1) degrees of freedom can then be computed as:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where, $\bar{d}$ is the mean difference, s_d the standard deviation of the individual differences, and n the number of pairs of observations. So, in this example we find:

$$t = \frac{4.7}{6.48 / \sqrt{14}} = 2.72$$

with 13 degrees of freedom. Comparing this to the tabular value of t from Appendix 1 Table 1 for the required level of significance, we find that the difference is significant at the 0.05 level ($t_{.05}$= 2.160). Consequently, the hypothesis that sprayed and unsprayed maize yield have the same mean is rejected at the 0.05 level of significance. Observe, however, that the hypothesis will not be rejected at the 0.01 level.

How many pairs (replicates) should be used? If too small a number is observed, the test may fail to detect a difference that is important; observing too many on the other hand is wasteful in terms of both time and money. This question is answered by specifying the smallest mean difference (D) worth knowing and solving the above equation for n.

$$n = \frac{t^2 s_d^2}{D^2}$$

The variance of the difference (s_d^2) is usually estimated from a previous experiment or from knowledge of the range to be encountered.

Unpaired

With unpaired sampling, individuals are assigned at random to two groups and then one of the treatments is applied to each group. Assume for example that a second test was made of the maize yield under the different treatments. This time 8 farms were selected at random and all strips in the field were sprayed. At the remaining 6 farms, maize is not sprayed. We can now distinguish two treatment groups, one with sprayed maize and one with unsprayed maize. Again, the yield in bushels per acre was measured. The results were as follows:

Sprayed		
67.8	101.9	87.1
85.9	109.3	88.7
70.3	64.7	
		n = 8
		Mean = 84.5
		Stand. Dev. = 16.11

Unsprayed		
76.7	74.3	80.4
88.6	93.7	62.4
		n = 6
		Mean = 79.4
		Stand. Dev. = 11.07

To test the hypothesis that there is no difference between the sprayed and unsprayed population means we compute:

$$t = \frac{\overline{X}_s - \overline{X}_u}{s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_u}}}$$

where, $\overline{X}_s$ and $\overline{X}_u$ are the mean yields for sprayed and unsprayed maize, respectively, n_s and n_u the number of observations in each group, and s_p the pooled within-group standard deviation. The pooled within-group standard deviation is computed as:

$$s_p = \sqrt{\frac{(n_s - 1)s_s^2 + (n_u - 1)s_u^2}{(n_s - 1) + (n_u - 1)}}$$

Thus,

$$s_p = \sqrt{\frac{(8-1)16.11^2 + (6-1)11.07^2}{(8-1)+(6-1)}} = 14.23$$

The *t* value is therefore computed as:

$$t = \frac{84.5 - 79.4}{14.23\sqrt{\frac{1}{8} + \frac{1}{6}}} = 0.664$$

This value of *t* has (n_s-1) + (n_u-1)=12 degrees of freedom. If it exceeds the tabular value of *t* from Appendix 1, Table 1 at a specified probability level, the hypothesis of equal treatment means would be rejected. For a 0.05 probability level, the tabular value of *t* is 2.179. Since the computed *t* value is less than this, the mean yield difference between sprayed and unsprayed maize is not significant.

In computing the *t* value we assumed that each group had the same population variance. This assumption can be checked with Bartlett's test for homogeneity of variance. If the assumption of equal variance can not be met, a variation of the *t*-test may be used. The *t* value would then be computed as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This *t* value has a certain number of degrees of freedom, which can be approximated by:

$$df = \frac{(a_1 + a_2)^2}{\frac{a_1^2}{(n_1 - 1)} + \frac{a_2^2}{(n_2 - 1)}}$$

where, $a_1 = \frac{s_1^2}{n_1}$ and $a_2 = \frac{s_2^2}{n_2}$. The number of degrees of freedom obtained will generally not be an integer, but can be rounded to an integer value. Other statistical textbooks may suggest different approximations.

To answer the question what should the size of the sample be in order to detect a difference *D* between the two treatment means, we first assume that the number of observations in each group is the same (n_s=n_u=n). The equation for *t* can then be written as:

$$t = \frac{D}{\sqrt{\frac{2s_p^2}{n}}} \quad \text{or} \quad n = \frac{2t^2 s_p^2}{D^2}$$

The within-group variance must be estimated from previous experiments, or from knowledge of the range encountered. Suppose for example that we plan to test at the 0.05 level and want to detect a mean yield difference of D = 20.0. From the previous test we found a pooled within-group variance of:

$$s_p^2 = (14.23)^2 = 202.49$$

Thus, we have:

$$n = \frac{2t^2 s_p^2}{D^2} = 2t^2\left(\frac{202.49}{20.0}\right)$$

Here we have a problem. In order to estimate n we need a value for t, but the value of t depends on the degrees of freedom which in turn depends on n. The only way to solve this problem is by trial and error. Say, we start with a guess that n_0=20. As t has $(n_s\text{-}1)+(n_u\text{-}1)=2(n\text{-}1)$ degrees of freedom, we find a t value of 2.02, and compute:

$$n_1 = 2(2.02)^2\left(\frac{202.49}{20.0}\right) = 83$$

Consequently, the proper value of n lies somewhere between 20 and 83 – much closer to 83 than to 20. Now, we can make a second guess at n and repeat the process. If we try n_2=80, t has 158 degrees of freedom and its value is therefore 1.98. Hence:

$$n_3 = 2(1.98)^2\left(\frac{202.49}{20.0}\right) = 79.3$$

It appears that n = 80. Consequently, we need a total of 160 observations, 80 for each treatment, to detect a difference in mean maize yield of 20 bushels per acre.

It is obvious from the example above that the unpaired test is less sensitive than the paired test, i.e. the unpaired test is less capable of detecting smaller differences between the treatment means. There are, however, many situations in which pairing is not practical. Pairing is done whenever the experimental units can be grouped into pairs such that the variation between pairs is appreciably larger than the variation within pairs. In the example above, there might be considerable variation in yield at the different farms due to soil quality, temperature, moisture, etc. With pairing, variation in maize yield due to these factors is cancelled out, leaving only the variation due to treatment (spraying or not spraying). In an unpaired test, this variation is dealt with by randomly assigning farms to one of the two groups. This, however, leads to a less sensitive test.

Analysis of variance

In this section we will consider the more general problem of comparing means of more than two populations. The statistical methods used are based on analyzing the variation in the observed values within groups as well as the variation between the group means. Based on these two estimates of variation, conclusions are drawn about the population means, and hence, the name *analysis of variance* (abbreviated *ANOVA*). The analysis of variance procedure is based on two assumptions: (i) each of the groups is an independent random sample from a normal population, and (ii) all groups have the same variance.

Here we cover two analysis of variance procedures: o*ne-way ANOVA* and *simple factorial ANOVA*. We start with the one-way ANOVA procedure. One-way analysis of variance is needed when only one variable is used to classify cases into different groups. Examples of one-way classifications are: crop variety trials in which groups are classified by different crop varieties, fertilizer trials in which groups are classified according to the rate of fertilizer applied, and farm size trials where groups are formed according the size of the farms, etc. When two or more variables are used to form groups, the factorial analysis of variance procedure is required. Factorial experiments take account of the interaction between the different variables, i.e. the response of an individual to one variable might be influenced by the level of the other variables. For example, maize yields following three rates of nitrogen fertilization might also depend on the amount of phosphorus used along with the nitrogen.

One-way ANOVA

A farmer wanted to compare the effects of five rates of nitrogen (kg/ha) on grain yield of rice. He laid out 25 plots and randomly applied each rate of nitrogen to 5 plots. The grain yields measured (t/ha) are as follows:

	Treatment: Nitrogen rate (kg/ha)					
	0	60	90	120	150	
	(N_1)	(N_2)	(N_3)	(N_4)	(N_5)	
	4.891	5.763	6.712	6.458	5.683	
	2.577	6.625	6.801	6.830	6.597	
	4.541	5.672	6.799	6.675	6.868	
	3.653	6.009	6.154	6.265	5.937	
	4.187	5.934	6.693	6.636	5.692	
Sum	19.849	30.003	33.159	32.864	30.777	146.652
Sample mean	3.970	6.001	6.632	6.573	6.155	5.866

Looking at the data, we see that there are differences among the treatment means. Treatments N_3 and N_4 seem to have higher averages than N_1, N_2 and N_5, and especially treatment N_1 seems to be a lot lower than the others. The question is can differences in the sample means as large as these occur strictly by chance if there is actually no real difference among treatments? Or, is it reasonable to believe that the treatments make a difference in the grain yield and the samples therefore come from populations that have different means? Problems like this are neatly handled by a statistical technique called *analysis of variance*. To make this analysis, it is convenient and customary to use a so-called analysis of variance table (or ANOVA table).

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean squares (*MS*)	F ratio
Treatments	t-1	*SSTR*	$MSTR = \frac{SSTR}{t-1}$	$\frac{MSTR}{MSE}$
Error	$t(n$-1)	*SSE*	$MSE = \frac{SSE}{t(n-1)}$	
Total	nt-1	*SST*		

Source of variation – There are several reasons why the grain yield of the 25 plots might differ. However, only one source of variation can be identified and evaluated, namely the variation attributable to the different rates of nitrogen. The unidentified variation is assumed to represent the variation inherent in the experimental material and is labeled error. Thus, in analysis of variance the observed variation is divided into two parts: one part is attributable to treatments and the other part is unidentified variation and is called error.

Degrees of freedom – The number of degrees of freedom is not difficult to determine. For the total, the degrees of freedom are the total number of observations minus one. In case the number of observations in each group is the same (n), the degrees of freedom for the total can be calculated as (nt-1). There are 25 plots, consequently the total has 24 degrees of freedom. For the treatments, the degrees of freedom are one less than the number of classes or groups distinguished (t). There are 5 groups (5 treatments), so there are 4 degrees of freedom for treatments. The remaining degrees of freedom (24 - 4 = 20) are associated with the error term. If n is the same in each group, the degrees of freedom for error can also be determined by $t(n\text{-}1)$.

Sums of squares – To calculate the sum of squares associated with each source of variation, we first need to know the so-called 'correction term' (*CT*). This is simply:

$$CT = \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{t} X\right)^2}{N} = \frac{146.652^2}{25} = 860.272$$

where, N is the total number of observations and $\sum_{i=1}^{n}\sum_{j=1}^{t} X$ the sum of all observations. The total sum of squares can then be computed as:

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{t} X^2 - CT = \left(4.891^2 + 2.577^2 + \ldots + 5.692^2\right) - 860.272 = 29.41$$

and the sum of squares attributable to treatments is:

$$SSTR = \sum_{j=1}^{t} \frac{\left(treatment\ totals\right)^2}{n_j} - CT = \left(\frac{19.849^2}{5} + \ldots + \frac{30.772^2}{5}\right) - 860.272 = 23.92$$

The sum of squares for error is obtained simply by subtracting sum of squares for treatments from the total sum of squares:

$$SSE = SST - SSTR = 29.41 - 23.92 = 5.49$$

Mean squares – The mean squares are now calculated by dividing the sum of squares by the associated degrees of freedom. The resulting mean squares are shown in the following analysis of variance table.

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean squares (*MS*)	F ratio
SSTR	4	23.92	5.98	21.79
Error	20	5.49	0.275	
Total	24	29.41		

F Ratio – We have now two estimates of the variation in the population: the treatments mean square (MSTR) and the error mean square (MSE). The first one measures variations between treatments (variation among group means) and MSE measures how much the observations in each group vary (the variation not attributable to the treatments). If the hypothesis is true that the population means of the five groups are equal, we expect no real

differences between the group means, apart from natural variation, and thus, the two estimates should be close to each other. If you divide one by the other, the ratio should be close to 1. The ratio of the treatments mean square by the error mean square is called an F statistic. In this example,

$$F = \frac{MSTR}{MSE} = \frac{5.98}{0.275} = 21.79$$

This number appears in the last column of the table. It certainly doesn't appear to be close to 1. To see if the hypothesis of equal means can be rejected, the computed *F* value is compared to the tabular *F* value from Appendix 1, Table 6. In this table, the columns correspond to the degrees of freedom for treatments and the rows correspond to the degrees of freedom for error. Thus, looking in the column headed 4 and then following down the column to the row labelled 20, we find a tabular *F* value of 2.87 for the 0.05 significance level and a value of 4.43 for the 0.01 level. As the computed *F* value exceeds 4.43, we conclude that the grain yield differences due to different nitrogen levels is significant at the 0.01 level and thus, the hypothesis that there is no difference between treatment means is rejected.

Multiple comparisons

The result of the above test only tells us that the population means of the five treatments are probably not all equal. It does not indicate, however, which pairs of groups have a different mean. In general, a farmer will not be satisfied knowing that differences between treatment means exist. He wants to know specifically which means differ from others. To establish this, special tests, called multiple comparison procedures, are needed. There are many multiple comparison procedures available (Winer et al. 1991). Some of these are Bonferroni's method, Duncan's multiple range test, Turkey's method for multiple comparisons, and Scheffe's method for multiple comparison. Here we will briefly discuss Scheffe's test, which is a method for posthoc comparison of means. Thus, after it is established that differences between population means exist, we might want to test if treatment N_1 differs significantly from treatment N_2, N_3, etc., that is, to compare all treatments pairwise. This is the same as testing if the linear contrasts:

$$\hat{Q} = \overline{N}_1 - \overline{N}_2$$
$$\hat{Q} = \overline{N}_1 - \overline{N}_3$$
$$\vdots$$
$$\hat{Q} = \overline{N}_4 - \overline{N}_5$$

differ significantly from zero. ($\overline{N}_1$ is the mean for treatment N_1, etc). There are other comparisons that might also be made. Suppose, for example, that treatment N_1, N_2 and N_3 involved some mechanical form of harvesting while rice under treatment N_4 and N_5 was harvested manually. The farmer might then want to test whether the average of N_1, N_2 and N_3 differs from the combined average of N_4 and N_5. This would be the same as testing whether the linear contrast:

$$\hat{Q} = \left(2\overline{N}_1 + 2\overline{N}_2 + 2\overline{N}_3\right) - \left(3\overline{N}_4 + 3\overline{N}_5\right)$$

differs significantly from zero. Notice that the coefficients add up to zero and that they are selected in such a way that the three means in the first group are on an equal basis with the two means in the second group. Using Scheffe's test, any linear contrast can be tested by computing:

$$F = \frac{n\hat{Q}^2}{k\left(\sum a_i^2\right)(MSE)}$$

where, n is the number of observations in each treatment, k the number of degrees of freedom for treatment, and a_I the coefficients. This figure is then compared to the tabular F value with the respective number of degrees of freedom for treatment and error. For example, in the pairwise comparison of treatment N_1 against treatment N_2, we would have:

$$\hat{Q} = \overline{N}_1 - \overline{N}_2 = 19.849 - 30.003 = -10.154$$

and

$$F = \frac{5(-10.154)^2}{4\left(1^2 + (-1)^2\right)(0.275)} = 234.33$$

with 4 and 20 degrees of freedom. This figure is much larger than the tabular F value and thus the hypothesis that the mean for treatment N_1 does not differ from the N_2 mean is rejected. In comparing treatments N_1, N_2, and N_3 with the treatments N_4 and N_5 we would have:

$$\hat{Q} = 2(19.849) + 2(30.003) + 2(33.159) - 3(32.864) - 3(30.777) = -24.901$$

and the F value is computed as:

$$F = \frac{5(-24.901)^2}{4\left(2^2 + 2^2 + 2^2 + (-3)^2 + (-3)^2\right)(0.275)} = 93.948$$

In comparing this value to the tabular F value, we would reject the hypothesis that the average mean together for treatment N_1, N_2 and N_3 is equal to the combined average of treatment N_4 and N_5. This is mainly due to the lower mean of treatment N_1.

If the number of observations for each treatment is not equal, the F value in Scheffe's test is computed as:

$$F = \frac{\hat{Q}^2}{k\left(\sum \frac{a_i^2}{n_i}\right)(MSE)}$$

It can be difficult, however, to find appropriate coefficients (a_i). For testing the hypothesis that there is no difference between the means of two groups of treatments, the positive coefficients are usually,

$$positive \quad a_i = \frac{n_i}{p}$$

where, p is the total number of observations in the group of treatments with positive coefficients. Similarly, the negative coefficients are:

$$negative \quad a_i = \frac{n_i}{o}$$

where, o is the total number of observations in the group of treatments with negative coefficients. To illustrate, suppose that the 25 plots were divided over the five treatments as follows: N_1 4 plots, N_2 6 plots, N_3 5 plots, N_4 3 plots, and N_5 7 plots. If we wanted to compare

the means of the treatments N_1, N_2 and N_3 with the means of the treatments N_4 and N_5, then $p = 4 + 6 + 5 = 15$, and $o = 3 + 7 = 10$. The contrast is then:

$$\hat{Q} = \left(\frac{4}{15}\overline{N}_1 + \frac{6}{15}\overline{N}_2 + \frac{5}{15}\overline{N}_3\right) - \left(\frac{3}{10}\overline{N}_4 + \frac{7}{10}\overline{N}_5\right)$$

One might wonder why in comparing all possible pairs of means we do not just use a *t*-test. The reason is that when we make many comparisons involving the same means, the probability increases that one mean will turn out to be significantly different from another. The more comparisons we make, the more likely it is that we will find one or more pairs to be statistically different, even if all population means are actually equal. Multiple comparison procedures adjust for the number of comparisons we are making. The more comparisons we make, the larger the difference between pairs of means must be for a multiple comparison procedure to find it significant. Therefore, we can be more confident in finding the true differences when using a multiple comparison procedure instead of a *t*-test.

Simple factorial ANOVA

In one-way analysis of variance we looked at the effect of one variable (nitrogen rate) on another variable (grain yield of rice). Often, however, it is desirable to consider the effect of two or more variables (also called factors) at a time on a response variable. Examples are comparing yield response of three different rice varieties (factor *a*) following four rates of pesticide use (factor *b*), or, comparing labour productivity on three types of farms (factor *a*) using different levels of mechanization (factor *b*). If there are *a* levels of factor A and *b* levels of factor B, then a complete study of all levels of factor A with all levels of factor B will include $a \cdot b$ different factor combinations. It is possible to think of each combination of factor levels as a treatment and then analyze the problem as a one-way analysis of variance problem. To illustrate, say that the measurements of the effect of nitrogen rates on grain yield of rice in the previous example were taken during the dry season. Now suppose that the farmer also wanted to take into account the effect of seasons and so the test was repeated during wet season. The results were:

	Factor A: Season									
	Dry season					Wet season				
Factor B: Nitrogen rate (kg/ha)	0 (N_1)	60 (N_2)	90 (N_3)	120 (N_4)	150 (N_5)	0 (N_1)	60 (N_2)	90 (N_3)	120 (N_4)	150 (N_5)
	4.891	5.763	6.712	6.458	5.683	4.999	6.351	6.071	4.818	3.436
	2.577	6.625	6.801	6.830	6.597	3.503	6.316	5.969	4.024	4.047
	4.541	5.672	6.799	6.675	6.868	5.356	6.582	5.893	5.813	3.740
	3.653	6.009	6.154	6.265	5.937	4.561	5.989	4.920	5.323	3.317
	4.187	5.934	6.693	6.636	5.692	3.843	6.482	5.751	4.210	4.153
Sum	19.849	30.003	33.159	32.864	30.777	22.262	31.720	28.604	24.188	18.693
Sample mean	3.970	6.001	6.632	6.573	6.155	4.452	6.344	5.721	4.838	3.739

In using the one-way ANOVA procedure we can now distinguish $2 \cdot 5 = 10$ treatments and proceed as before. It is, however, of more interest to treat the factors separately. In this way, we could identify what the effect of nitrogen rate is on the yield, what the effect of the two seasons is on the yield, and if there is interaction between the two factors. The statistical technique to evaluate these questions is a refinement of the one-way analysis of variance. The analysis of variance table for this study is:

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean squares (*MS*)	F ratio
Factor A	a-1	SSA	$MSA = \frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Factor B	b-1	SSB	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
Interaction	(a-1)(b-1)	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{MSE}$
Error	ab(n-1)	SSE	$MSE = \frac{SSE}{ab(n-1)}$	
Total	nab-1	SST		

The first column lists the sources of variation. The total observed variation in the grain yield of rice is subdivided into four components: variation due to factor A (season), factor B (nitrogen rate), their interaction, and unidentified variation dubbed error. The degrees of freedom for factor A and B, listed in the second column, are one less than the number of classes. For example, since there are two seasons, factor A has one degree of freedom. Similarly, factor B has 4 degrees of freedom since there are 5 levels of nitrogen. The degrees of freedom associated with the interaction term are determined by the product of the degrees of freedom of each of the individual factors (i.e. 1·4=4). The total degrees of freedom is one less than the total number of observations (49). The remaining degrees of freedom are associated with the error term (49-9=40). If each class has an equal number of observations, it can also be calculated as $ab(n-1)$. The sums of squares attributable to each of the four components listed in the third column of the table are computed in a similar way as done in the one-way ANOVA procedure. We start by computing the correction term:

$$CT = \frac{\left(\sum_{i,j,k} X\right)^2}{N} = \frac{272.119^2}{50} = 1480.975$$

where, subscript i represents the season, j the level of nitrogen and k the sample observation. The term $\sum_{i,j,k} X$ is thus the sum of all observations. The total sum of squares is then computed as:

$$SST = \sum_{i,j,k} X^2 - CT = \left(4.891^2 + 2.577^2 + \ldots + 4.153^2\right) - 1480.975 = 65.832$$

and the sums of squares for seasons, nitrogen, and the seasons-nitrogen interaction can be computed as follows:

$$SSA = \sum_i \frac{(season\ total)^2}{no.of\ obs.\ per\ season} - CT = \frac{146.652^2}{25} + \frac{125.467^2}{25} - 1480.975 = 8.976$$

$$SSB = \sum_j \frac{(nitrogen\ total)^2}{no.of\ obs.\ per\ level\ of\ nitrogen} - CT = \frac{42.111^2}{10} + \ldots + \frac{49.47^2}{10} - 1480.975 = 29.019$$

$$SSAB = \sum_{ij} \frac{(treatment\ total)^2}{no.of\ obs.\ per\ treatment} - CT - SSA - SSB$$

$$= \frac{19.849^2}{5} + \ldots + \frac{18.693^2}{5} - 1480.975 - 8.976 - 29.019 = 16.105$$

The sum of squares attributable to the error term is then easily obtained as:

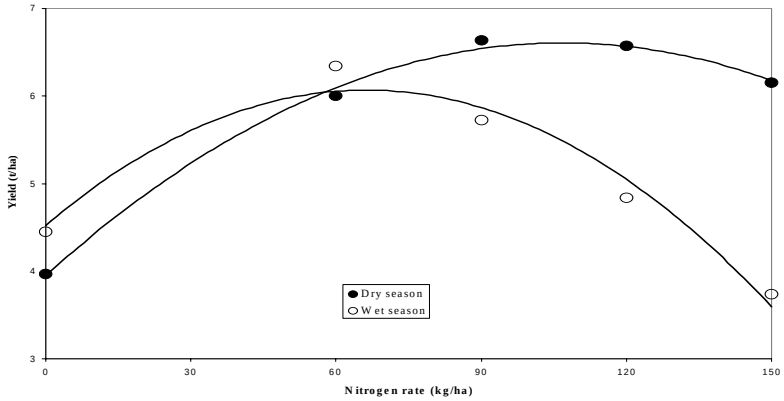
$$SSE = SST - SSA - SSB - SSAB = 11.731$$

The mean squares shown in the fourth column of the following completed analysis of variance table are obtained by dividing each sum of squares by its degrees of freedom.

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean squares (*MS*)	F ratio
Factor A	1	8.976	8.976	30.607
Factor B	4	29.019	7.255	24.738
Interaction	4	16.105	4.026	13.729
Error	40	11.731	0.293	
Total	49	65.832		

The hypothesis that all treatment means are equal is only true if and only if there is no significant effect from seasons, nitrogen and interaction. For each of these an *F* value is computed by dividing the mean squares of each source of variation by the mean square of error. In comparing the computed *F* values with the tabular *F* values from Appendix Table 6, we see that they are all significant at the 0.01 level. Thus, at a first look, it appears that grain yield responds significantly both to crop season and to nitrogen application. However, the F value associated with season-nitrogen interaction is 13.729 and, therefore, highly significant. What does this mean? Consider Figure 6, which is a plot of the group means of the data.

Figure 6 Mean yields.



The horizontal axis represents the rate of nitrogen and the vertical axis measures the yield. Notice how the mean yields not only relate to the level of nitrogen and the crop season but also to the particular combination of the values of the two factors. Yields in the dry season respond better to higher nitrogen rates, while the yields in the wet season are higher if lower nitrogen rates are used. Thus, the yield for each level of nitrogen depends on the time of the year. It may therefore be misleading to recommend a level of nitrogen without mentioning in which crop season it should be applied. Following regression techniques, explained in the next section, we obtain the following two regression equations:

$$\textit{Dry season}: \quad \hat{Y}_D = 3.9507 + 0.0497N - 0.0002\,N^2$$

$$\textit{Wet season}: \quad \hat{Y}_W = 4.5240 + 0.0466N - 0.0004\,N^2$$

With these estimated nitrogen response functions, the optimum nitrogen rates in both seasons can be computed as:

$$\frac{d\hat{Y}}{dN} = b + 2cN = 0$$

Thus, the estimated optimum nitrogen rates that maximize yield are:

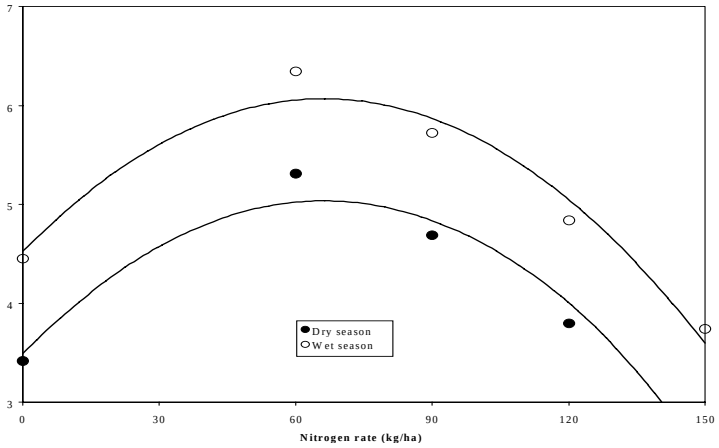
$$\textit{Dry season}: \quad N = \frac{-0.0497}{(2)\cdot(-0.0002)} = 124.25\ kg\ /\ ha$$

$$\textit{Wet season}: \quad N = \frac{-0.0466}{(2)\cdot(-0.0004)} = 58.25\ kg\ /\ ha$$

Consequently, the results seem to indicate that we need different nitrogen recommendations for dry and wet seasons.

If there were no interaction between season and nitrogen a plot similar to the one shown in Figure 7 might have resulted. In this figure, the difference between yields in the two seasons is the same for the 5 levels of nitrogen.

Figure 7 Mean yields with no interaction.



In this case, the effects of the two factors can be tested individually. The *F* value associated with seasons provides a test of the hypothesis that the crop season does not affect the yield. Similarly, the F value associated with nitrogen will test the hypothesis that the level of

nitrogen has no significant effect on the yield. If, however, interaction is present it is no use to test the hypothesis individually since the two factors jointly affect the yield.

Analysis of variance techniques can be used with any number of grouping variables. In this example, we considered only two factors, season and nitrogen. The number of factors can, however, easily be extended and include for example, different rice varieties, levels of technology, climatic conditions, and weed control methods. Also, in our example each class had the same number of observations, which greatly simplified the analysis and its interpretation. This, however, does not have to be so. When unequal sample sizes occur there is a problem with dividing the total sum of squares into nice components that sum to the total. Various techniques are available for calculating sums of squares if this is the case (Kleinbaum and Kupper 1978).

Regression analysis

Regression analysis is a powerful tool for analyzing the relationships among variables. It is one of the most widely used statistical techniques in agricultural research. Regression analysis describes the effect of one or more variables, referred to as the *independent variable*, on one single variable, referred to as the *dependent variable*, by expressing the latter as a function of the former. Regression analysis can be classified according to the number of variables involved and the functional relationship between the variables. If only two variables are involved, one dependent and one independent, the procedure is called *simple*. Otherwise, it is called *multiple*. If the functional relationship between the variables is linear the procedure is termed *linear*. If not, it is called *non-linear*. To reveal the underlying functional relationship between two variables, it is helpful to make a scatter diagram. In a scatter diagram the data is are plotted in a two-dimensional space, where the axes represent the variables. If the two variables are strongly related, then the data points form a systematic shape (e.g., a straight line or a clear curve). If the variables are not related, then the points form an irregular "cloud." Figure 8 illustrates some commonly encountered patterns.

Figure 8 Scatter diagrams.

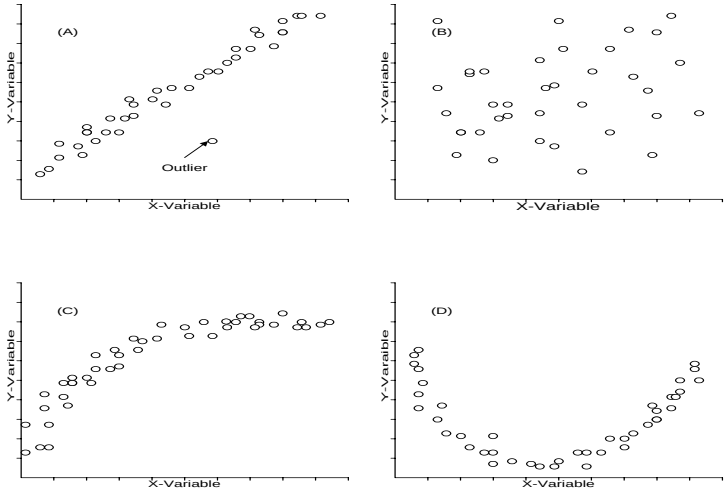


Figure 8(a) shows a linear relationship, (b) no relationship, and (c) and (d) show non-linear relationships. Scatter diagrams give an indication of what type of mathematical function best describes the relationship between the variables. Furthermore, the plot might indicate points that are suspiciously different from the others. In Figure 8(a) there appears to be an *outlier*. Dealing with outliers can be difficult. It is usually tempting to outright discard an outlier. However, the decision to exclude an outlier should be made with extreme caution. Apart from making a scatter diagram, it is often useful to quantify the strength of association between two variables by calculating the *correlation coefficient* or *covariance*. The correlation coefficient between two variables *Y* and *X* is defined as:

$$R = \frac{s_{xy}}{s_x \cdot s_y}$$

where, s_x and s_y are the standard deviations of the two variables and S_{xy} is the *covariance* of *X* and *Y*. The covariance is defined as:

$$s_{xy} = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\left(n-1\right)}$$

The correlation coefficient and the covariance are measures of the strength of the linear relationship between *X* and *Y*. The covariance will be positive if large (small) values of one variable are associated with large (small) values of the other variable, as in Figure 8(a). If large (small) values of one variable are associated with small (large) values of the other variable the covariance will be negative. If there is little or no linear association between the variables, as in Figure 8(b) and (d), the covariance will be close to zero. A serious deficiency in using covariance is that the numerical size of the covariance is not meaningful. It depends on the units in which the variables *X* and *Y* are measured. For example, the covariance will be much larger if we measure something in millimeters instead of, say, inches. The correlation coefficient is free of such scale effects. It can vary from –1 to +1. A correlation of 0 indicates that there is no linear association. There may, however, be a very strong nonlinear relationship as in Figure 8(d). A correlation of –1 or +1 would suggest a perfect linear association. As for the covariance, a positive correlation implies that large (small) values of *X* are associated with large (small) values of *Y*. The correlation will be negative if large (small) values of *X* are associated with small (large) values of *Y*.

Linear relationships

Once it has been determined that it is reasonable to assume linearity, we can make use of linear regression procedures. The goal of linear regression procedures is to fit a straight line through the points in the scatter diagram. Specifically, it is the computation of a line so that the squared deviations of the observed points from that line are minimized. This procedure is also referred to as linear least squares estimation. Assumptions for linear regression are that the errors (e_i), i.e. departures from regression, of the sample observations are normally distributed and independent, and that there is a constant variance for the entire regression line.

Simple linear regression

Simple linear regression involves one dependent (*Y*) and one independent variable (*X*) that are related in a linear fashion. The functional form of a linear relationship between a dependent variable *Y* and an independent variable *X* is:

$$Y = \alpha + \beta X$$

Thus, the *Y* variable is expressed in terms of a constant (α) and a slope (β) times the *X* variable. The constant is also referred to as the intercept, and the slope as the regression coefficient or B coefficient. Simple linear regression deals with the estimation and tests of significance concerning the two parameters α and β. Note that, as simple regression analysis is performed under the assumption of linearity, it does not provide a test as to whether the best functional relationship between *X* and *Y* is indeed linear. To illustrate the procedure of simple linear regression consider the data shown in Table 6 on the demand for sugar over a small range of prices.

Table 6 Price and sales of sugar.

Price ($/kg)	50	55	60	65	70	75	80
Sales (kg)	128	112	118	97	76	83	74

If we plot these data in a scatter diagram, we can see that the observed data points do not all fall on a straight line. Consequently, many lines could be fitted to these data. The problem is which line to select among the many possibilities. According to the method of least squares, the best line is the one that minimizes the sum of squared vertical distances from the observed data points to the line (e_i). This will be achieved as we take α and β to be:

$$\beta = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

$$\alpha = \overline{Y} - \beta\overline{X}$$

which yield a β of –1.871 and an α of 219.93. Substituting these estimates in the general linear regression equation gives:

$$\hat{Y} = 219.93 - 1.871X$$

where, $\hat{Y}$ is used to indicate that we are dealing with an estimated value of *Y*. With this equation we can estimate what the demand for sugar will be given its price.

Figure 9 Regression line.

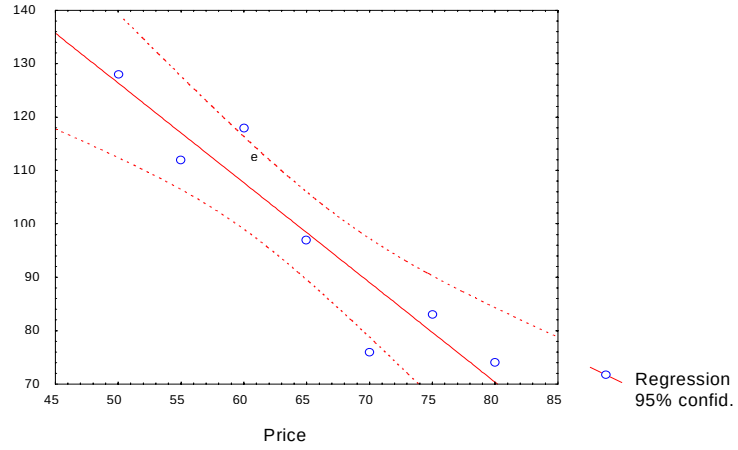


Figure 9 shows the least square or regression line superimposed on the scatter plot. The intercept and the regression coefficient estimated from a single sample are likely to differ from the population values and vary from sample to sample. A measure of variation of the regression line is the *standard error of estimate*, s_ε, which is defined as:

$$s_\varepsilon = \sqrt{\frac{\left(Y_i - \alpha - \beta X_i\right)^2}{n-2}} = 8.1223$$

A frequently tested hypothesis is that there is no linear relationship between X and Y, that is the slope of the population regression line is zero. To test this we use the t statistic:

$$t = \frac{\beta}{s_\beta}$$

where, s_β is the standard error of the slope and is computed as:

$$s_\beta = \frac{s_\varepsilon}{\sqrt{\sum X^2 - n\overline{X}^2}} = \frac{8.1223}{\sqrt{4225 - 7 \cdot 65^2}} = 0.30699$$

The computed t-value is then –6.09599 and has n-2 degrees of freedom. Comparing this value to the tabular t-value in Appendix 1 Table 1, the hypothesis is rejected and we can conclude that the independent variable is significantly related in a linear way to the dependent variable. To test the hypothesis that the intercept is zero, the statistic is:

$$t = \frac{\alpha}{s_\alpha}$$

where, s_α is the standard error of the intercept which is defined as:

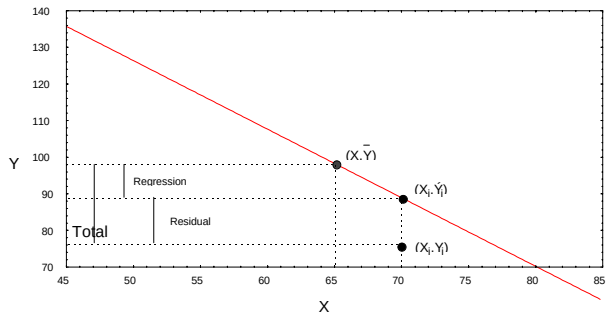
$$s_\alpha = s_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X - \overline{X}\right)^2}} = 8.1223\sqrt{\frac{1}{7} + \frac{65^2}{700}} = 20.18933$$

The t-value is therefore 10.89331, which is larger than the tabular value with 5 degrees of freedom. The intercept α is thus significantly different from zero.

A different way to see how well the regression line fits the data is to use the analysis of variance test. With this test, the total observed variability in the dependent variable is subdivided into two components, namely that attributable to regression and that which is not (the residual, i.e. unexplained). Consider Figure 10. The distance between the observed value (Y_i) and the mean value ($\overline{Y}$) can be subdivided into two parts:

$$Y_i - \overline{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \overline{Y}\right)$$

Figure 10 Components of variability.



The difference between the observed value (Y_i) and the value predicted by the regression line ($\hat{Y}_i$) is called the *residual from regression*. It will be zero if the regression line passes through the point. The second component $\left(\hat{Y}_i - \overline{Y}\right)$ is the difference between the predicted value and the mean of the *Y*'s. This difference is explained by the regression in that it represents the improvement in the estimate of the dependent variable achieved by the regression. Without regression, the mean of *Y* would have been used as an estimate. The total sum of squares is computed as:

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \overline{Y}\right)^2$$

The first term on the right hand side is the *residual sum of squares* and the second term is the *regression sum of squares*. These sums of squares are shown in the analysis of variance table under the heading sums of squares.

Source of variation	Degrees of freedom*	Sums of squares	Mean squares**	F ratio
Regression	1	2451.57	2451.57	37.16
Residual	5	329.86	65.97	
Total	6	2781.43		

* As there are 7 values of Y, the total sums of squares has 6 degrees of freedom. The regression has one df. This leaves 5 df for the residual.

** The mean squares for each entry are the sum of squares divided by its degrees of freedom.

After computing the mean squares the regression can be tested by:

$$F = \frac{MS\ regression}{MS\ residual} = 37.16$$

As this figure is much greater than the tabular *F* value (0.01 level) with 1 and 5 degrees of freedom, the regression is said to be significant at the 0.01 level. Observe that the square root of the *F* value is 6.09599, which is the value of the *t* statistic for the slope. Therefore, either *t* or *F* values can be computed to test that there is a linear relationship between *Y* and *X*.

A measure of how well a regression fits the sample data is the *coefficient of determination*. It computes the proportion of the total variation in *Y* that is associated with the regression. The coefficient of determination (R^2) is equal to the square of the correlation coefficient:

$$R^2 = \frac{SS\ regression}{SS\ total} = \frac{2451.57}{2781.43} = \left(R\right)^2 = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = \left(\frac{-218.33}{(10.80)(21.53)}\right)^2 = 0.88$$

If all observed *Y*'s fall exactly on the regression line, R^2 is equal to one. If there is no linear relationship between the dependent and independent variables, R^2 is equal to zero. An R^2 of zero does not, however, mean that there is no association between the variables. It only indicates that there is no linear relationship. In our example R^2 is 0.88, which means that 88% of the variation in *Y* is associated in a linear way with *X*. The sample R^2 tends to be an optimistic estimate of how well the regression fits the population. The regression usually does not fit the population as well as it fits the sample from which it is derived. The statistic *adjusted* R^2 attempts to correct R^2 so that it more accurately reflects the regression for the population. Adjusted R^2 is given by:

$$R_s^2 = R^2 - \frac{p(1-R^2)}{n-p-1} = 0.86$$

where, p is the number of independent variables (in the demand for sugar example we have only one, namely the price).

Confidence limits on the regression line can be calculated for estimated values of $\hat{Y}_i$. Estimates will be most precise in the middle of the data, i.e. near the mean value of X. The further away from the mean, the greater the error and the wider the interval. Furthermore, the width of the interval differs depending on the type of $\hat{Y}_i$ being estimated: an average Y, indicating the limits in which the true mean of Y for a given X will lie, or an individual Y, indicating the limits for a single value of Y. Obviously, means can be estimated with more precision than a single value of Y and thus the interval for means will be narrower. The confidence limits are:

$$\hat{Y} \pm t\sqrt{(MS\ residual)\left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right)} \quad \textit{for a mean estimate}$$

$$\hat{Y} \pm t\sqrt{(MS\ residual)\left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right)} \quad \textit{for a point estimate}$$

where, X_0 is a selected value of X, and the degrees of freedom for t are the same as the degrees of freedom for the residual sum of squares. If we use the standard error of estimate, s_ε (equal to the square root of the mean square residual), the degrees of freedom are equal to n-2. For example, in computing confidence limits on the mean of Y we pick X_0=50. For $\hat{Y}$ we find then 126.36, and the 95% confidence limits:

$$= 126.36 \pm 2.571\sqrt{(65.97)\left(\frac{1}{7} + \frac{(50-65)^2}{700}\right)}$$

$$= 112.13 \text{ to } 140.59$$

For other values of X_0 we would get:

		95% Confidence Limits	
X_0	$\hat{Y}$	Lower	Upper
42	141.33	121.53	161.12
57	113.26	103.15	123.37
70	88.93	80.11	97.75
83	64.60	48.35	80.85

In Figure 9, these points have been plotted and connected by smooth curves.

Some caution is in order when using regression analysis procedures. First, statistical significant does not mean that it is also 'practically' significant. For example, you might find a significant relationship between the income of a farmer and the length of his hair. In practice

this, of course, does not make much sense. Second, a statistical relationship does not imply a cause and effect relationship. Does the price of sugar determine its demand or does demand determine the price? Lastly, one should be careful about extrapolation beyond the sample observations because in that range the error is undefined.

Multiple linear regression

It frequently happens that a variable *Y* is related to more than one independent variable. If this relationship can be estimated, it enables us to make more accurate predictions of the dependent variable than would be possible with simple linear regression. If we assume that the independent variables affect the dependent variable in a linear fashion and are independent of one another, the method of multiple linear regression analysis can be used. Multiple linear regression is merely an extension of simple linear regression by including more than one independent variable. The general form of the multiple linear regression model is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Multiple linear regression equations are frequently determined with the aid of a computer since the computations are cumbersome. Here the calculation methods are illustrated for a case with two independent variables, using the data on grain yield (Y), plant height (X_1), and tiller number (X_2), based on Gomez and Gomez (1984).

Variety Number	Grain yield (kg/ha) Y	Plant height (cm) X_1	Tiller (no./hill) X_2
1	5,755	110.5	14.5
2	5,939	105.4	16.0
3	6,010	118.1	14.6
4	6,545	104.5	18.2
5	6,730	93.6	15.4
6	6,750	84.1	17.6
7	6,899	77.8	17.9
8	7,862	75.6	19.4
9	7,188	89.5	18.3
10	6,730	94.7	16.8
11	7,348	74.0	19.1
12	6,230	108.3	15.1
13	6,145	103.7	16.4
14	6,331	98.2	15.6
15	7,026	85.0	19.1
Sum	99,488	1,423	254
Mean	6,632.53	94.87	16.93

With k=2, the multiple linear regression equation is expressed as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

According to the method of least squares, the best estimates of β_1 and β_2 are obtained by simultaneously solving the set of *least square normal equations*:

$$b_1 \sum x_1^2 + b_2 \sum x_1 x_2 = \sum x_1 y$$

$$b_1 \sum x_1 x_2 + b_2 \sum x_2^2 = \sum x_2 y$$

where, x_i and y are the corrected values of X and Y, that is, $x_i = X - \overline{X}_i$, and $y = Y - \overline{Y}$. After solving the normal equations for the for the X coefficients thee constant term α is obtained by solving:

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

In case of k independent variables the X coefficients are found by solving k simultaneous normal equations:

$$\begin{aligned} b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \ldots + b_k \sum x_1 x_k &= \sum x_1 y \\ b_1 \sum x_1 x_2 + b_2 \sum x_2^2 + \ldots + b_k \sum x_2 x_k &= \sum x_2 y \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \\ b_1 \sum x_1 x_k + b_2 \sum x_2 x_k + \ldots + b_k \sum x_k^2 &= \sum x_k y \end{aligned}$$

Given the X coefficients, the constant term is computed as:

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 - \ldots b_k \overline{X}_k$$

Note that the normal equations for the multiple linear model include the solution for the simple linear regression. In case of one independent variable we would have one normal equation:

$$b_1 \sum x_1^2 = \sum x_1 y$$

Hence,

$$b_1 = \frac{\sum x_1 y}{\sum x_1^2} = \frac{\sum \left(X_1 - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sum \left(X_1 - \overline{X}\right)^2}$$

In fact, simple linear regression can be seen as a special case of multiple linear regression with only one independent variable.

Computing the corrected sums of squares and their products for each of the (k+1) variables, Y, X_1 and X_2 in our example yields the following set of normal equations:

$$\begin{aligned} 2{,}599.13\, b_1 - 266.82\, b_2 &= -98{,}482.23 \\ -266.82\, b_1 + 40.35\, b_2 &= 11{,}892.33 \end{aligned}$$

Solving these equations gives the solutions for b_1 and b_2 as:

$$b_1 = -23.77 \text{ and } b_2 = 137.51$$

Given the values of b_1 and b_2 we can now compute the estimate of the constant term as:

$$\begin{aligned} a &= \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 \\ &= 6{,}632.53 - (-23.77)(94.87) - (137.51)(16.93) = 6{,}559.51 \end{aligned}$$

Thus, the estimated multiple linear regression equation relating plant height (X_1) and tiller number (X_2) to yield (Y) is:

$$\hat{Y} = 6{,}559.51 - 23.77\, X_1 + 137.51\, X_2$$

The analysis of variance procedure can be used to test the overall model, that is, to test the hypothesis that $\beta_1 = \beta_2 = 0$. The analysis of variance table for this test is:

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F ratio
Regression	2	3,976,237	1,988,119	31.75
Residual	12	751,312	62,609	
Total	14	4,727,549		

The degrees of freedom for the total are equal to the number of observations minus one (15-1=14). The number of degrees of freedom for the regression is equal to the number of independent variables, in this case 2. The remaining degrees of freedom are associated with the residual. The sum of squares for the total is computed as:

$$SST = \sum y^2 = \sum \left(Y - \overline{Y}\right)^2 = 4{,}727{,}549$$

The regression sum of squares for any least square regression is equal to the sum of the estimated X coefficients (b_i) times the right hand side of their normal equations. Thus,

$$\begin{aligned} SSR &= b_1\left(\sum x_1 y\right) + b_2\left(\sum x_2 y\right) \\ &= \left(-23.77\right)\left(-98{,}482.23\right) + \left(137.51\right)\left(11{,}892.33\right) = 3{,}976{,}237 \end{aligned}$$

The residual sum of squares is obtained by subtracting:

$$SSE = SST - SSE = 751{,}312$$

As usual, the mean squares for regression and residual are computed by dividing the respective sums of squares by their degrees of freedom. The hypothesis is tested by computing the F ratio as:

$$F = \frac{MSR}{MSE} = \frac{1{,}988{,}119}{62{,}609} = 31.75$$

and comparing this value to the tabular F value from Appendix 1, Table 6. The tabular F values with 2 and 12 degrees of freedom are 3.88 at the 5% level and 6.93 at the 1% level of significance. Because the computed F value exceeds the tabular F value at the 1% level, the hypothesis that all X coefficients are zero can be rejected, and the regression is said to be significant at the 1% level.

If the hypothesis is rejected it might be of interest to test the individual terms of the regression. For example, we might want to test the hypothesis that tiller number does not contribute to the prediction of the yield, that is, b_2 is zero. If this proves to be so, we can rewrite the equation in terms of only X_1, using simple linear regression technique. The method used to test the contribution of a particular variable (or set of variables) is the *marginal F test*. The marginal F test considers two models: (1) a model containing all the independent variables and (2) a model containing all the variables except the one (or ones) we want to test for significance. For both models, the regression sums of squares are calculated and the difference between the two is the gain due to the variable(s) being tested. The mean square for the gain is then tested against the mean square error of the full model. In testing the significance of X_2, we have already established that the regression sum of squares for the full model is 3,976,237 with 2 degrees of freedom and the residual is 751,312 with 12 degrees of freedom. For the regression of Y on X_1 alone, the normal equation is:

$$\begin{aligned} b_1 \sum x_1^2 &= \sum x_1 y \quad \text{or,} \\ 2{,}599.13\, b_1 &= -98{,}482.23 \\ b_1 &= -37.89 \end{aligned}$$

The regression sum of squares associated with X_1 alone is computed as:

$$SS = b_1\left(\sum x_1 y\right) = \left(-37.89\right)\left(-98{,}482.23\right) = 3{,}731{,}537$$

with 1 degree of freedom. The gain due to X_2 after X_1 is the difference between the regression sums of squares of the full and reduced model:

$$SS\ gain = 3{,}976{,}237 - 3{,}731{,}537 = 244{,}700$$

with (2-1)=1 degree of freedom. The F ratio is the computed as:

$$F = \frac{MS\ gain}{MSE} = \frac{244{,}700}{62{,}609} = 3.91$$

The test is usually presented in an analysis of variance table:

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F ratio
Regression (X_1,X_2)	2	3,976,237		
Regression (X_1)	1	3,731,537		
Gain (X_2)	1	244,700	244,700	3.91
Residuals	12	751,312	62,609	
Total	14	4,727,549		

The computed F value is not significant at the 1% level. Consequently, the hypothesis that b_2=0 can not be rejected and we could rewrite the regression into a simple linear regression in which yield is predicted using only one independent variable, namely, height.

As a measure of how well the multiple regression fits the data, the *coefficient of multiple determination* is computed:

$$R^2 = \frac{SSR}{SST} = \frac{3{,}976{,}237}{4{,}727{,}549} = 0.84$$

This value measures the proportion of change in Y explained by the independent variables (84% in this case). It can be adjusted to reflect the fact that a degree of freedom is lost for every additional independent variable brought into the model. The adjusted R^2 is computed as:

$$R_s^2 = R^2 - \frac{p(1-R^2)}{n-p-1} = 0.84 - \frac{2(1-0.84)}{15-2-1} = 0.81$$

where, p is the number of independent variables in the equation.

Just as with simple linear regression we can put confidence limits on a multiple regression. This, however, requires the computation of *c*-multipliers. The *c*-multipliers are the elements of the inverse of the matrix of corrected sums of squares and their products as they appear in the normal equations. It is therefore the inverse of the following matrix:

$$\begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^2 \end{bmatrix} = \begin{bmatrix} 2{,}599.13 & -266.82 \\ -266.82 & 40.35 \end{bmatrix}$$

The inverse of this matrix is:

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} 0.0012 & 0.0079 \\ 0.0079 & 0.0772 \end{bmatrix}$$

The *c*-multipliers are used in the calculation of confidence limits either on a mean value of Y or on a single predicted value of Y. The general equations for calculating confidence limits in case of k independent variables are:

$$\hat{Y} \pm t\sqrt{MS\ residual\left(\frac{1}{n} + \sum_i \sum_j c_{ij}\left(X_i - \overline{X}_i\right)\left(X_j - \overline{X}_j\right)\right)} \quad \textit{for a mean estimate}$$

$$\hat{Y} \pm t\sqrt{MS\ residual\left(1 + \frac{1}{n} + \sum_i \sum_j c_{ij}\left(X_i - \overline{X}_i\right)\left(X_j - \overline{X}_j\right)\right)} \quad \textit{for a point estimate}$$

where, i and j are 1,2,…k and the degrees of freedom for t are the same as the degrees of freedom associated with the residual sum of squares. In the example, if we specify X_1=98.7 and

X_2=14.9, then the predicted yield is $\hat{Y}$ =6,262, and the 95% confidence limits for the mean predicted value are:

$$6{,}262 \pm 2.179\sqrt{62{,}609\left(\begin{array}{l}\frac{1}{15}+c_{11}\left(98.7-\overline{X}_1\right)^2+c_{12}\left(98.7-\overline{X}_1\right)\left(14.9-\overline{X}_2\right)\\ +c_{21}\left(14.9-\overline{X}_2\right)\left(98.7-\overline{X}_1\right)+c_{22}\left(14.9-\overline{X}_2\right)^2\end{array}\right)} = 5{,}973 \text{ to } 6{,}550$$

The *c*-multipliers may also be used to compute the estimated regression coefficients. The general equation is:

$$b_j = \sum_i c_{ji}\left(\sum x_i\, y\right)$$

where, $\sum x_i\, y$ is the right hand side of the i^{th} normal equation. To illustrate, b_2 in the example would be calculated as:

$$b_2 = c_{21}\left(-98{,}482.23\right)+c_{22}\left(1{,}892.33\right)=140.0$$

which is the same as before except for rounding off problems.

It is now also possible to compute *t* values for each estimated regression coefficient b_i as:

$$t_i = \frac{b_i}{s\left(b_i\right)}$$

where, $s(b_i)$ is the standard error of each estimated regression coefficient and is computed as:

$$s\left(b_i\right)=\sqrt{c_{ii}\left(MSE\right)}$$

The standard errors of b_1 and b_2 in the example are:

$$s\left(b_1\right)=\sqrt{0.0012\left(62{,}609\right)}=8.668$$

$$s\left(b_2\right)=\sqrt{0.0772\left(62{,}609\right)}=69.523$$

and the *t* values are:

$$t_1=\frac{-23.77}{8.668}=-2.742$$

$$t_2=\frac{137.51}{69.523}=1.978$$

both with 12 degrees of freedom (=df *MSE*).

The hypothesis that β_i=0, at α level of significance is tested by comparing each computed *t* value to the tabular *t* value. The hypothesis is rejected if the absolute value of the computed *t* value is greater than the corresponding tabular *t* value. For this example, the tabular *t* value with 12 degrees of freedom at the 5% level of significance is 2.179. The results show that only the absolute value of the computed t_1 value is greater than the tabular *t* value at the 5% level. This indicates that only plant height influences the yield in a significant way and the variable tiller number can be dropped from the regression equation. Note that the square of t_2=3.91 is equal to the *F* value obtained in the marginal *F* test.

Nonlinear relationships

When relationships among variables under consideration are not linear, such as in Figure 8(c) and (d), linear regression procedures are inadequate and we have to turn to nonlinear regression analysis. There are numerous functional forms that can describe a nonlinear relationship between a dependent variable and the independent variables. The chief difficulty in nonlinear regression is the selection of a suitable equation. If the scatter diagram resembles a

mathematical function that is clearly recognizable, it can be fitted to the data. Figure 8(c), for example, is a logarithmic function of the form $Y = a + b\ln(X)$ and Figure 8(d) is a quadratic function of the form $Y = a + b_1 X + b_2 X^2$. If it is not obvious which function is to be fitted to the data, it is best to try several potential models and select the best alternative among them. Here we will focus on one nonlinear regression technique, involving the linearization of the nonlinear form.

Simple nonlinear regression

A nonlinear relationship between two variables can be linearized by creating a new variable or by transforming one or both variables. Creating a new variable is most commonly applied to the k^{th} degree polynomial:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$$

Such an equation can be linearized by creating k new variables: Z_1, Z_2,…,Z_k, to form a multiple linear equation of the form:

$$Y = \alpha + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_k Z_k$$

where, $Z_1=X$, $Z_2=X^2$,…, and $Z_k=X^k$.
The nonlinear functions:

$$\frac{1}{Y} = \alpha + \beta X \text{ and } Y = \alpha + \frac{\beta}{X}$$

have the linearized forms:

$$Y' = \alpha + \beta X \text{ and } Y = \alpha + \beta X'$$

where, $Y' = \frac{1}{Y}$ and $X' = \frac{1}{X}$.

Functions that are nonlinear in their coefficients can sometimes be linearized by logarithmic transformation. Examples of some commonly encountered nonlinear forms and their linear transformation are presented in Table 7.

Table 7 Linear transformations of nonlinear models.

Model	Equation	Linear transformation
Exponential	$Y = \alpha e^{\beta X}$	$\ln(Y) = \ln(\alpha) + \beta X$
Growth	$Y = e^{(\alpha + \beta X)}$	$\ln(Y) = \alpha + \beta X$
S	$Y = e^{\left(\alpha + \frac{\beta}{X}\right)}$	$\ln(Y) = \alpha + \frac{\beta}{X}$
Compound	$Y = \alpha \beta^X$	$\log(Y) = \log(\alpha) + \log(\beta) X$
Power	$Y = \alpha X^\beta$	$\log(Y) = \log(\alpha) + \beta \log(X)$

After linearization the regression procedures for linear regression as described in the previous section can be directly applied. To illustrate, consider the following data on the yield of a crop (Y) and the growing season rainfall (X):

Observation	Yield (t/ha)	Rainfall (mm)	$Z = X^2$
1	10	98	9,604
2	17	125	15,625
3	30	145	21,025
4	40	183	33,489
5	40	201	40,401
6	47	223	49,729
7	43	258	66,564
8	52	269	72,361
9	48	298	88,804
10	53	323	104,329
11	55	354	125,316
12	50	367	134,689
Mean	40.42	237	63,494.67
$\sum x^2 = 87{,}908$		$\sum z^2 = 19{,}909{,}080{,}739$	
$\sum xy = 12{,}676$		$\sum zy = 5{,}534{,}248$	
$\sum xz = 41{,}295{,}972$			

It is assumed that the yield is a quadratic function of the growing season rainfall; that is,

$$Y = \alpha + \beta_1 X + \beta_2 X^2$$

The function is linearized by introducing a new variable Z which is defined as $Z=X^2$. The linearized form is:

$$Y = \alpha + \beta_1 X + \beta_2 Z$$

To estimate the regression coefficients of this function we can simply apply the multiple linear regression procedures of the previous section. The set of normal equations is:

$$b_1 \sum x^2 + b_2 \sum xz = \sum xy$$

$$b_1 \sum xz + b_2 \sum z^2 = \sum zy$$

The sum of squares and their products for each of the variables Y, X, and Z are computed and given in the table. Filling in the equations and solving for b_1 and b_2 gives:

$$b_1(87{,}908) + b_2(41{,}295{,}972) = 12{,}676$$

$$b_1(41{,}295{,}972) + b_2(19{,}909{,}080{,}739) = 5{,}534{,}248$$

$$b_1 = 0.5317 \text{ and } b_2 = -0.0008$$

And the estimate for α is then computed as:

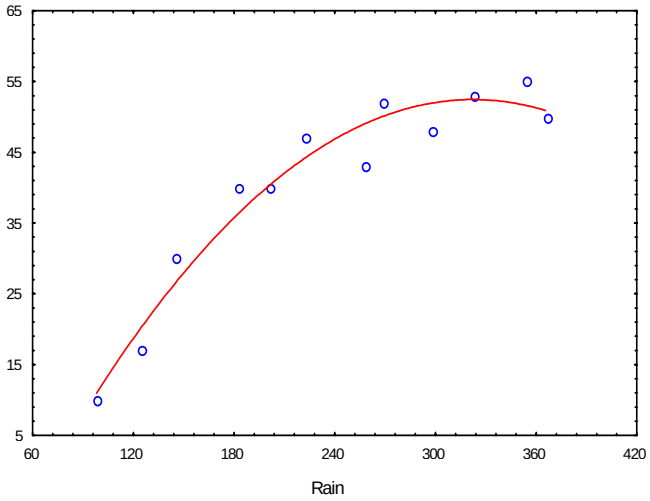
$$a = \overline{Y} - b_1 \overline{X} - b_2 \overline{Z} = 40.42 - 0.5317(237) - (-0.0008)(63494.67) = -33.22$$

Replacing the variable Z for the nonlinear component yields the following nonlinear regression equation:

$$\hat{Y} = -33.22 + 0.5317X - 0.0008X^2$$

The graphical representation of the example is shown in Figure 11.

Figure 11 Scatter diagram with nonlinear regression curve.



Multiple nonlinear regression

Multiple nonlinear regression procedures are used when a relationship between a dependent variable Y and k independent variables X_1, X_2,…,X_k does not follow a linear relationship. For example, with two independent variables, X_1 and X_2, a multiple nonlinear relationship exists if either one or both of the two variables exhibits a nonlinear relationship with the dependent variable. For instance, if both independent variables are related to Y in a quadratic manner, the corresponding multiple nonlinear regression equation would have the following general form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

A multiple nonlinear relationship may also occur as a result of interaction between two independent variables. For example, with two independent variables X_1 and X_2, each of which separately affects Y in a linear fashion, the multiple regression equation may be nonlinear if the effect of variable X_1 on Y varies with the level of variable X_2, and vice versa. In such a case the multiple regression equation may be represented by:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

If both of the foregoing cases occur simultaneously, that is, at least one of the k independent variables has a nonlinear relationship with the dependent variable and at least two independent variables interact with each other, the multiple nonlinear equation in case of two independent variables may be:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \beta_6 X_1^2 X_2 + \beta_7 X_1 X_2^2 + \beta_8 X_1^2 X_2^2$$

The analytical techniques for multiple nonlinear regression are the same as for simple nonlinear regression. The multiple nonlinear form is first linearized. Then the multiple linear regression procedures of the previous section can be directly applied. For example, the nonlinear equation:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

is linearized as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z$$

where, $Z = X_1 X_2$, the interaction variable. And the Cobb-Douglas function:

$$Y = \alpha X_1^{\beta_1} X_2^{\beta_2} \ldots X_k^{\beta_k}$$

has the linearized form:

$$Y' = \alpha' + \beta_1 X_1' + \beta_2 X_2' + \ldots \beta_k X_k'$$

where, $Y' = \log Y$, $\alpha' = \log \alpha$, and $X_i' = \log X_i$.

Statistical computer software

Many of the statistical procedures described in the previous sections can easily be accomplished using a statistical software package. However, there are many statistical packages available which makes it difficult to choose a package most suitable for one's requirements. Here an overview will be given of some of the most popular packages. The overviews are based on Prof. James Curral's reviews which can be found on CTI's website, http://www.stats.gla.ac.uk/cti/.

SPSS

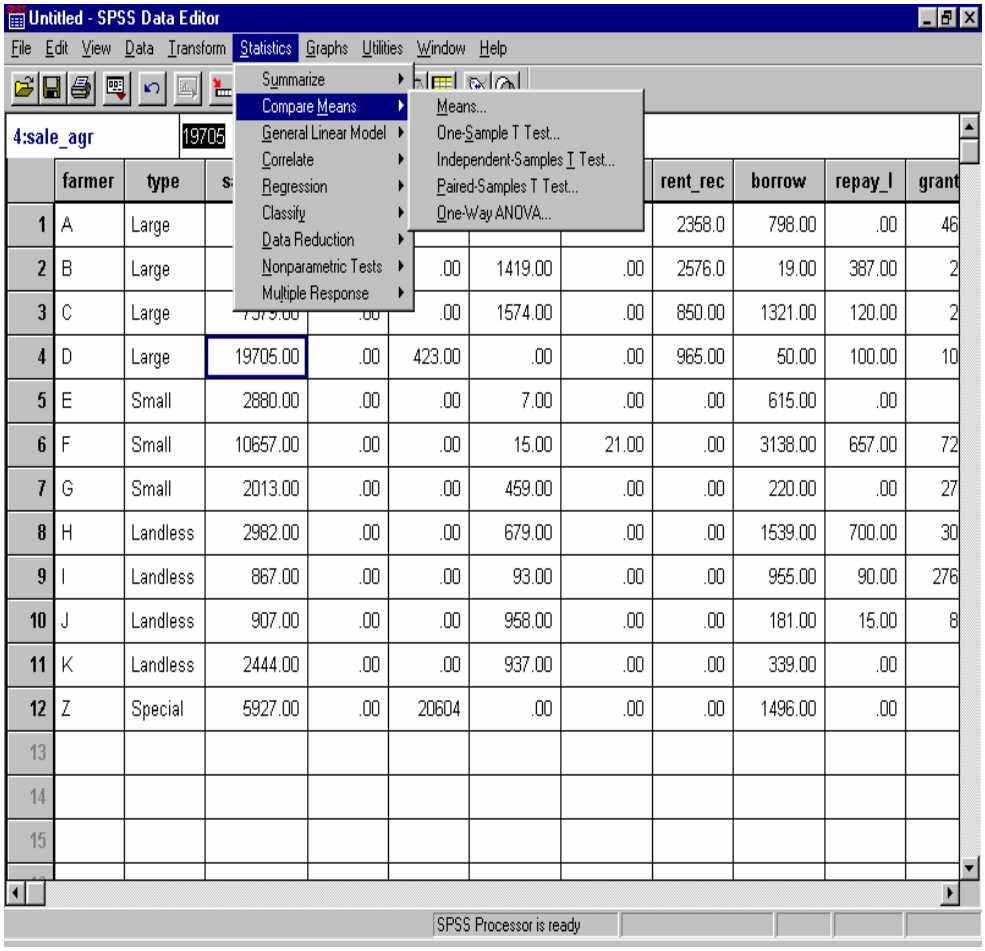
SPSS is a very popular statistical package mainly due to its ease of use. It has its roots in the social sciences (Statistical Package for Social Sciences) and, therefore, its major strength lies in the analysis of questionnaires and surveys. SPSS has been around for many years; it started off on mainframes, made it to DOS, OS/2 and finally to Windows. This review will focus on the standard version of SPSS for Windows, release 7.5. Analyzing data in SPSS is straightforward. Firstly, data are brought into SPSS by opening a previously saved SPSS data file, entering data directly in the data editor, or importing the data from a spreadsheet, database or text data file. Secondly, a procedure is selected from the menu bars to perform statistical analysis or to create a chart. Thirdly, the variables used in the analysis are specified in a procedure specific dialog box, and finally the results can be examined and the appearance altered.

Data

SPSS opens with the *Data Editor* window, which displays the contents of the active working file in a spreadsheet format (Figure 12). Each row is a single case (for example, each farmer in a survey is a case), and each column is a single variable (for example, farmer's income, farm size). Only one data sheet can be open at a time. Data can be entered and modified in the data editor, but to perform calculations or enter formulae the transform menu has to be used. SPSS supports many types of data formats, including numeric, date, text, currency, etc. However, to SPSS all data are real! This means that nominal and ordinal variables, as well as continuous ones, are treated has real numbers. Text (or string) variables are treated only as labels and cannot effectively be employed as categorical variables. In data sheets, variables may be displayed either in their numeric or label form. Consequently, categorical variables can look as though they are text variables even though in reality they are numeric. Variable names are restricted to 8 characters. However, each variable may also have longer pieces of text (variable labels) associated with it, which is used in output and labelling of charts. Within a variable a text label can be associated with any numeric value.

Missing data can either be represented by 'impossible' values or by leaving gaps in the data set.

Figure 12 SPSS data editor.



	farmer	type	s						rent_rec	borrow	repay_l	grant
1	A	Large							2358.0	798.00	.00	46
2	B	Large			.00	1419.00	.00		2576.0	19.00	387.00	2
3	C	Large			.00	1574.00	.00		850.00	1321.00	120.00	2
4	D	Large	19705.00	.00	423.00	.00	.00		965.00	50.00	100.00	10
5	E	Small	2880.00	.00	.00	7.00	.00		.00	615.00	.00	
6	F	Small	10657.00	.00	.00	15.00	21.00		.00	3138.00	657.00	72
7	G	Small	2013.00	.00	.00	459.00	.00		.00	220.00	.00	27
8	H	Landless	2982.00	.00	.00	679.00	.00		.00	1539.00	700.00	30
9	I	Landless	867.00	.00	.00	93.00	.00		.00	955.00	90.00	276
10	J	Landless	907.00	.00	.00	958.00	.00		.00	181.00	15.00	8
11	K	Landless	2444.00	.00	.00	937.00	.00		.00	339.00	.00	
12	Z	Special	5927.00	.00	20604	.00	.00		.00	1496.00	.00	
13												
14												
15												

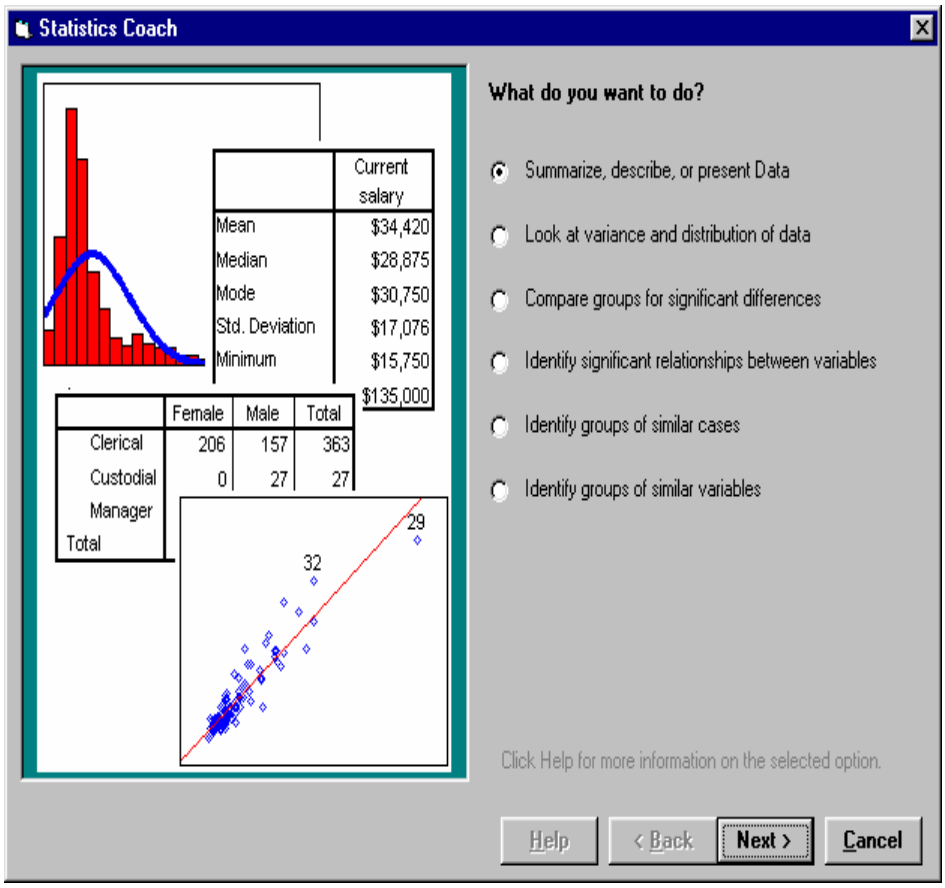
Statistics

The statistical procedures in the base module of the standard version range from standard summary statistics (frequencies, descriptive, cross tabulation), *t*-tests and one-way analysis of variance, linear regression and correlation analysis (Figure 12). Further analyses are available in modules that can be purchased separately, such as the Professional Statistics module (multidimensional scaling, clustering, etc.), the Advanced Statistics module (survival regression, MANOVA, etc.), the Tables module (complex camera-ready tables), the Trends module (time series), the Categories module (correspondence and conjoint analyses), the Chaid module (automatic interaction detection), the Lisrel, and the SPC module (quality control).

To run a statistical procedure a category is chosen from the statistical menu leading to the respective dialog boxes in which variables and options can be selected. The menus now largely replace the SPSS syntax. Syntax is the way in which commands were given to SPSS in the past. It is, however, still available for those who prefer this way of working.

A nice feature is the *Statistics Coach*. The statistical coach under the help menu helps one to choose the correct statistics procedures based on simple questions and non-technical choices (Figure 13).

Figure 23 SPSS statistics coach.

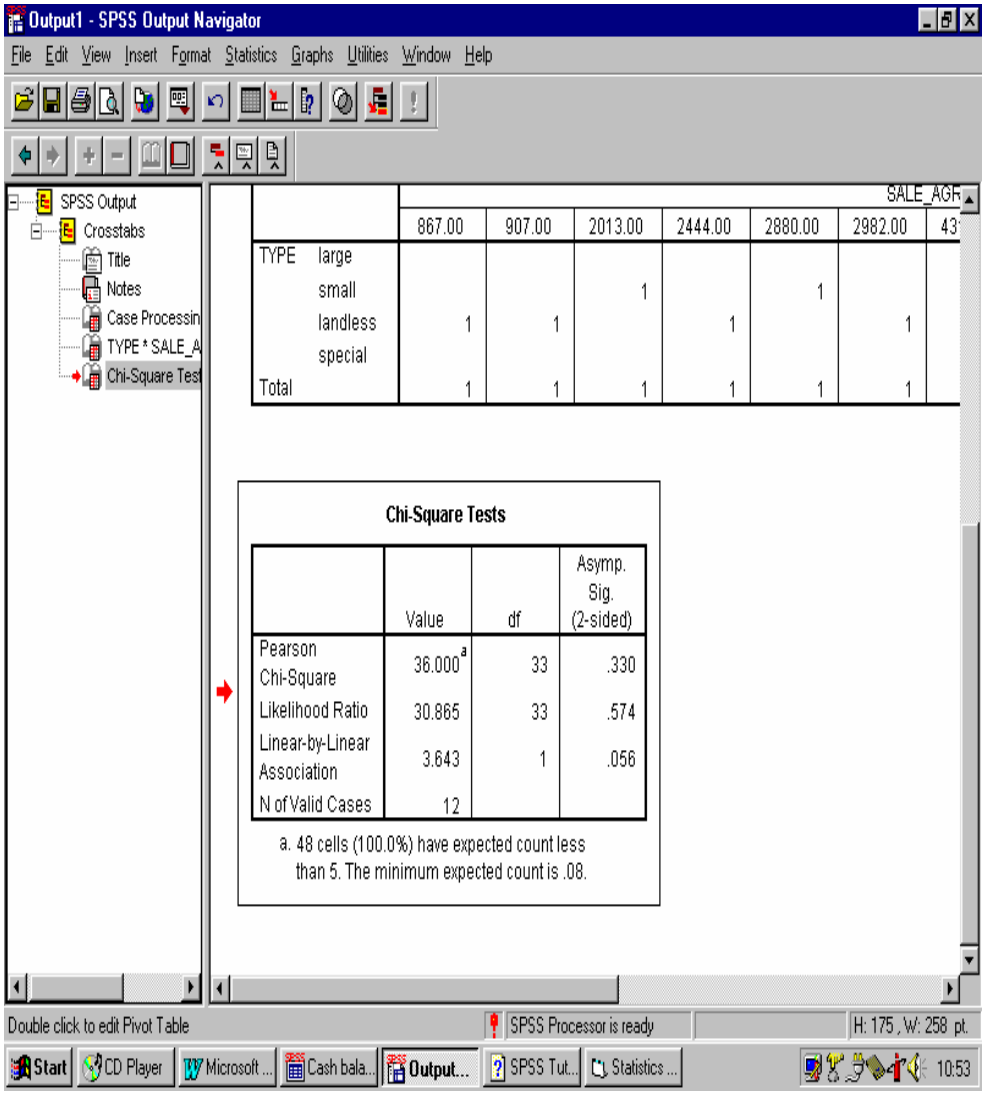


The help menu can be accessed at any time and most dialog boxes have a help button that directly displays a help topic relevant to that dialog box. For help on individual dialog box controls and output terms, click on the term you want to know about with the right mouse button and choose 'What's this?'

Output

Results are displayed in the *Output Navigator*. The output navigator window consists of two panes. The narrow pane on the left shows the structure of the output and hide, reveal or delete elements of the output. The larger pane on the right contains the actual output (Figure 14). Tables can be edited using the *Pivot Table Editor*. Double click on the table you want to pivot and select pivoting trays from the pivot menu. Now columns can be moved, rows and columns can be interchanged, and many aspects of the text appearance modified. The results can be placed in word processing programs using traditional 'cut and paste' techniques or by 'Object Linking and Embedding (OLE).

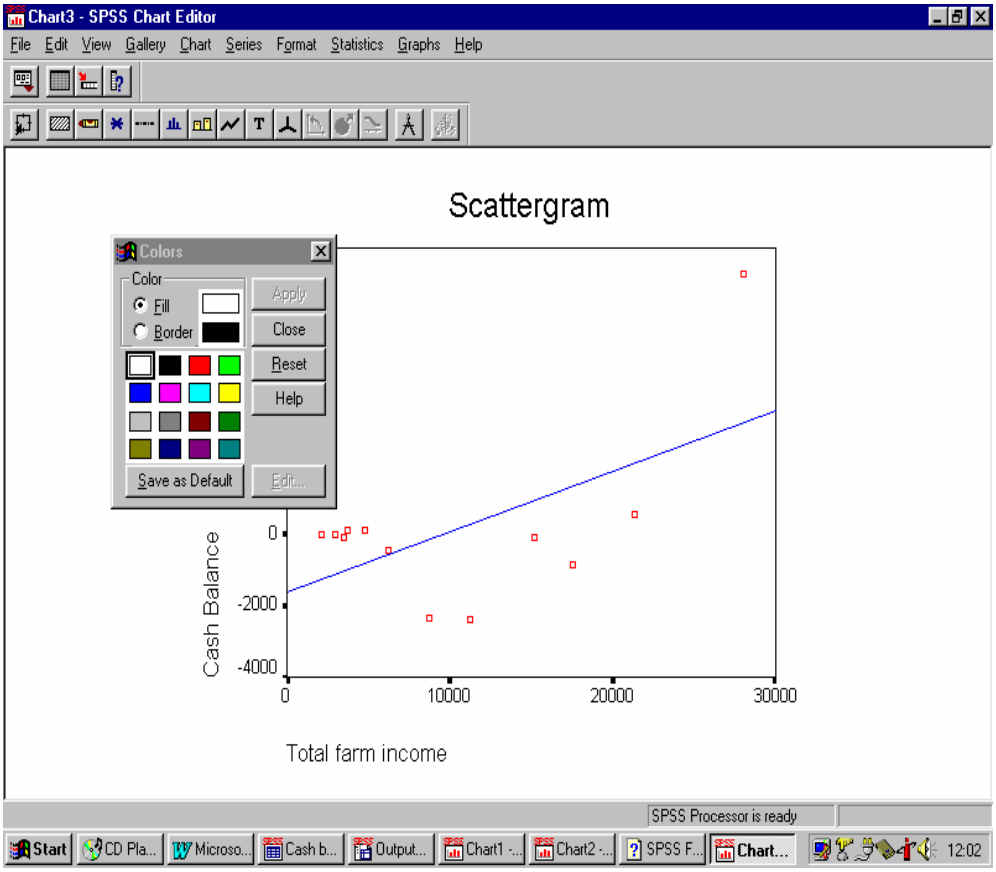
Figure 14 SPSS output navigator.



Graphics

In SPSS it is possible to produce a wide range of graph types. The graph menu shows 16 types of graphs, each with many sub-types and variants. In the respective dialog box, much of what one requires can be specified. However, the appearance of the chart can also be modified subsequently in the *Chart Editor*. The graphs created are written to the Output Navigator. A simple double click with the mouse button on the chart launches the chart editor. Figure 15 shows a scatter plot being edited.

Figure 15 SPSS chart editor.



SAS

SAS is probably one of the largest statistical computing packages available. Here it is only possible to give a very brief outline of the system. Documentation on SAS, however, is widely available and the manuals for SAS are extensive and generally excellent. SAS also provides a range of introductory manuals.

There are several different modes in which SAS may run, ranging from a batch mode to a menu system. The standard user interface makes use of windows. There are usually three sets of windows on the screen: a *program editor*, a *log window*, and an *output window*. The program editor is just a text editor in which SAS commands are typed. It is possible to add, copy, paste and delete text, search, input from and export to external files, etc. Once the program is written it is submitted to SAS. The log window shows details of how the program is running, gives warnings and error messages. All the output appears in the output window. In case of errors, the program can be recalled into the editor, corrected and subsequently submitted again.

Data sets

All SAS statistical facilities make use of data that have been put into a SAS data set. A SAS data set is a named file of data, with information conveniently stored for SAS, but not readable by any other software.

The data are stored in a rectangular way, with rows representing cases and columns representing variables. Variables are named by the user and case names can also be defined. The maximum number of variables depends on the operating system, but runs in the several thousands. The number of cases is limited only by disc storage space. The data may be character or numeric. The file may be temporary i.e. it is erased at the end of the SAS session, or permanent (available for use in later sessions).

Data sets are created either as output from statistical procedures or in a '*Data step*'. The Data step provides very flexible facilities for reading data from external files or SAS data sets, selecting or dropping observations, transforming and creating new variables and merging or sub-setting data sets. It contains a very wide range of operations and functions for manipulating and transforming numerical as well as character data. A full range of arithmetical and mathematical facilities, as well as probability functions, random number generators, financial functions and functions for manipulating special data, such as date and time, are provided.

Procedures

After a data set has been created SAS procedures are used to present and analyse the data. Each procedure has a name and is invoked by typing the procedure name and data set on which it is to operate. For example, the procedure PRINT is used for printing, and to print all the data in a data set named SURVEY would require the command:

```
PROC PRINT DATA = SURVEY;
```

There are many procedures available but those most useful in statistics can be roughly categorized as:

- Utility and data management. These perform functions such as sorting and transposing data sets, etc.
- Simple summaries and tables. These provide standard summary statistics, one or multi-way tables, etc.
- Statistical procedures. These range from basic procedures such as *t*-tests and linear regression to advanced techniques like multivariate analysis, econometrics or categorical data modelling.
- Matrix language. Performs matrix calculations.
- Graphics. Provides graphics procedures for a wide range of plots, bar charts, 2 and 3D scatter diagrams, surface and contour plots.

A recent, most welcome addition for the learner to SAS is ASSIST, a system for taking the inexperienced user through both data input steps and statistical procedures. ASSIST generates SAS codes that can be saved and modified.

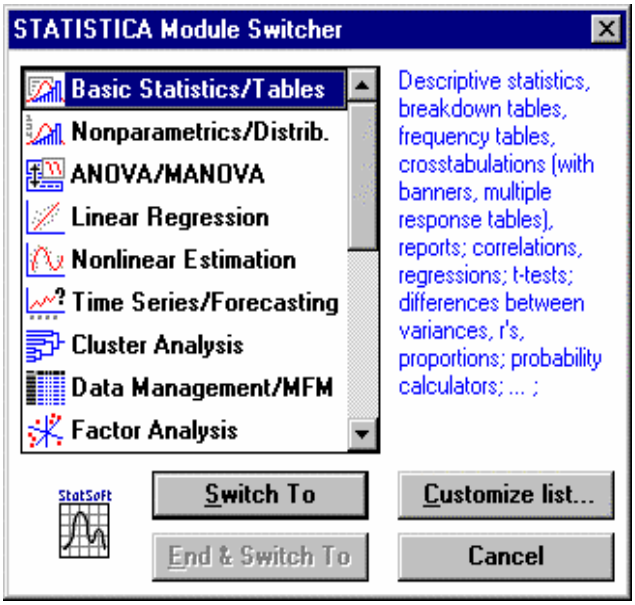
STATISTICA

Statistica is a combined statistics, graphics and database management package. The program is available in two versions: the full version (Statistica) and the basic version (Quick Statistica). The full version encompasses basic and advanced statistical modules and excellent graphics capabilities. Quick Statistica consists of four basic statistical modules and includes the full graphics capabilities of Statistica. This review will focus upon the basic version. Statistica is an extremely versatile and powerful program. It comes with three manuals: Volume I covers

general conventions and statistics; Volume II covers the graphics capabilities and statistical command language; and there is a Quick Reference Manual. To appreciate the many features of Statistica it is worth delving into these manuals as these features are not always directly evident from the user interface. In Statistica there are several ways of carrying out the same operation and there are many options available for most of the statistical procedures. The volume of options that appears on the screen for every statistical operation or feature of a graphical operation might seem a little overwhelmingly at first but let this not put you off as everyone familiar with windows can easily generate basic statistics and create simple graphs in a 'point and click' fashion. The default analysis can easily be produced by clicking OK several times after having selected the variables. Helpful are the brief descriptions which appear at the bottom of the screen if you click once and hold on the icons or buttons. Help can be accessed at any point in the program offering a comprehensive account of statistical procedures. An excellent feature in the help menu is the *Statistical Advisor*. Based on your answers to successive questions about the nature of your research, the statistical advisor suggests which statistical methods should be used and where to find them in Statistica. The program can be controlled in four ways, the most important being the interactive or batch mode. Running the program in the batch mode entails learning the Statistica Command Language, that is you develop programs that will instruct Statistica what to do. This review is based on the interactive mode since it is preferred by many.

The program starts with a dialog box which presents a menu of the available modules (Figure 16).

Figure 16 Statistica's statistical modules.



Quick Statistica encompasses only five basic modules, one dealing with data management and four dealing with statistics namely, basic statistics and tables, nonparametrics and distributions, ANOVA/MANOVA, and multiple regression. Each module contains a group of related procedures, which appear listed within a start-up panel when a module is opened.

Data management

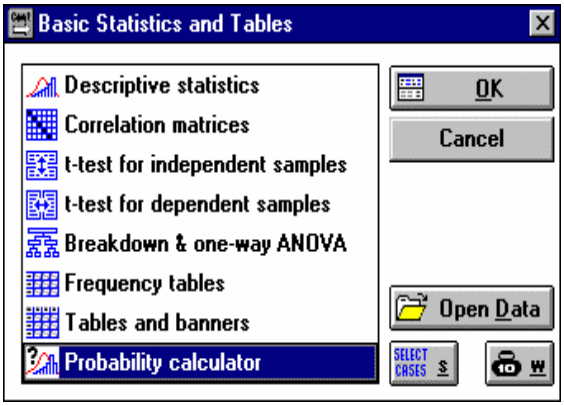
In Statistica the general structure of data files is similar to that of database management programs. Each file can be thought of as a table, where rows represent records, and each record has the same number of fields (columns). Data are organized in *cases* (the records of a database management program or the rows of a spreadsheet) and *variables* (equivalent to fields or columns of a spreadsheet). The Statistica spreadsheet window offers a wide selection of data editing and restructuring options. Statistica supports double notation of variables, where each value can simultaneously have a numeric and text identity. There can be a practically unlimited number of cases and up to 300 variables in one data file. Files up to 32,000 variables can be processed in the Megafile Manager (MFM). The *Data Management/MFM module* allows you to directly input data into its own spreadsheet or import/export a wide variety of data file formats. It includes a nice suite of data editing, restructuring, transforming, splitting, sorting, and merging facilities. Variables can be shifted and standardized, and missing values can be replaced by means. In this module the megafile manager can be accessed for data which need to be transformed, aggregated, extracted, or cleaned before they can be directly accessed by any statistical or graphics procedures of Statistica (e.g., data embedded inside very long text values or data organized into very long records).

You can also establish a DDE (dynamic data exchange) link between a source file (e.g., an Excel spreadsheet) and a Statistica data file, so that when changes are made to the data in the source file, the data will be automatically updated in the Statistica spreadsheet.

Statistics

The four basic statistical modules contain a fairly extensive set of tests and statistical procedures that will cover most requirements. Each module can be used simultaneously within a separate window and can operate on separate sets of data.The *Basic Statistics and Tables module* contains the options shown in Figure 17.

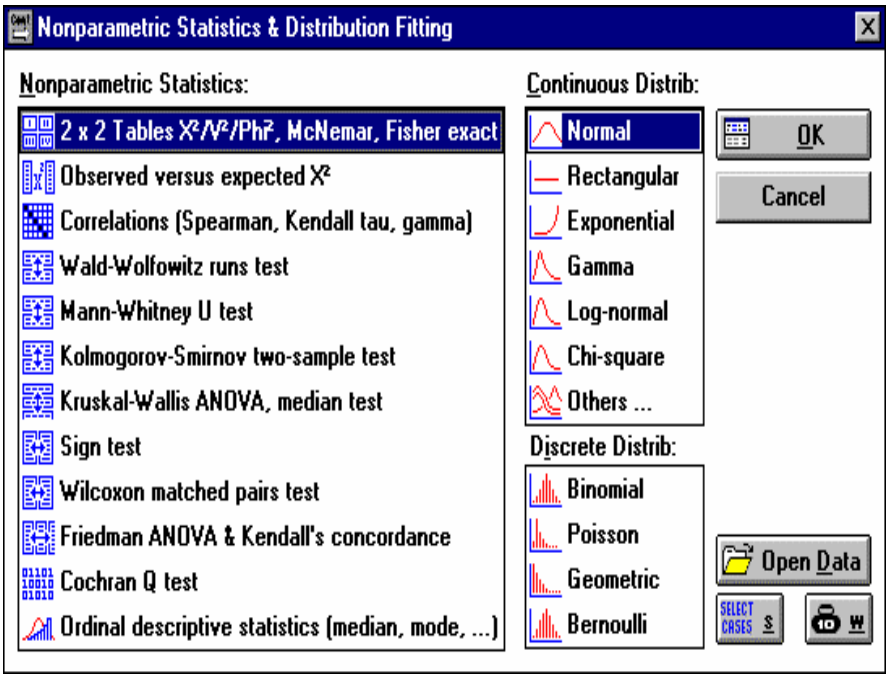
Figure 17 Basic statistics and tables module.



The statistics included in this module are usually used in the initial, exploratory phase of data analysis. It contains a whole set of summary statistics (mean, median, standard deviation, etc.), correlation matrices, *t*-tests for dependent and independent samples, one-way analysis of variance, frequency tables and histograms, cross tabulations, stub-and-banner tables and a probability calculator to compute significance levels and critical values.

The *Nonparametric and Distributions module* contains a wide range of non-parametric tests shown in the first column of the start-up panel (Figure 18).

Figure 18 Non-parametric tests.

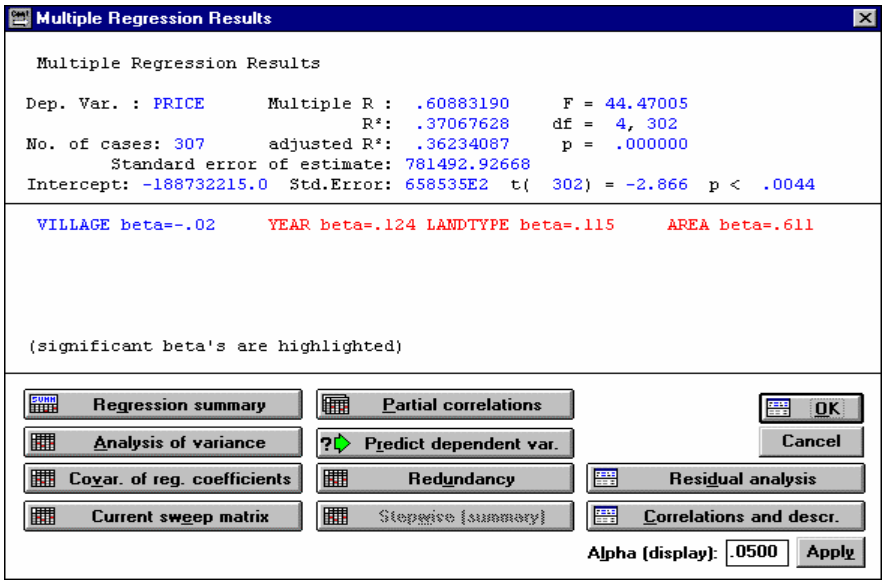


The Nonparametric and Distribution module also allows one to test the goodness-of-fit of data to the respective hypothesized distribution selected in the second column of the start-up panel. It is also possible to enter any set of arbitrary expected values (i.e. 'make up' a hypothesized distribution) and test the fit of an observed set of values (observed vs. expected frequencies).

The *ANOVA/MANOVA module* performs general univariate and multivariate analysis of variance, covariance for independent groups and repeated measures designs with fixed and changing covariates. A range of a priori and post-hoc contrast analysis tests is also available, and there are specialized options for customizing error terms for analysis of variance and carrying out canonical and discriminant function analyses.

For multiple regression techniques the *Linear Regression module* is used. This module performs stepwise forward, backward, hierarchical and ridge regression analyses. Procedures for fitting fixed non-linear regression (e.g. polynomial) equations are also available. Various procedures for residual analysis are readily available. Figure 19 shows the multiple regression results.

Figure 19 Regression results.



The full version of Statistica offers many more modules for more advanced statistical procedures, such as factor analysis, cluster analysis, reliabilities, time series, log-linear analysis and multidimensional scalings. However, the range of statistical procedures within the basic version of Statistica can satisfy most research requirements.

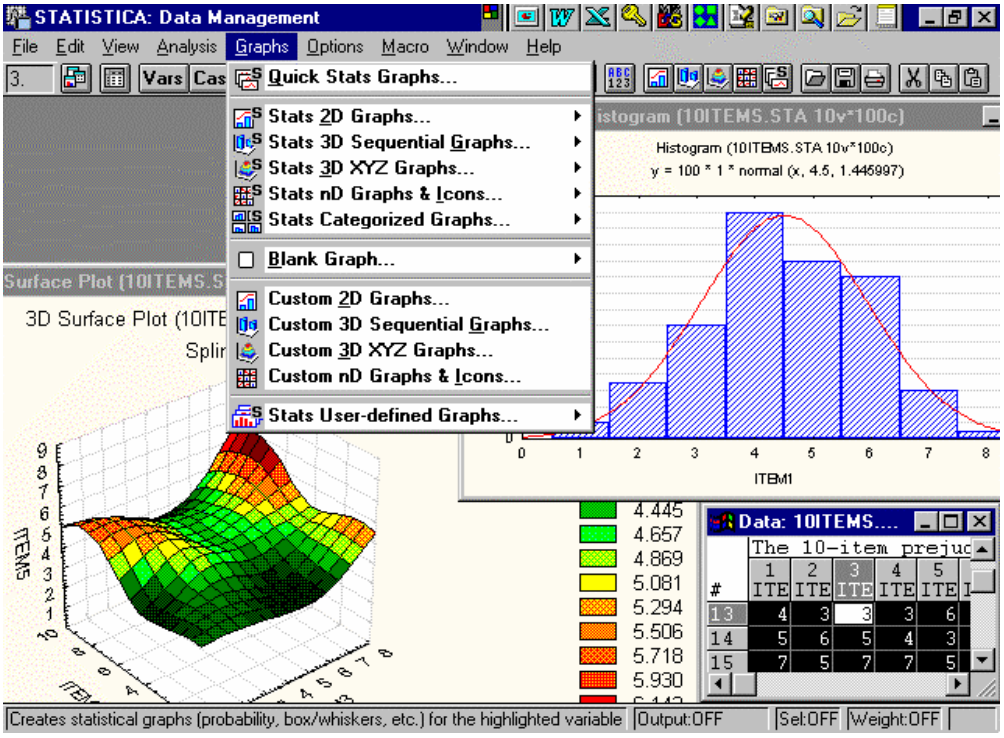
Graphics

The graphics in Statistica are excellent and the graphics capabilities are easily applied to the results of statistical operations. Apart from the specialized graphs, which are automatically offered in the output of statistical procedures, there are two general types of graphs accessible from menus and toolbars (Figure 20). These are:

- custom graphs, and
- stats (and Quick Stats) graphs.

The two types of graphs differ in terms of the source of data which they illustrate. With custom graphs a specified selection of the spreadsheet can be plotted. There are four general categories (custom 2D graphs, custom 3D sequential graphs, …) and after selection a respective dialog box will open allowing you to specify a block of data within the current spreadsheet and specify the type of graph. Stats graphs are pre-defined statistical graphs. They do not depend on highlighted blocks of data as they process data directly from the current data file. They represent either standard methods to graphically summarize data (e.g. various scatter plots, histograms, plots of central tendencies) or standard graphical analytical techniques (e.g. categorized normal probability plots, plots of confidence intervals of regression lines). Quick Stats graphs include the most commonly used types of Stats graphs. There is also the possibility to open a blank graphical screen, which enables one to draw and create figures.

Figure 30 Custom and stats graphs.



In addition to the above some scrollsheets offer other, more specialized statistical graphs (e.g. plots of interaction in ANOVA, icon plots of regression residuals). As mentioned above, specialized graphs which are related to a specific type of analysis are directly accessible from the results dialog. For instance, the descriptive statistics procedure within the Basic Statistics/Tables module offers 2D and 3D scatter surface plots, detrended normal probability plots, 3D bivariate distributions as well as the usual box and whisker plots and histograms.

Once a graph or figure is drawn the ability to edit and customize the presentations appears endless. With a simple double click with the mouse on any feature it can be manipulated in almost anyway you can think of. Statistica's output can be easily imported into Word for Windows and therefore proves invaluable for report writing.

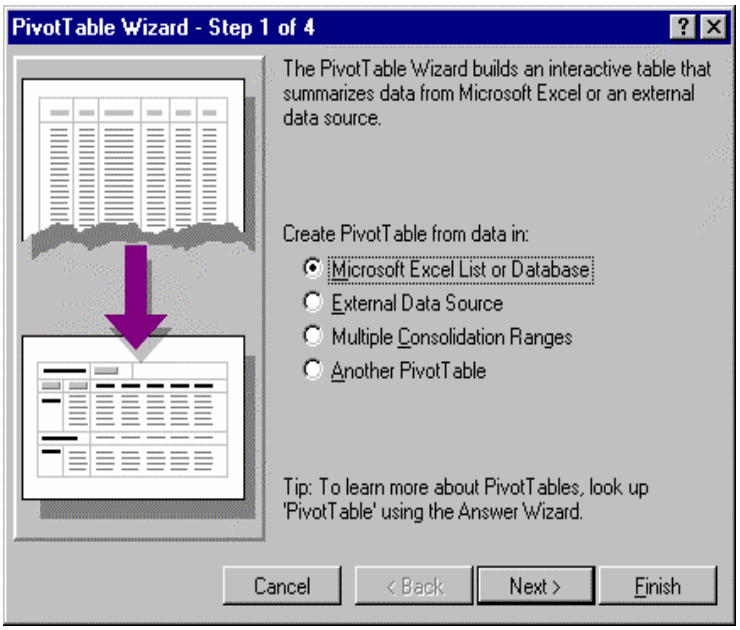
Excel

If you are used to working with spreadsheets, Excel can be an option for statistical work. It has a great deal to offer at the elementary level, however it is of limited use to serious statistical analysis. For that one ought to appreciate the limitations of spreadsheets, programs and consider the use of a proper statistical software package. Excel is currently unsuitable for simple exploratory data analysis; neither does it offer more advanced graphical techniques such as scatterplot arrays and data brushing. There is no stepwise regression, no non-parametric tests, no principle components analysis, no multi-factor analysis of variance, and no generalised linear models, which are all standard features of the statistical packages introduced above.

Data handling

Possibly the most compelling reason for using excel for statistical work is the ease with which data can be entered and manipulated. Data are easily imported from other applications such as Lotus 123, Quattro Pro, dBase, Access and ASCII text files. Excel is particularly good at handling list data, laid out in columns with headers (name farmer, income, assets,…), such as may result from a survey. A *data form* is automatically created and can be accessed via the data menu. A *data filter* can be used to hide all data not satisfying certain criteria, allowing the researcher to focus on particular subsets (e.g. small farmers with less than 2.0 hectares of land). Unfortunately, it is not possible to calculate functions just on filtered data. Data can be sorted in ascending or descending order. In Excel the workbook contains multiple sheets, rendering the workspace effectively three dimensional. Several data sets can be accommodated on separate sheets in the same workbook, along with charts and tables. Switching between the sheets is as easy as turning the pages of a book. Another excellent feature of Excel is the *Pivot Table Wizard* for creating cross tabulations. The data laid out originally in columns can be rearranged in different ways (Figure 21). The row and column labels can be dragged to change the table outlay. Tables can show counts, summary statistics (mean, min, max, stand. dev.) and percentages.

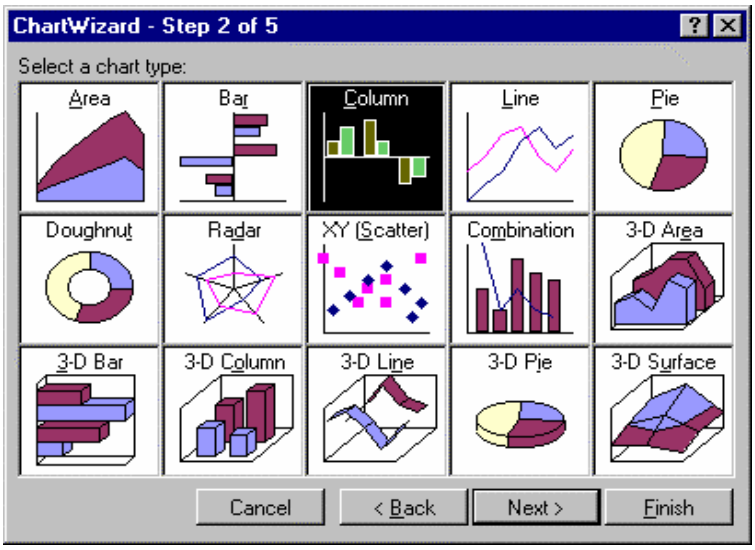
Figure 21 Pivot table.



Graphics

Excel supports a whole range of chart options presented in Figure 22. There are area charts, bar and line charts, pie, doughnut and radar charts, scattergrams and three dimensional charts. However, you won't find boxplots, dotplots, stem-plots or labelled XY-plots. Histograms can only be drawn with equal class intervals and the class limits are positioned at the midpoints.

Figure 22 Excel charts.



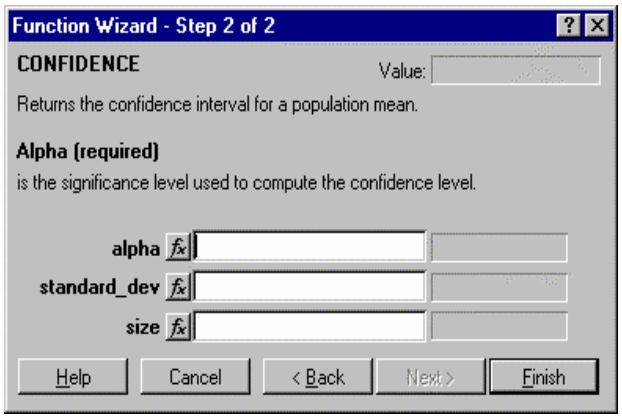
Using the *chart wizard*, creating charts is very straightforward. Charts can be embedded on the same worksheet which contains the data or placed on a separate chart sheet. An exciting feature is the two-way dynamic link between chart and data. Not only does the chart respond to change in the source data, but also the source data change when points on the chart are moved with the cursor. With the trendline option one can insert a trend/regression line of the type linear, polynomial, logarithmic, exponential or moving average.

Statistical functions

Excel has more than 70 statistical functions. These include the usual descriptive statistics and a few less common ones, such as the mean absolute deviation, harmonic mean and general percentiles. For regression analysis there are functions for calculating the various sums of squares and products if you want to calculate from first principles, and functions such as SLOPE, INTERCEPT, RSQ, STEYX which give the regression results straight away. The LINEST function returns the slope and intercept. All main probability functions are present in either density, cumulative or inverse cumulative form - some have all three. These are extremely useful. For example, NORMDIST(1.96) returns the tail probability 0.975 and NORMSINV(0.975) returns the 1.96 Z-value. Statistical tables therefore become redundant.

Luckily names of statistical functions do not have to be memorized. Just click the f_x button and the F*unction Wizard* pops up which prompts for the various arguments and then pastes the function into the selected cell. Figure 23 shows the function wizard box for calculating confidence intervals.

Figure 24 Function wizard.

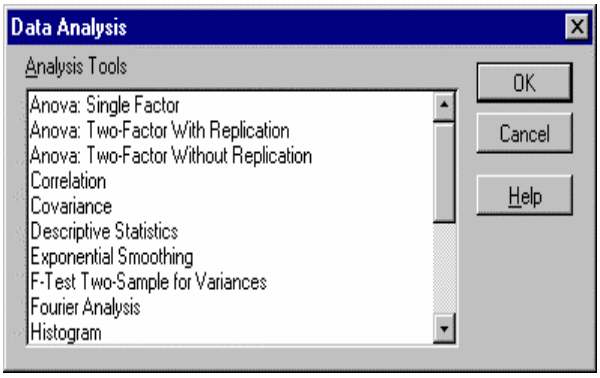


The onscreen help facility is a comprehensive and easy way to search for help on using Excel's statistical functions. It explains the syntax of the functions, contains examples and explains the results.

Add-ins

Excel comes with a range of statistical add-in macros. After selection, the *data analysis* option appears in the tool menu giving a range of analysis tools some of which are displayed in Figure 24.

Figure 25 Data analysis tools.



Add-ins produce non-formula based output which can be displayed on a separate worksheet. A drawback is that one has no indication of how the output is calculated. Thus, undermining one of the great benefits of spreadsheets, namely their transparency - being able to inspect a cell and see what formula generated it. Another disadvantage is that the output is not dynamically linked to the data. Consequently, the output will not be updated when data are changed.

Bibliography

Freese, F. 1967. Elementary Statistical Methods for Foresters. Agricultural Handbook 317. U.S. Department of Agriculture, Forest Service.

Gomez, K.A.; and Gomez, A.A. 1984. Statistical Procedures for Agricultural Research, 2nd edition. IRRI, Los Baños, Philippines.

Hays, W.L. 1988. Statistics, 4th edition. CBS College Publishing, New York.

Kachigan, S.K. 1986. Statistical Analysis: An Interdisciplinary Introduction to Univariate and Multivariate Methods. Radius Press, New York.

Kendall, M.; and Stuart A. 1979. The Advanced Theory of Statistics. Hafner, New York.

Kleinbaum, D.G.; and Kupper, L.L. 1978. Applied Regression Analysis and Other Multivariable Methods. Duxbury Press, Boston, Mass.

McFedries, P. 1994. Excel 5 Super Book, Sams Publishing, Indianapolis.

SAS Institute Inc. 1982. SAS User's Guide: Statistics. 1982 edition.

Snedecor, G.W. 1956. Statistical Methods, Applied to Experiments in Agriculture and Biology. The Iowa State University Press.

SPSS Inc. 1993. SPSS for Windows, User's Guide, Release 7.5. Chicago.

Statsoft Inc. 1995. Statistica for Windows, User's Guide. Tulsa.

Winer, B.J.; Brown, D.R.; and Michels, K.M. 1991. Statistical Principles in Experimental Design, 3rd edition. McGraw-Hill, New York.

Linear Programming and Multiple Goal Linear Programming for Agricultural Planning

Siemon Hollema[*]

Introduction to linear programming

Farmers constantly have to make decisions on *what*, *when*, *how*, and *how much* to grow. These decisions are made subject to available physical and financial resources. From experience, the farmer knows very well when to go to the fields, when to start planting, sowing and harvesting, etc. and, for example, how much fertilizer to apply. In other words the farmer is constantly engaged in regulating his activities and practices planning. In complex planning situations, the technique of linear programming (LP) can be helpful. In its simplest form, LP is a method of determining an optimal combination of farm activities that is feasible with respect to a set of fixed farm constraints.

An example

The easiest way to explain LP is by using an example. This example is taken from Van Ittersum et al. (1997). A farmer with no possibility to irrigate wants to achieve maximum harvestable dry matter production from two crops: wheat and potatoes. One hectare of wheat yields 2 tons dry matter and one hectare of potatoes yields 5 tons.
The farmer faces three constraints:

1. He has only 6 hectares of arable land.
2. Because of government regulations, the farmer is not allowed to grow more than 4 hectares of wheat.
3. On a certain plot, potatoes may not be grown more than once every two years.

Naturally, the areas for wheat and potato production cannot be negative. The mathematical formulation of this problem is as follows:

Maximize	$Z = 2X_1 + 5X_2$	(in tons)	
Subject to:			
	$X_1 + X_2 \leq 6$	(total area constraint)	(1)
	$X_1 \leq 4$	(government regulation on wheat)	(2)
	$X_2 \leq 3$	(crop rotation constraint: potatoes)	(3)
	$X_1 \geq 0$	(non-negative constraint)	(4)
	$X_2 \geq 0$	(non-negative constraint)	(5)

where, X_1 = number of hectares under wheat (ha)

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

X_2 = number of hectares under potatoes (ha)

This is called an optimization model. The objective is Z, the maximum possible harvestable dry matter production. The maximization of Z is restricted by 5 constraints. This problem can also be written in its primal form:

$$\text{Max} \sum_j c_j X_j$$

$$\sum_j a_{ij} X_j \leq b_i, \quad \text{for all i}$$

$$X_j \geq 0, \quad \text{for all j}$$

In LP the following terminology is used. X_j (the area planted for each crop) is called the decision variable, c_j (the respective yields) are the objective coefficients, a_{ij} (the amount of land needed to produce one unit of output) are the technical coefficients or input/output coefficients, and b_i (the amount of land available) corresponds to the constraint. The objective coefficients together with the decision variables form the objective function. The objective function represents the aim of optimization and measures how good a certain combination of decision variables is. The constraints relate the available resources with the resource used.

The use of LP is subject to several restrictions. First of all, the objective functions and constraints should be linear expressions. If the relationships are nonlinear, use can either be made of non-linear programming or the non-linear relationship is split up into several linear segments. Secondly, all parameters have a fixed value and assumed to be known. Thirdly, the variables are real and continuous. There is a technique, called mixed-integer programming, where the variables can also have integer values only.

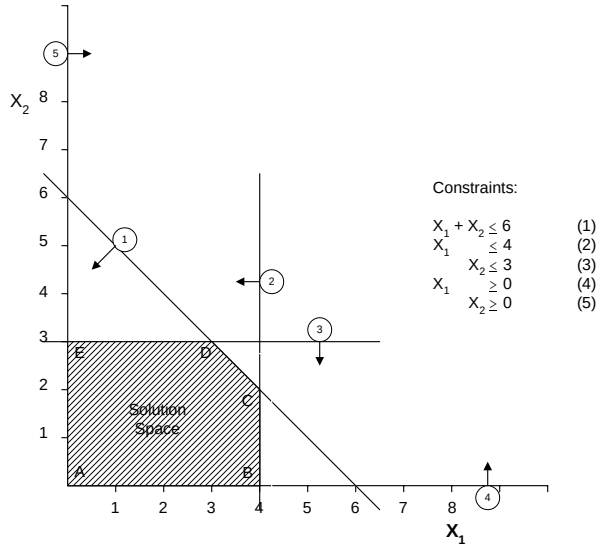
The problem above has only two variables. This means that a solution can be found with help of a graphical method, an algebraic method using the simplex algorithm or using an LP computer software package. As the algebraic method can be quite tedious especially if more variables are involved and since computers are readily available it will not be discussed here. For solving the LP problem we will make use of two different computer software packages. This will be dealt with later, where it will be explained how simple LP problems can be solved by using *Excel*. As *Excel* is a spreadsheet program, it is not meant for complex LP problems. However, as most people are familiar with *Excel* it will be easy to master. GAMS (General Algebraic Modeling System) will also be introduced. GAMS is currently the most popular and versatile software package available for building and solving complex models.

A graphical solution

The farm model can be solved graphically because it has only two variables: the number of hectares planted with wheat and the number of hectares planted with potatoes. For models with three or more variables the graphical method cannot be used. However, the two dimensional presentation allows us to draw general conclusions that will serve as a basis for solving more complex problems.

The first step in the graphical method is plotting the feasible *solution space*. In this area all constraints are satisfied simultaneously. Figure 1 depicts the required solution space.

Figure 1 The solution space.

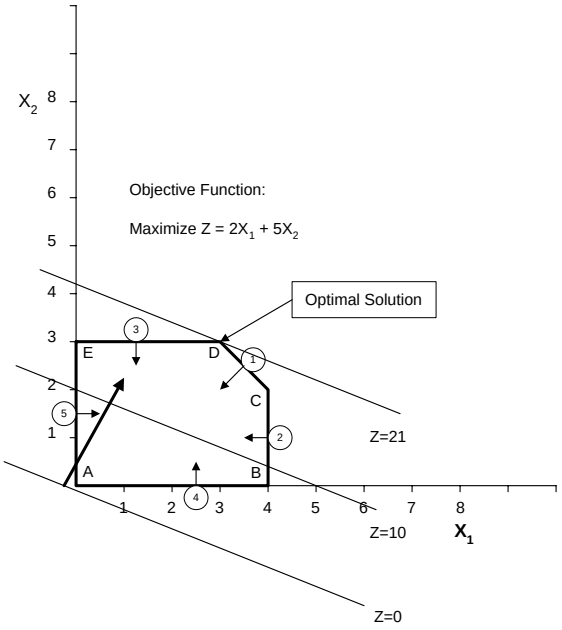


The two variables are plotted on the x- and y-axes. As X_1 and X_2 both face a non-negativity restriction, the solution is confined to the positive quadrant. The space enclosed by the remaining constraints is determined by first changing the ≤ sign to the = sign, thus yielding a straight-line equation. Each straight-line equation is then plotted in the X_1, X_2 plane. The direction of the arrows indicates the activation of the inequalities when each constraint holds. An easy way to determine the direction of the arrows is to use the origin (0,0) as a reference point. If the origin satisfies the inequality, the feasible direction should include the origin; if not, the feasible direction should be on the other side. Applying this procedure to the example yields the solution space ABCDE. The solution of the problem must thus be somewhere in this region, satisfying all the constraints. Although there is an infinity of feasible points in the solution space, the optimal solution can be found by observing the direction in which the objective function increases. This is illustrated in Figure 2, which shows that the optimal solution occurs at point D since moving the objective function further upward would render an infeasible solution. Since point D is the intersection of constraints 1 and 3, the values of X_1 and X_2 are determined by solving the following equations simultaneously:

$$X_1 + X_2 = 6$$
$$X_2 = 3$$

which yields an optimal solution of X_1 and X_2 both 3 of hectares. Consequently, the highest obtainable dry matter production is 21 tons.

Figure 2 Graphical representation of the optimal solution.



It can be proven that the optimal solution will always occur at one of the feasible corner points. The choice of the specific corner will, in the first place, depend on the slope (the coefficients) of the objective function. The fact that a solution will always be in one of the corner points is the key idea in solving LP problems in general. It means that we don't have to concern ourselves with the fact that the solution space has an infinity of solutions but can concentrate on a finite number of corner points.

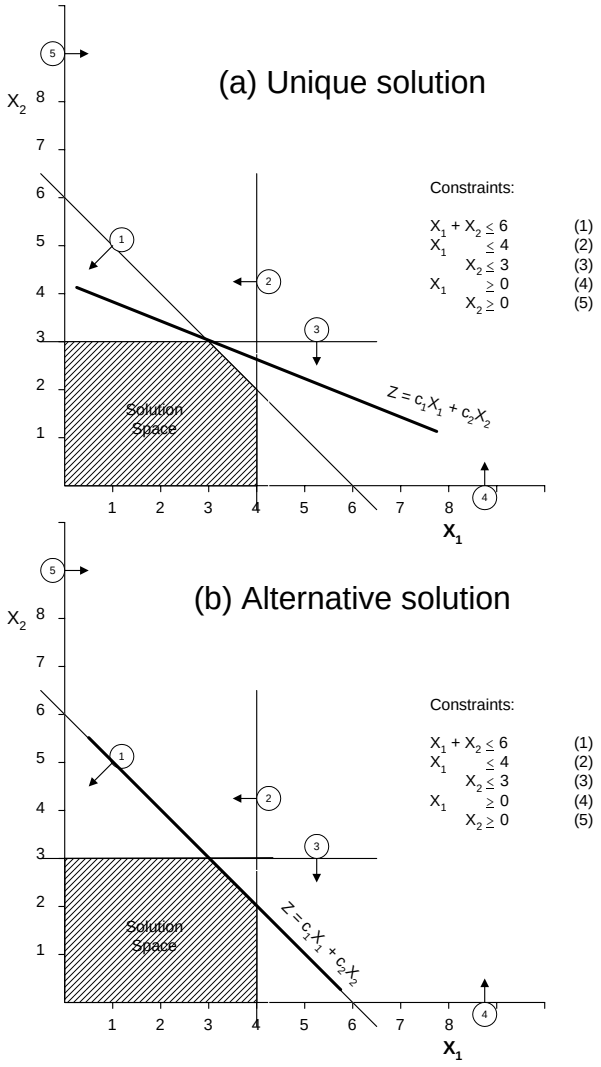
Types of solutions

An LP problem can yield 4 types of solutions:

i) a unique solution
ii) an alternative solution
iii) an unbounded solution
iv) no feasible solution

Each type is illustrated in Figure 3.

Figure 3 Types of solutions.



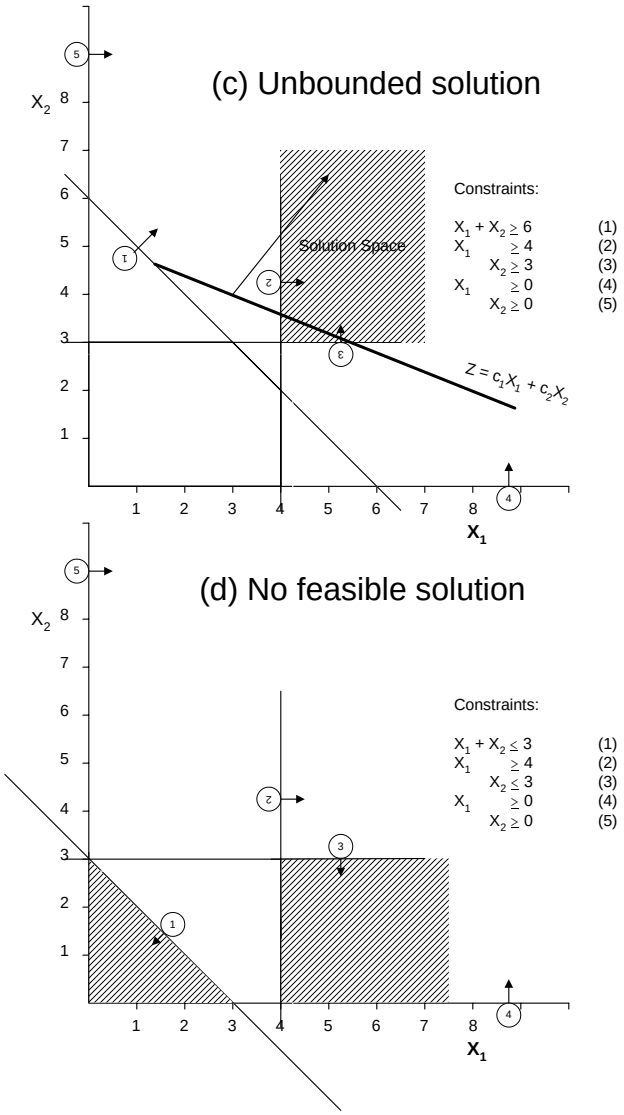


Figure 3(a) shows a unique solution. There is only one combination of variables which yields the maximum outcome. This is the ideal output of an LP problem. Figure 3(b) shows the possibility of alternative solutions. The objective coefficient c_1 and/or c_2 have changed and the objective function now coincides with the first constraint. There are now many possible solutions for X_1 and X_2, all yielding the same optimal outcome. Observe, however, that the corner solutions C and D are still valid alternatives. In Figure 3(c), the first three inequality constraints are changed from smaller than or equal to, to greater than or equal to. Consequently, the solution space is not bounded. This means that the value of the objective function can infinitely increase without being restricted by the constraints.

The last figure shows the possibility of an infeasible solution. No point can be found which satisfies all constraints simultaneously. Consequently, a solution to the problem is not possible.

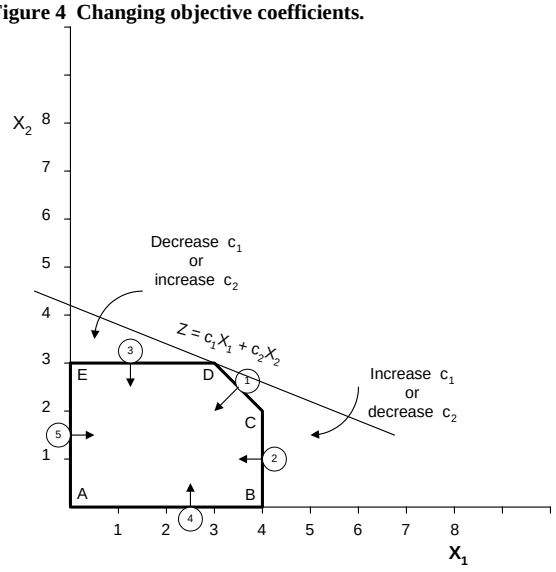
Sensitivity analysis

Sensitivity analysis is a very important part of linear programming. It serves two goals. First of all, data included in LP models are often not very accurate. If the errors in the data used are only marginal this might not be such a problem. Also, when errors are considerable but do not change the optimal solution we don't have to worry. But if a certain figure which is not exactly known plays a crucial role in determining the solution we have to investigate how the solution changes when the value of the parameter changes. Secondly, sensitivity analysis is also used to investigate intentional changes in the values of the parameters. For example, what happens if prices change or if more land can be used, etc. In this section a graphical procedure will be used to explain the basic elements of the technique.

How much change is allowed in the objective coefficients?

A change in the objective coefficients affects the slope of the objective function. This is illustrated in Figure 4. It shows that the effect of an increase/decrease in c_1 and c_2 is to rotate the line representing the objective function in a clockwise or counterclockwise direction around the current optimum point D. The figure shows that, as long as the slope of the objective function remains between those of line DE and CD, the optimum will not change and will stay at point D. This is termed c*oefficient ranging*. If the slope coincides with the line DE or CD, alternative solutions are possible. Further changes in the objective coefficients will change the solution to point C or E. A related term is called *reduced cost*. Imagine the solution is found at point E. This means that the farmer will only grow potatoes. Reduced cost will then indicate how much the yield of wheat has to increase so that the farmer will also grow wheat. In other words, reduced cost indicates the amount the objective coefficients have to increase or decrease in order that the corresponding variable takes place in the optimal solution. The reduced costs in the example are both zero as both crops will be grown.

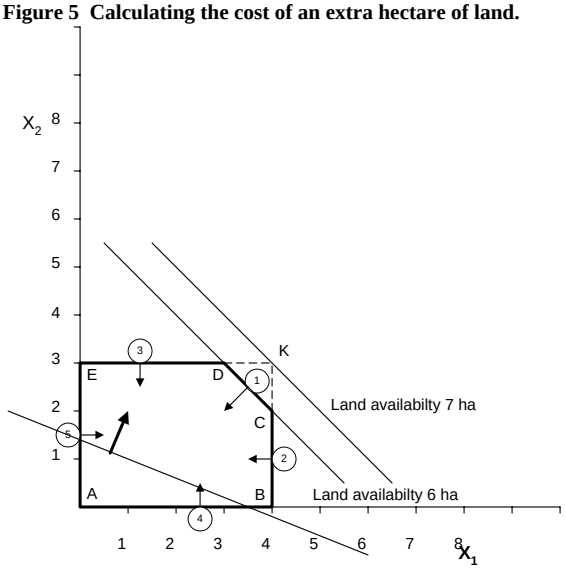
Figure 4 Changing objective coefficients.



What is the cost of an extra unit of resource?

This part of sensitivity analysis deals with changes on the right-hand side of the constraint equations. If the right-hand side represents a limited resource (for example, land), what is then the effect of changing the availability of this resource on the value of the objective function? Consider the first constraint expressing the availability of land (6 hectares). Any change in the availability of this resource will cause the associated constraint line 1 to shift in a parallel way. This is shown in Figure 5 where 7 instead of 6 hectares of land are available. The result of an increase in the availability of land is that the value of the objective function increases at a constant rate as long as the optimum solution is determined by the intersection of line 1 and 3. The value of an extra unit of land is equal to the increase in the value of the objective function. This is called the *shadow price*. If the number of hectares is increased past the amount associated with point K, the constraint will be redundant, as the optimal solution will then be determined by the lines 2 and 3. Hence, any further increase in the number of hectares available will have no affect on the value of the objective function and will be of no value to the farmer. Similarly if the amount of land is decreased below the value associated with point B, the optimum solution will no longer be determined by the intersection of line 1 and 3. The linear relationship between the optimum objective value and the changes in the resource land will be destroyed. From Figure 5 it is evident that the second constraint (government regulation on wheat) can be increased to a maximum of three hectares grown with wheat without changing the optimal solution. Consequently, any increase/decrease in the range $(3, \infty)$ will not affect the optimal solution or the optimal value of the objective function. This is called the *right hand side ranging*. The corresponding value per unit in this range is zero.

Figure 5 Calculating the cost of an extra hectare of land.



Solving the model on computer

Solving linear programming problems in Excel

Linear programming problems can be solved in *Excel* using the *Solver* option. The general guidelines for solving LP problems are as follows.

Setup

Load the spreadsheet program *Excel*. The *solver* is listed under the *Tool* menu. If not, you must install the *Solver*.

Linear programming with Excel.

Consider the linear programming problem of maximizing $Z = \sum_{j=1}^{n} c_j X_j$,

subject to $\sum_{j=i}^{n} a_{ij} X_j \leq b_i$ all i = 1 to *m*

and $X_j \geq 0$ all *j* = 1 to *n*

For a given farm situation the symbols can have the following meaning:

X_j = the level of the *j*th farm activity, such as the acreage of maize grown. *n* denotes the number of possible activities.

c_j = the expected net selling price (e.g. dollars per acre maize grown).

a_i = the quantity of the *i*th resource (e.g. acres of land or days of labour) required to produce one unit of the *j*th activity. *m* denotes the number of resources.

b_j = the available quantity of the *i*th resource (e.g. acres of land or days of labour).

There are many ways to formulate this in *Excel*. The conventional method is shown below in a so-called *linear programming tableau*.

Row name	Columns						
Farm activity	X_1	X_2	.	.	X_n	LHS	RHS
Selling price	c_1	c_2	.	.	c_n		Maximize
Resource constraints:							
1 Land	a_{11}	a_{12}	.	.	a_{1n}		$\leq b_1$
2 Labour	a_{21}	a_{22}	.	.	a_{2n}		$\leq b_2$
.	.	.		.	.		.
.	.	.	.	.	.		.
m	a_{m1}	a_{11}	.	.	a_{mn}		$\leq b_m$
Objective							……..

The data matrices c_j, a_{ij}, b_i (these are the givens) as well as the variable vector X_j (this is the variable) can be displayed exactly in this way in *Excel*.

Load a worksheet.

- Choose Units. Choose the units for measuring things relatively small and reasonably comparable. For example, if revenue in the data set is typically a few million dollars, measure revenue in millions of dollars rather than dollars.
- Enter the Data Matrices. Type in the data as shown in the preceding tableau. Thus, data c_j (selling price) is typed in as a row; the maximum available quantity of resource i (b_i) is typed in as a column on the right hand side (RHS); and the required amount of each resource to produce one unit of the *j*th activity (a_{ij}) forms a matrix.
- Enter the Variable Vector. Just type in any convenient values into the cells X_1 to X_n. For example use the value 1. *Excel* calls these cells *changing cells*.
- Define the Objection function. Identify a cell, for example the one in the lower right hand corner which shows "……." in the tableau, to contain the o*bjection function*. Position the mouse there and define the function

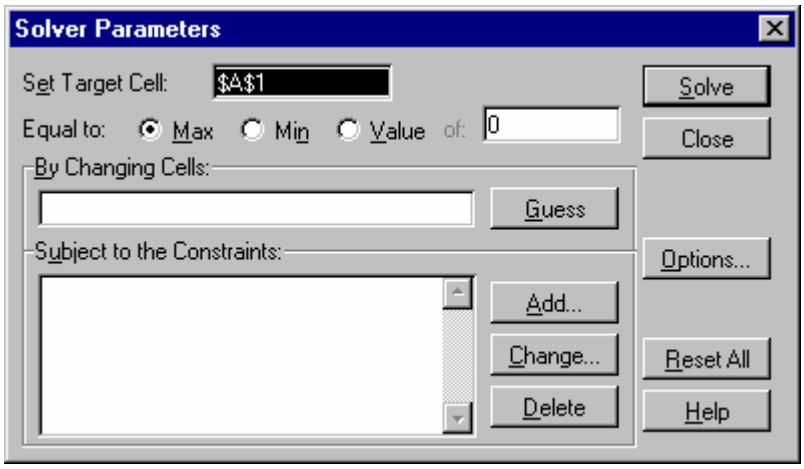
  $$Z = \sum_{j=1}^{n} c_j X_j$$

  using *Excel*'s matrix function, **=SUMPRODUCT (array 1, array 2)**. This function returns the sum of the products of the corresponding array components. The arrays are selected by dragging the mouse over the ranges to be multiplied. In this case array 1 is range $c_1 \ldots c_n$, and array 2 is range $X_1 \ldots X_n$.
- Define the Constraint Function. Identify a range of cells to contain the constraint function $\sum_{j=i}^{n} a_{ij} X_j$

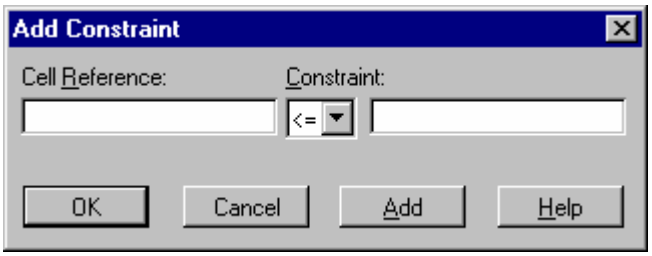
  in the column LHS. Define the function in each cell using the *Excel* matrix function **=SUMPRODUCT (array 1, array 2)**. For example the land constraint is calculated by multiplying range $a_{11} \ldots a_{1n}$ with range $X_1 \ldots X_n$.
- Save the worksheet. Under the *file* menu, the Save As… option appears. Give the worksheet a name and save it accordingly.

Load the Solver

With the Worksheet displayed, load the *Solver* from the *Tool* menu. This brings up the *Solver Parameter* dialog box.

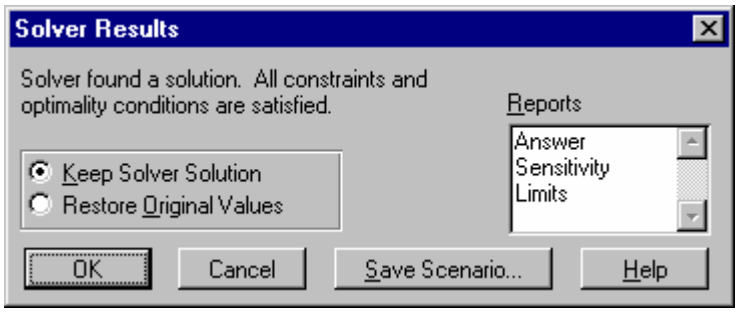


- Objective Function. Select with the mouse the cell which contains the *Objection Function* (the cell in the lower right hand corner) in the *Set Target Cell* box.
- Maximization/Minimization. Fill in the *Equal to* line, by clicking the circle **O** just before Max or Min. This assures that the objective function is to be maximized or minimized.
- Variables. In the *By Changing Cells* box, select the range X_1 … X_n with the mouse. This assures that the level of activities is changed to their optimal level.
- Constraints. To add a constraint, click on the Add.. button. *Excel* displays the *Add constraint* dialog box.



- The middle box specifies whether the constraint is ≤, =, or ≥. The left-hand box specifies the LHS and the right-hand box specifies the RHS of the constraint. Make the appropriate selection in the middle box for the constraint at hand. Then enter the cells of the left- and right-hand sides of the constraint in the other two boxes and click the OK button. The constraint then appears in the *Subject to the Constraints* box. Repeat for each constraint. For example: to enter the constraint x≥0, click the Add button, select ≥ from the middle box, position the mouse in the left box and select the range X_1 … X_n there. Next, position the mouse in the right box and type 0 there. Click the OK button to complete the selection. The inequality x≥0 then displays. If a mistake is made, select a constraint to be changed or delete and click the Change… button or Delete button, depending on the objective.
- Options. Click the Option… button and select the *Assume Linear Model* option if this has not been done already. Close the *Option* dialog box by clicking OK.

- Solve Problem. Now the problem is ready to be solved. Click the Solve button. If the *Solver* declares that it has found a solution, consider the next step. If not, several things might have gone wrong. Errors might have occurred which must be corrected. Another possibility is that you have run out of time or number of iterations. If so, they will have to be increased in the *Option* dialog box (see Options).
- Reports. If the *Solver* declares it has found a solution, a *Solver Results* dialog box appears on your screen.



It is possible to produce three reports, viz., *Answers report, Sensitivity report,* and *Limits report*. Just select the reports desired with the mouse and click OK. *Excel* displays each report on a separate sheet.

The *Answer Report* displays information about the model's target cell, changing cells and constraints. For the target cell and changing cells, *Solver* shows original and final values. For the constraints, the report shows the final value and two values called the *binding* and the *slack*.

The *Sensitivity Report* attempts to show how sensitive a solution is to changes in the model's formulae.

The *Limits Report* displays the target cell and its value, as well as the changing cells and their values, upper and lower limits, and target results.

- Printing. Save the reports to file and print the Worksheet and Report sheets.

Solving the example

Following the steps above, the spreadsheet for our LP example will, after we have solved the problem, appear as follows:

	A	B	C	D	E	F	G
1							
2		**Activity**	**X1**	**X2**	**LHS**	**RHS**	
3		ha	3	3		**Maximize**	
4		Yield	2	5			
5							
6		**Constraints:**					
7		Land constraint	1	1	6	6	
8		Market constraint	1		3	4	
9		Rotation constraint		1	3	3	
10		Non-neg.	1		3	0	
11		Non-neg.		1	3	0	
12							
13		**Objective**				21	
14							

This shows a similar result to the graphical solution. The areas planted with wheat and potatoes are both 3 hectares and the value of the objective function is 21 ton. This can also be seen in the answer report created by *Excel*.

Microsoft Excel 7.0 Answer Report
Worksheet: [Book1]Sheet1
Report Created: 9/15/97 15:22

Target Cell (Max)

Cell	Name	Original Value	Final Value
$E$13	Objective maximize	7	21

Adjustable Cells

Cell	Name	Original Value	Final Value
$B$3	ha X1	1	3
$C$3	ha X2	1	3

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$D$7	Land constraint	6	$D$7<=$E$7	Binding	0
$D$8	Market constraint	3	$D$8<=$E$8	Not Binding	1
$D$9	Rotation constraint	3	$D$9<=$E$9	Binding	0
$D$10	Non-neg.	3	$D$10>=$E$10	Not Binding	3
$D$11	Non-neg.	3	$D$11>=$E$11	Not Binding	3

The first section of the answer report shows the objective function, its cell reference, the original value and the maximum value. The decision variables (adjusting cells) were original given a value of 1. Consequently, the original value of the objective function was 7. In the second section, these values have changed to their optimum value of 3. The last section gives the constraints which are binding and non-binding. The non-binding constraints have a slack value, indicating the unused amount.

The next report created is the sensitivity report. This one of the most interesting parts of *Excel*'s output. The first section shows the decision variables and their coefficients. The reduced cost for both variables is zero. Reduced cost is a term used in linear programming to indicate how much the objective coefficient has to increase/decrease in order for the variable to take place i.e. to have a positive value in the optimal solution. As both crops are produced, the reduced cost of each is zero. The last two columns show the allowable increase and decrease in the objective coefficient without changing the optimal solution, as discussed earlier with the graphical solution.

The second section deals with the constraints. Its shows the level of the resource used, i.e. the value of the inequality constraint and its associated shadow price. The shadow price is the marginal value of a resource. It indicates how much the value of the objective function will increase when the constraint is relieved with one unit. Only the land constraint and the market constraint have a shadow price since they are the only binding constraints. The last two columns show the allowable increase and decrease in the right-hand side of the equation without changing the optimal solution as discussed before.

Microsoft Excel 7.0 Sensitivity Report
Worksheet: [Book1]Sheet1
Report Created: 9/15/97 15:22

Changing Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$B$3	ha X1	3	0	2	3	2
$C$3	ha X2	3	0	5	1E+30	3

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$D$7	Land constraint	6	2	6	1	3
$D$8	Market constraint	3	0	4	1E+30	1
$D$9	Rotation constraint	3	3	3	3	1
$D$10	Non-neg.	3	0	0	3	1E+30
$D$11	Non-neg.	3	0	0	3	1E+30

The last report which can be created in *Excel* is the Limits Report. It simply shows the value of the objective function and the value of the decision variables together with their lower and upper limits and contribution to the target result.

Microsoft Excel 7.0 Limits Report
Worksheet: [Book1]Sheet1
Report Created: 9/15/97 15:22

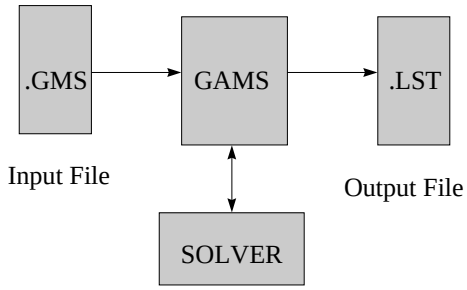
	Target	
Cell	**Name**	**Value**
$E$13	Objective maximize	21

	Adjustable		Lower	Target	Upper	Target
Cell	**Name**	**Value**	**Limit**	**Result**	**Limit**	**Result**
$B$3	ha X1	3	1.66978E-12	15	3	21
$C$3	ha X2	3	1.66978E-12	6	3	21

Modeling in GAMS

One of the most popular mathematical programming packages is GAMS (General Algebraic Modeling System). In this section we briefly describe how to set up the above linear programming example for GAMS. For more information, refer to the Chapter on Mathematical Programming in GAMS: A Course Manual for Agricultural Planning. GAMS is a file-based system. This means that an input file is created with an editor of your choice. Next GAMS is called upon and it checks the input file for possible mistakes. If no mistakes are found, a solver is used to solve the model. Irrespective if mistakes have been made or not, GAMS will create an output file with the extension *.lst* (Figure 6).

Figure 6 Overview of GAMS.



To set up an LP problem, GAMS makes use of certain keywords which determine the structure of the problem. The keywords are:

SET	:	defines the indices of the model
SCALAR	:	used for data entry (not defined over a set)

PARAMETER	:	used for data entry (defined over one set)
TABLE	:	used for data entry (defined over two or more sets)
VARIABLE	:	defines the decision variables
EQUATIONS	:	defines the equations
MODEL	:	identifies the model
SOLVE	:	solves the model

For most of these commands the format is similar e.g.:

```
KEYWORD     NAME(DOMAIN)        TEXT   /DATA/ ;
```

First, a keyword is used to open the section. Then the set (parameter, scalar, etc.) is declared by giving it a name and defining its domain. Optional explanatory text is possible. Data and set elements are defined using two forward slashes. Each section is ended with a semicolon. In GAMS the LP example has the following setup:

```
SET
j       Crops                           /wheat, potatoes/
i       constraints                     /land, government, rotation/;

PARAMETERS
c(j)    yield (objective coefficient)   /wheat 2, potatoes 5/
b(i)    constraints                     /land 6, government 4, rotation 3/;

TABLE
a(i,j)  input-output coefficients
                wheat   potatoes
land            1       1
government      1
rotation        1               ;

VARIABLES
X(j)    area planted for each crop (decision variable)
Z       Harvestable dry matter production (objective variable);
POSITIVE VARIABLE X;

EQUATIONS
OBJECTIVE           The objective function
CONSTRAINT(i)               The three constraints;

OBJECTIVE..         SUM(j,c(j)*X(j))=E=Z;
CONSTRAINT(i)..             SUM(j,a(i,j)*X(j))      =L=B(i);

MODEL EXAMPLE /ALL/;
SOLVE EXAMPLE USING LP MAXIMIZING Z;
```

Two sets are declared, one for the crops (j) and one for the constraints (i). For data entry parameter and table statements are used. There are two parameters: yield (c(j)) and land availability (b(i)), each with its own domain i.e. respectively the crops and constraints. The table statement is used to define the input/output coefficients, which in this case are either 1 or zero. Two variables are being declared: one decision variable (X(j)) and one variable which is to be maximized. As areas cannot be negative, X is limited to positive values. Equations are defined in two parts. In the first part each equations is declared by giving it a name and in the second

part the mathematical relationship is been defined. The model statement gives the LP example a name and groups the equations together and, finally, the solve statement solves the model.

After the model is solved, GAMS creates an output file with the extension .lst. This file can be opened using an editor. The output file consists of two parts: a compilation part and a solution part. The compilation part is the output created during the initial check of the input file. The solution part is the output created by the solver. Here the answer to the problem will be found. Part of the solution report is shown below.

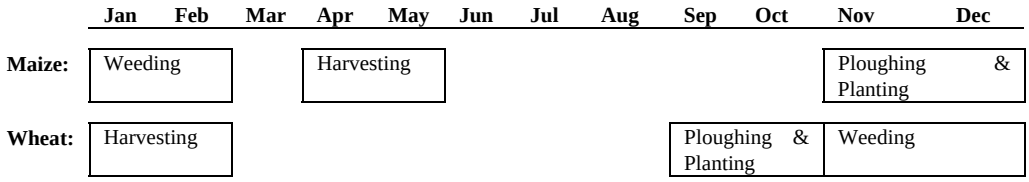
```
GAMS 2.25.087  386/486 DOS                    04/21/98 11:59:15  PAGE     7
EXAMPLE LINEAR PROGRAMMING
Solution Report  SOLVE EXAMPLE USING LP FROM LINE 31

---- VAR X       area planted for each crop (decision variable)

                 LOWER          LEVEL UPPER MARGINAL

WHEAT            .          3.000  +INF    .
POTATOES         .          3.000  +INF    .


                 LOWER          LEVEL UPPER MARGINAL

---- VAR Z       -INF       21.000  +INF    .

 Z               Harvestable dry matter production (objective variable)


**** REPORT SUMMARY :      0  NONOPT
                           0  INFEASIBLE
                           0  UNBOUNDED
```

The optimal level for each agricultural activity *i.e*. the area planted with wheat or potatoes is 3. This maximizes the objective variable Z which shows a value of 21 tons.

Exercise in linear programming

Imagine a village in Southeast Asia with around five hundred inhabitants. An average family, which consists of 5 people, owns 4 ha of land and 1 ox-plough. The farmer produces maize and wheat. In order to maximize net revenue the farmer has to decide how to divide his land between these two crops.

The following information is available. The timing of agricultural activities for maize and wheat is shown in below.



Ploughing & Planting: Making use of an ox-plough it takes 7 days of 4 hours to plough one hectare of land. Ploughing needs to be down twice. Once for the clearing of the land and once before planting. Labour necessary for clearing, preparation and planting is as follows:

i) Maize: 150 man-hours per hectare
ii) Wheat: 100 man-hours per hectare
iii) Weeding: Weeding takes on average about 200 man-hours per ha. This is the same for maize and wheat.
iv) Harvesting: Harvesting is done manually. It takes 180 man-hours to harvest one hectare of maize and 150 man-hours for wheat.

Based on the available data on yields by development region, the yields of maize and wheat are estimated to be 1,800 kg/ha and 1,000 kg/ha respectively. Based on 1991 - 1994 prices, the selling price of 1 kg of maize is 3,000 Rp and wheat 5,000 Rp.
The costs of production, including seed costs and use of ox-plough and manure, for maize are 4,000,000 Rp/ha, and for wheat are 2,500,000 Rp/ha.

Only one person of the family is farming full-time (8 hours a day), while 2 of them are too young to work and two are available for agricultural activities only 4 hours a day. One month has 25 working days.

Exercise I. Formulate a linear programming model to maximize the total net revenue of the family. Solve it on the computer using either Excel or GAMS.

Exercise II. Due to difference in altitude, the land available for maize and wheat consists of two parcels each of 2 ha with different soil fertility. On one of them the yields of crops and inputs are the same as before, while on the other parcel the yield of wheat is 20% higher. Labour inputs for planting and weeding for both crops are 20% higher, while the labour input for harvesting wheat increases 10%. It takes one extra day of 4 hours to plough the land. The operating costs on this area are 4,500,000 Rp/ha for maize and 3,000,000 Rp for wheat.

Modify the linear programming model and solve it.

Exercise III. For the production of maize use can be made of a fertilizer instead of manure. This will increase the yield of maize 30% i.e. to 2,340 kg/ha. The cost of manure is 500,000 Rp/ha while the cost of fertilizer is 1,500,000 Rp/ha. If fertilizer is used more labour is needed for preparing the land, namely 250 man-hours per hectare. Also labour for harvesting increases to 200 man-hours per hectare. Modify the linear programming model of exercise I to investigate if fertilizer should be used.

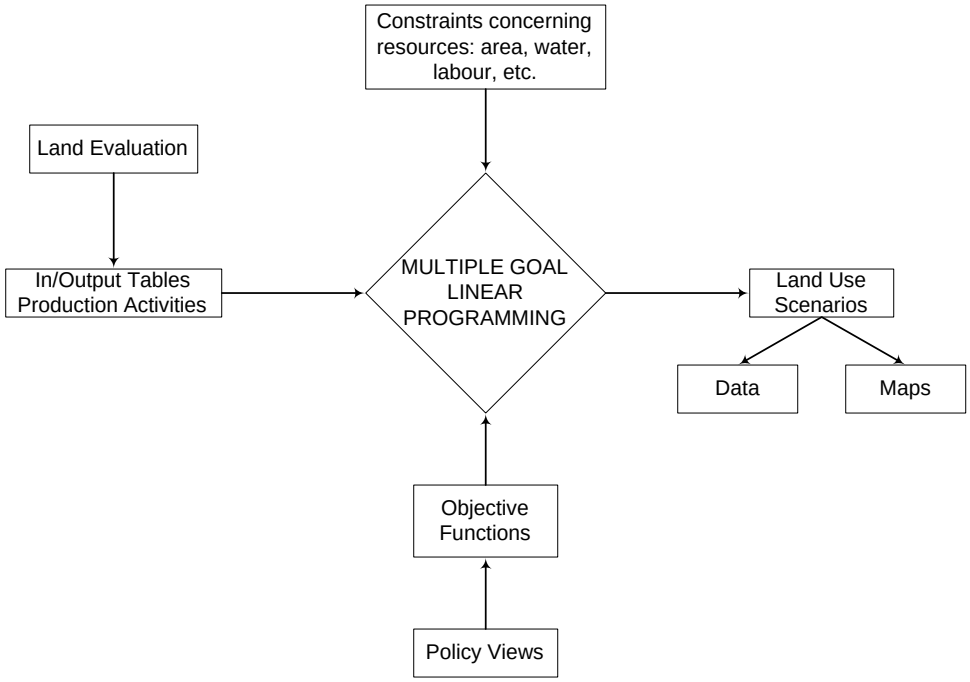
Solutions to the exercises can be found in Appendix 2.

Introduction to multiple goal linear programming in land use analysis

In the methodology for explorative land use studies, linear programming techniques can be used, which helps to select the best options from the alternative uses of land. The difference between linear programming (LP) and multiple goal linear programming (MGLP) is that in LP only one objective is maximized while in MGLP more objective functions are defined which are maximized successively until a satisfying solution is found for all objective functions. The reason that more objective functions are considered is that there are conflicting views of how to best utilize the land. As land becomes increasingly scarce, these conflicting interests become

more and more apparent. MGLP reveals and quantifies the trade-offs between different perceptions of how to utilize land. The consequences of different land uses can be made clear by creating scenarios. MGLP is thus a strong tool for analyzing different policy measures related to land use (Figure 7).

Figure 7 The general building blocks of MGLP for explorative land use studies.



Policy views and objective functions

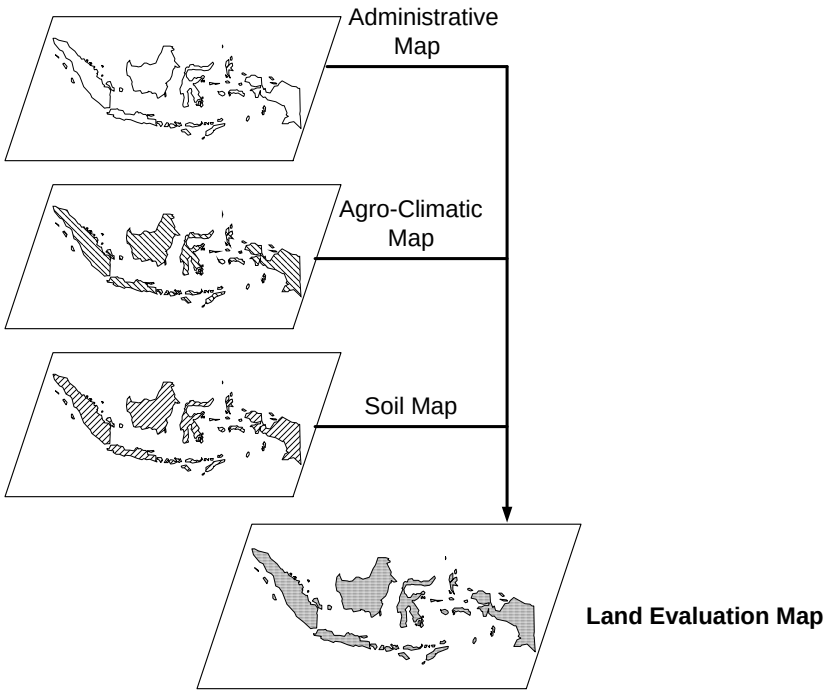
Various policy views concerning land use in the region under study should be identified. For example, self-sufficiency in food supply, environmental issues, nature conservation, employment generation, etc. They can be distilled from policy documents, public opinion, interviews with policy makers and representatives of social organizations, etc. The next step is to operationalize these views by means of objective functions which can be maximized or minimized. The objective functions used in the MGLP procedure must be to a certain extent mutually conflicting, because, if maximizing one objective automatically causes the maximization of other objectives, they don't need to be included. On the other hand, if the objective functions are totally mutually conflicting, the results will be meaningless, as a gain in one objective will automatically mean a loss in the other.

Land evaluation

Land evaluation units (LEU) are employed in land use studies. Every unit has a unique combination of administrative region, soil type and agro-climatic conditions. The different land evaluation units can be established using a geographical information system. Figure 8 shows the

basic procedure. An administrative boundary map is combined with an agro-climatic map and a soil map.

Figure 8 Map overlay procedure (GIS - Indonesia).



In the next step, the characteristics of soil and climate in each LEU are confronted with the requirements for various forms of land use. This confrontation can be a qualitative one, saying that a certain LEU is suitable or not for a certain land use, or a quantitative one, which specifies how much of the LEU is suitable for a certain form of land use.

Quantification of input/output coefficients

The input/output coefficients represent the quantification of different production activities in a specific LEU. They tell us how much of the required input is needed to produce one unit of a certain output given the climatic and soil conditions. In an explorative land use study, potential input/output coefficients are used not the observed ones. The question we have to ask ourselves is: What is possible?

Constraints

The constraints contain the information on the availability of resources, such as land, labour, water, capital, etc. They can also be used to quantify more normative constraints like food or income requirements.

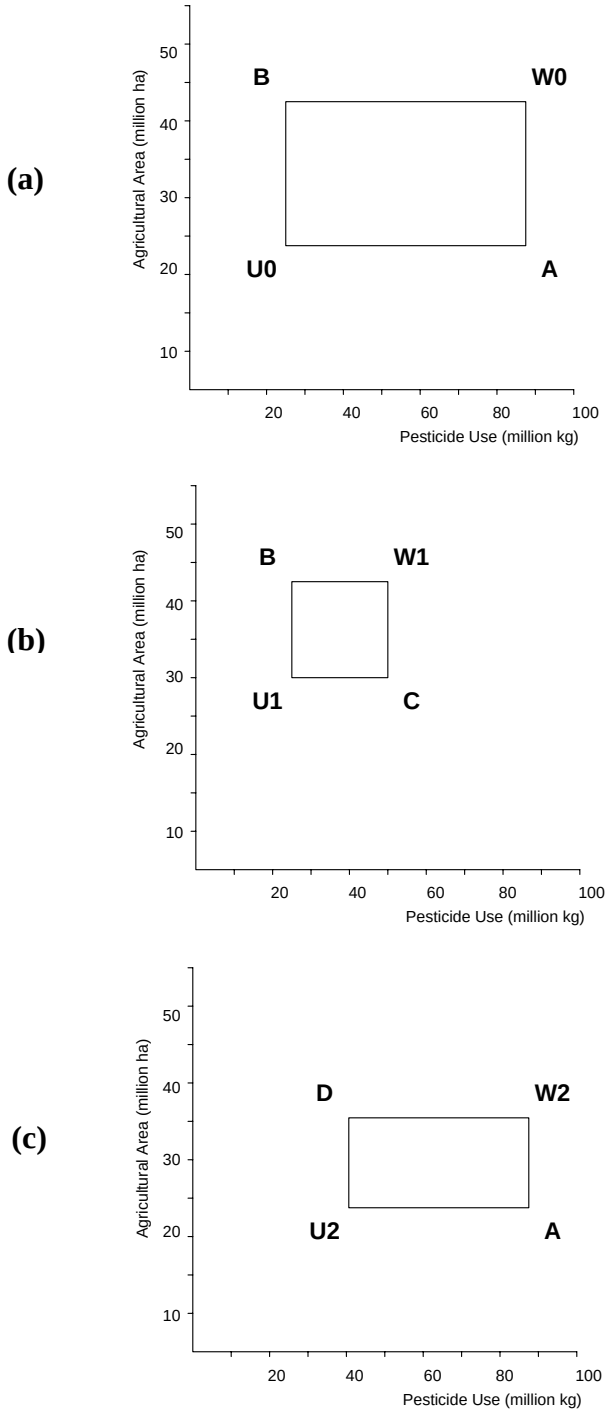
MGLP

The MGLP procedure consists of a number of optimization rounds. Each round comprises several optimizations. In each run one objective function is optimized while the others serve as constraints. Upper and lower limits can be put on these “goal constraints”. A MGLP procedure starts with a so-called zero round. Each objective function is optimized without putting any upper or lower boundary on the goal constraints. In this zero round the feasible space, or the playing field, is established. The extreme values of the objective functions marking the playing field are important for choosing upper and lower limits on goal constraints in scenarios.

Next, the worst and best value of each objective function is selected. In this way the initial freedom of choice, i.e. the best value minus the worst value, is made explicit to each stakeholder. The next step is to select the objective with the worst value which is considered by the stakeholder as unacceptable. For this objective function an upper or lower boundary is formulated. Subsequently, each stakeholder is confronted with the results of a new series of optimization runs. Again the objective with the worst unacceptable value is selected and lower or upper boundaries are formulated. This procedure continues until ideally each stakeholder is satisfied. A possible solution is thus found in an interactive way. That is why this procedure is sometimes also referred to as interactive multiple goal linear programming (IMGLP). The IMGLP procedure can be illustrated using an example with only two objective functions (i) minimization of the use of pesticides and (ii) minimization of agricultural area (Spharim et al. 1992).

The results for the zero round are presented in Figure 9(a). The minimum amount of agricultural land without placing an upper limit on the use of pesticides is 24 million ha. The consequent use of pesticide is then 88 million kg (point A). The minimum amount of pesticide used without an upper limit on agricultural land is 22 million kg. This coincides with an agricultural area of 43 million ha (point B). If A and B coincide, both objectives are completely tied. This means that maximizing one will automatically realize the other. There is thus no conflict between the objectives. Point W0 represents the amount of pesticide used when the agricultural area is minimized and the agricultural area when the use of pesticide is minimized. This point represents the worst combination. A stakeholder does not have to accept a worse outcome. Point U0 is the best combination; it combines the lowest pesticide use with the smallest agricultural area. Unfortunately, however, this combination is impossible to realize as we are dealing with two partially conflicting objectives.

Figure 9 Multiple goal linear programming example.



Now assume that the stakeholder does not accept a pesticide use of more then 50 million kg. The most unfavorable combination of objective achievement is then W1 (50, 43) as in Figure 9(b). Point C is determined by a second optimization run, minimizing the use of agricultural area given the upper limit on pesticide use. The minimum area is now 30 million ha. The point representing the best outcome of both objectives will consequently move up (point U1).

In Figure 9(c), it is assumed that the stakeholder is satisfied with an agricultural area of 35 million ha. The minimum use of pesticide is then 40 (point D). If the stakeholder is now satisfied with the solution the procedure stops. If not it continues until an appropriate solution is found.

If more objectives are included, the basic principle is the same as with two objectives, however, the number of optimizations which are needed to arrive at a satisfactory solution increases rapidly with the number of objectives. It is then convenient to present the outcomes for each optimization round as illustrated below (Table 1). In this table the first objective has been limit, while the boundaries on the others are equal to their worst values.

Table 1 Outcome table for MGLP example with multiple objectives.

Round no. …	Upper / lower boundary	Results of the optimization for objective			Worst value	Best value
		1	i	n		
Objective 1	< or > X	B1	.	.	X	B1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Objective i	< or > Wi	.	Bi	.	Wi	Bi
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Objective n	< or > Wn	.	.	Bn	Wn	Bn

Scenarios and presentation

By giving more or less weight to the various policy objectives, different scenarios can be developed. By doing this the trade-offs between the different policy goals become clear. Data can be presented in tables and graphs, and maps can be created to illustrate the implications.

Solving a case study

The assignment is to undertake an explorative land use study in region ‘X’, an area located in West Java, Indonesia. The aim is to reveal and quantify the trade-offs between different perceptions of regional development. Various policy views concerning land use can be identified from policy documents issued by government and donor organizations. Region ‘X’ consists of a unique ecosystem. Consequently, there is a high pressure from donor agencies for nature conservation. Another dominant policy view aims at promoting income in the agricultural sector.

The entire region is considered as ‘one big farm’ *i.e.* the effects of differences in farm size and type are not considered. Only three crops, namely rice, maize and cassava are taken into consideration.

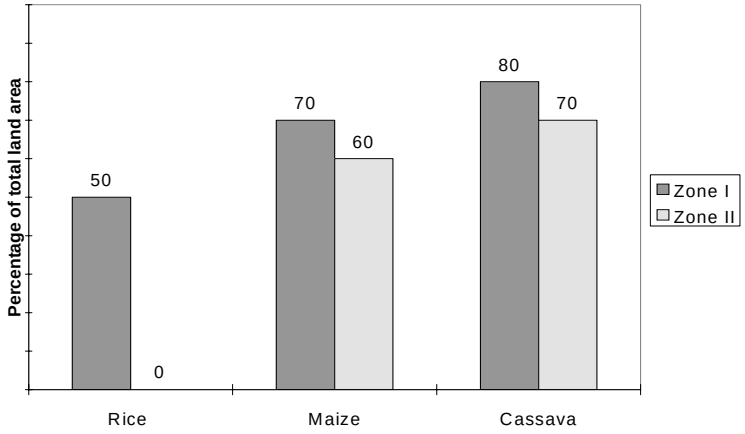
This study starts with a qualitative part to get an idea of the suitability of the region for different crops. In GIS, an administrative boundary map was combined with a climatic map and a soil map. This resulted in a map of *land evaluation units*, each comprising a unique

combination of soil type, climatic conditions and administrative regions, called *zones*. Our region can roughly be divided into 2 zones. The zones differ in soil quality and also in climatic conditions. The total areas calculated in GIS are shown below.

Region 'X'	
Zone	Total Area
I	1,000 ha
II	800 ha

Subsequently, each zone was confronted with the soil and climatic requirements of the three crops under consideration. From the qualitative selection procedure, it becomes apparent that (part of) a zone is either suitable or not for cultivating rice, maize and cassava. Figure 10 shows the percentage of each zone suitable for rice, maize and cassava production.

Figure 10 Percentage suitable for agricultural production.



With the help of a crop growth model, the potential yields of the three crops were calculated given the defined weather and soil conditions (Table 2).

Table 2 Potential yield.

Zone	Rice	Maize	Cassava
I	6,000 kg/ha	4,200 kg/ha	9,000 kg/ha
II	-	3,900 kg/ha	8,500 kg/ha

The expected prices per kilogram of rice, maize and cassava are Rp 1,000, Rp 650 and Rp 250, respectively. Conservation areas can be used to collect firewood and therefore yield an estimated financial return of Rp 500,000 per hectare.

Each land use type has its own output production and input use and timing, depending on the climate and the suitability of the land. The required inputs to realize the outputs were estimated on the basis of a survey and by using expert knowledge and literature. In zone I, rice and maize occupy the land from June until December, and the growing season for cassava starts in March and lasts until October. In zone II, maize occupies the land from March to September while cassava occupies the land for 8 months starting in January. Table 3 shows the labour

requirements, necessary to grow one kilogram of rice, maize or cassava. For simplicity it is assumed that the collection of firewood does not require any labour input.

Table 3 Labour requirements (man-days per hectare).

	Zone I			Zone II	
Month	Rice	Maize	Cassava	Maize	Cassava
Jan					17.0
Feb					2.0
Mar			18.5	10.6	3.0
Apr			2.0	4.3	17.5
May			3.0	6.5	3.5
Jun	22.9	10.8	18.7	44.0	17.3
Jul	122.3	4.5	4.0	2.0	2.5
Aug	35.9	5.0	18.0	25.5	1.0
Sep	8.4	44.2	3.0	26	
Oct	6.4	2.5	1.0		
Nov	43.9	26.8			
Dec	20.4	27.0			

Use is also made of equipment (such as the plough, etc.) and fertilizer. The input requirements are roughly the same for both zones and are given in rupiah per hectare.

Table 4 Use of equipment and fertilizer (rupiah per hectare).

Crop	Equipment	Fertilizer
Rice	7,000	31,000
Maize	4,000	18,000
Cassava	4,500	15,000

According to the agricultural population census of 1995, the labour supply in zone I is 500 and in zone II 750. Labour can freely move between zones. Seasonal labour from outside the region is not taken into consideration. One month has 25 working days. Total working capital available in the region is Rp 38,000,000.

Assignment

With a group of about four people execute the following steps:

i) Model preparation
- operationalize the policy views by means of o*bjection functions* which can be maximized or minimized.
- Identify and quantify the variable(s), constraints, objective coefficients and input-output coefficients.

ii) Model construction
- develop a multiple goal linear programming model in GAMS format.

iii) Model utilization
- Construct the *playing field* by solving the model and then determine the worst and best values for each of the objection functions.
- Discuss which value of the objective is unacceptable.

- Formulate a tighter boundary for that objective and optimize again. (This can be repeated until all group members are satisfied.)
- Generate two land use scenarios by giving more or less weight to each of the objection function.

iv) Presentation
- Make graphs or maps for the different land use scenarios.
- Present the outcomes of your land use analysis study.

Solution to the exercise can be found in Appendix 3.

Bibliography

F. Veinott, Arthur, Jr. 1994. Introduction to Operation Research I. ENGR 62/OR 152.

Brooke, A.; Kendrik, D.; and Meeraus, A. 1992. GAMS, A User's Guide, Release 2.25. The Scientific Press Series, Massachusetts.

Hazell, P.B.R.; and Norton, D. 1986. Mathematical Programming for Economic Analysis in Agriculture. Macmillan, New York.

Ittersum, van M.K.; Ridder, de N.; Rheenen, Bakker, T.; Touré, M.S.M.; and Sissoko, K. 1997. Land Use Analysis using Multiple Goal linear Programming, A Course Manual. Rapports PSS No 31, Wageningen.

McFedries, P. 1994. EXCEL 5 Super Book, Sams Publishing, 1st Edition.

Netherlands Scientific Council for Government Policy. 1992. Ground for Choices, Four Perspectives for the Rural Areas in the European Community. The Hague.

Schweigman, C. 1985. Operations Research Problems in Agriculture in Developing Countries. Khartoum University Press, Khartoum.

Spharim, I.; Spharim, R.; and de Wit, C.T. 1992. Modelling agricultural development strategy. *In* Th. Alberda et al. (eds), Food from Dry Lands, An Integrated Approach to Planning of Agricultural development, Kluwer Academic Publishers, The Netherlands, pp. 159-192.

Taha, H.A. 1992. Operations Research, An Introduction. 5th edition, MacMillan, New York.

Mathematical Programming in GAMS: A Course Manual for Agricultural Planning

Siemon Hollema[*]

Introduction

Modeling has always taken an important place in agricultural economics. It contributes to a better understanding of agricultural production systems, and it can be useful in indicating alternative responses to different conditions and in showing possible future developments. Modeling can take place on different levels. Farm level modeling, for example, can assist farmers in efficiently adapting to changing economic and technological conditions, while sector models can be used to show the response of a sector to different policies, technologies, etc. Also economy-wide models can be useful for national planning.

Mathematical programming models provide a natural framework to analyze quantitative information about the agricultural production system. Agriculturists are used to thinking in terms of input/output coefficients and constraints, while economists are concerned with farm optimization strategies given resource and market conditions.

The General Algebraic Modeling System (GAMS), developed by Brooke, Kendrick and Meerhaus (1992) is designed to facilitate formulation of mathematical programming models. Such models optimize a specified objective subject to various production constraints. GAMS models problems in a highly compact and natural way. The user can change the formulation quickly and easily. Using GAMS, data are entered only once. Models are described in concise algebraic statements which are easy to understand. GAMS is in a way self-explanatory as it pinpoints errors made and their location. GAMS is a DOS based system. It can also run inside a DOS window under Windows. GAMS requires at least 2 Mb of RAM and 200 KB of real mode memory, but at least 4 Mb is recommend to solve larger models. A 386 math co-processor is required. GAMS is a file-oriented system; fortunately however, no special editor is required. Each user can use his favorite editor or word processor.

It takes some time to fully understand GAMS possibilities. The objective of this course manual is to guide the user through the main features of GAMS with the help of a farm-household example. This manual draws heavily on the GAMS USER'S GUIDE (1992) and a course manual written in French by Deybe (1995). To keep the manual as compact as possible, only the main features of GAMS have been selected. Some topics are discussed extensively while others are only briefly touched upon. Knowledge of basic mathematical programming is required to easily follow this manual. For those who want to refresh their knowledge of mathematical programming, Schweigman (1985) and Hazell and Norton (1986) are recommended.

[*] UN/ESCAP CGPRT Centre, Bogor, Indonesia.

The GAMS language

Definition of the problem

Take the problem of a farmer who has to decide on the division of his land between four principal crops, *rice*, *maize*, *cassava* and *soybean* and the amount of labour used. The objective of the farmer is to obtain the highest possible income. He owns four hectares of land, and his family can supply labour equivalent to 25 days per month. It is possible to hire labour for $4 per day. Family members can also do off-farm work for $3 per day. The cost and returns of the four crops are shown in Table 1.

Table 1 The costs and returns for four crops.

Crop	Yield (tons/hectare)	Costs ($/hectare)	Price ($/ton)
Rice	1.5	80	350
Maize	2	5	70
Cassava	3	50	125
Soybean	1	5	200

Rice occupies the land from March until November, maize from May until October, while the growing season for cassava and soybean starts in November and lasts until April. The estimated number of man-days per hectare for each crop is indicated in Table 2.

Table 2 Labour required by month.

Crop	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Rice			5	5	9	2	1.5	2	1	26	12	
Maize					4.3	5.04	7.16	7.97	4.41	1.12		
Cassava	5.16	5	19.6	2.42							11.2	4.68
Soybean	0.75	0.75	0.75	16							7.5	0.75

The mathematical formulation of this problem has the following format:

- indices:

  j = crops
  t = months

- given data:

  W = wage rate for hired labour
  Woff = wage rate for off-farm labour
  Y_j = Yield per hectare of crop j
  P_j = Price of crop j per ton
  C_j = Costs per hectare for crop j
  $l_{j,t}$ = Labour input in mandays per month
  $a_{j,t}$ = Land requirement in hectare per month

- decision variables:

  X_j = number of hectares planted with crop j,
  where, $X_j \geq 0$, for all j
  $Loff_t$ = the amount of family labour working off-farm in every month
  $Lhired_t$ = the amount of hired labour in every month

- constraints:

  Land constraint: $\sum_j X_j \cdot a_{j,t} \leq 4$ for all t

  Labour constraint: $\sum_j X_j \cdot l_{j,t} + Loff_t - Lhired_t \leq 25$ for all t

- objective function:

  Maximize $\sum_j X_j \left(Y_j P_j - C_j\right) + \sum_t Loff_t \cdot Woff - \sum_t Lhired_t \cdot W$

This so-called optimization model gives the optimal allocation of land and labour among alternatives so that the farmer's income is maximized. In mathematical programming the following terminology is used. The cultivation of crops are termed *activities*, the resources available correspond to *constraints*, the return on activities form the *objective coefficients*, and the relations between input and output are called the *technical coefficients* or *input/output coefficients*.

To define this problem in GAMS, certain keywords which help the program to recognize the different components of the mathematical model are used. These keywords are:

SET
PARAMETER
TABLE
SCALAR
VARIABLE
EQUATION
MODEL
SOLVE
DISPLAY (optional)

Now the above problem will be defined in GAMS language using these keywords.

Sets

Sets are the building blocks of a GAMS model. They correspond to the indices, indicated in the algebraic representation of the model. Indices identify the different activities, time periods, resources, etc. (for example, the four cropping activities mentioned, or the months of the year). Other examples include the seasons of the year, types of land, types of farms, kinds of livestock, etc. The keyword used for indices is **SET** or **SETS**. GAMS does not distinguish between plural or singular forms.

It is necessary to define the indices before use. In our example four crops are distinguished. In order not to constantly repeat the names of the crops, they are combined in one single set called J which refers to all four crops. To do so, the following SET statement is written:

```
SET
  J   Crops   /Rice, Maize, Cassava, Soybean/
```

In general the syntax of a SET statement is as follows:

```
SETS
  (Identifier 1)   Text    /element 1, element 1, …, element 1/
  (Identifier 2)   Text    /element 2, element 2, …, element 2/
      .
      .
      .
```

```
(Identifier N)    Text    /element N, element N, …, element N/;
```

The declaration of the indices starts with the word **SET** or **SETS**. The name of the identifier, followed by a space (TAB cannot be used as GAMS does not recognize this) and text (optional). The identifier can have a maximum of 10 characters, and a space between the characters is not allowed. Neither are the following symbols ?&#-+=@%()_’’’!,. The first character has to be a letter. The elements are separated by a comma or press enter to go to the following line. The names of the elements can have a maximum of 10 characters, and a + or - symbol is allowed. This can be handy if there is reason to use two names for the same element. For example, Basmati rice can then be written as Bas-Rice. It is also possible to add explanatory text to every element. The logic is the same as with text used for the identifier: /element1 text, element1 text, …, element1 text/. The element and added text are separated by a space. To terminate the SET statement (and all other statements) place a **semicolon (;)** at the end.

To account for seasonal use of resources, the months of the year have been included in the model. The sets in the GAMS model will thus look as follows:

```
SETS
   J   crops                  /Rice, Maize, Cassava, Soybean/
   T   months of the year     /Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec/;
```

If the elements of a set consist of a sequence, type in the first and the last element combined with an asterisk. For example, in a simulation model there might be 10 annual periods from 1991 to 2000. Instead of typing in all the ten years, the elements of this set can be written as:

```
SETS
   Y   Year   /1991* 2000/;
```

which means that the set includes the ten elements 1991, 1992, …, 2000. It is also possible to define the years as **/Y1 * Y10/** which include the elements Y1, Y2, …, Y10.

It is sometimes necessary to have more than one name for the same set. The set Y, for example, can be given another name by using the **ALIAS** statement.

```
ALIAS (y, ye) ;
```

Now that the sets have been declared and defined objective coefficients and the technical coefficients of the model will be introduced.

Data entry: parameter, scalar and table

Data can be entered in three different ways:

i. For individual numbers, use the keyword **SCALARS**
ii. For vectors, use the keyword **PARAMETERS**
iii. For matrices, use the keyword **TABLES**

The general structure is as follows:

```
KEYWORD     NAME(DOMAIN)        TEXT    /DATA/;
```

The structure is similar to the SET declaration. The statement for data entry is opened with one of the three keywords depending on the kind of data to be entered. First the statement is declared by giving the parameter a name to define its domain. Parameters, in general, have

only one dimension while tables have more. Scalars have a dimension of zero and consequently do not have a domain. Optionally, text can be added. Next the data are entered. For example, in defining the return on agricultural activity, three parameters which are named yield, price and cost can be distinguished. They have a different value for each respective crop. To define these parameters we have several possibilities:

```
1.  PARAMETER YIELD(J)  Crop yield in tons per hectare          /Rice 1.5, Maize 2, Cassava 3,
    Soybean 1/;
2.  PARAMETER PRICE(J)  Crop price in Dollars per ton
                                                               /Rice    350
                                                               Maize    70
                                                               Cassava  125
                                                               Soybean  200/;
3.  PARAMETER COST(J)   Cash costs in Dollars per hectare      /Rice = 80, Maize = 5, Cassava = 50,
    Soybean = 5/;
```

Parameter initialization requires a list of data elements, each consisting of a label and an assigned value. Slashes must be used at the beginning and end of the list. Either use a comma or a return to separate data elements. An equals sign or a blank may be used to separate a label from its associated value.

GAMS always verifies if the labels and domain correspond to the declaration in the SET statement. If they are not the same, GAMS will give an error message.

For parameters with dimensionality zero, the statement **SCALAR** is used. This means that there are no associated sets. Defining a scalar is similar to defining a set. In the example four scalars distinguished, which can be defined in the following way:

```
SCALAR
     LAND      Farm size (hectares)                          /4./
     FAMLAB    Total family labour supply (days per month)   /25/
     Woff      Wage rate for off-farm labour (dollars per day) /3./
     W         Wage rate for hired labour (dollars per day)  /4./;
```

Except for the one dimension lists where the PARAMETER statement is required, most data are entered in GAMS using the TABLE statement. Here it is used to include our technical coefficients $l_{j,t}$ and $a_{j,t}$. The rules for forming simple tables are straightforward. Taking for example the labour requirement, the following table can be created:

```
TABLE     LABREQ(T,J)  Labour requirement in man-days per hectare

          Rice      Maize     Cassava   Soybean
Jan                           5.16      .75
Feb                           5.        .75
Mar       5.                  19.6      .75
Apr       5.                  2.42      16.
May       9.        4.3
Jun       2.        5.04
Jul       1.5       7.16
Aug       2.        7.97
Sep       1.        4.41
Oct       26.       1.12
Nov       12.                 11.2      7.5
Dec                           4.68        .75   ;
```

The structure of this table declaration is similar to the previous declaration, starting with a keyword, identifier name with the domain, followed by optional text. The table represents a matrix with two dimensions: the months of the year (T) and the different crops (J). Notice that the first domain corresponds to the rows of the table, while the second domain corresponds to columns. The values corresponding to each combination of the labels can be typed in a straightforward way. Blank entries imply that the default value zero will be associated with that label combination. Labels can be left out if the entries are zero or not needed. The list of values is not between slashes "/", although the semicolon has to be used to terminate the table statement. GAMS gives an error message when the labels do not correspond with the initial set elements. If it is uncertain in what column a certain value goes, GAMS will protest with an error message.

If a table has too many columns to fit nicely on a single line, then the columns which do not fit can be entered below. For example the table above could be written as:

```
TABLE     LABREQ(T,J)  Labour requirement in man-days per hectare

          Rice         Maize
Jan
Feb
Mar       5.
.
.
Oct       26.          1.12
Nov       12.
Dec

+         Cassava      Soybean
Jan       5.16         .75
Feb       5.           .75
Mar       19.6         .75
.
.
Oct
Nov       11.2         7.5
Dec       4.68           .75  ;
```

The crucial item is the plus sign "+" above the row labels and to the left of the column labels. The row labels (months) have been duplicated. Tables can be continued as many times as necessary.

One more table has to be made to finish the data entry, namely the land requirements. Recall from the example that rice occupies the land from March until November, maize from May to October and cassava and soybean from November until April. This can be included using the following table:

```
TABLE     LANDREQ(T,J) Months of land occupation by crop (hectares)

          Rice         Maize        Cassava      Soybean
Jan                                 1.           1.
Feb                                 1.           1.
Mar       1.                        1.           1.
Apr       1.                        1.           1.
May       1.           1.
Jun       1.           1.
Jul       1.           1.
Aug       1.           1.
```

```
Sep     1.      1.
Oct     1.      1.
Nov     1.              1.      1.
Dec                     1.      1.  ;
```

Now all the data of the model have been defined. In the next section, **VARIABLES** will be introduced. Variables are those identities whose values are unknown until after the model has been solved.

Variables

Variables, or in other words, the level of each activity chosen, are determined endogenously. In an optimization model, a combination of alternative activities, which gives an optimal solution, is determined. GAMS therefore needs a definition of variables which represent the level of every activity. The format is the same as with SETS but in this instance the keyword **VARIABLES** is used. Each variable is given a name, a domain if appropriate and (optionally) text.

```
VARIABLES       NAME1(DOMAIN)     TEXT
                NAME2(DOMAIN)     TEXT
                        .
                        .
                NAMEn(DOMAIN)     TEXT
```

In the farm model, the following VARIABLES can be declared.

```
VARIABLES
        Y           Farm-household income (the objective)
        X(J)        Number of hectares cultivated with crop J
        LHIRED(T)   Temporary labour hired
        LOFF(T)     Family labour workin off-farm    ;
```

One variable, Y (the farm household income) is declared without a domain because it is a scalar quantity. Every GAMS optimization model must contain at least one such variable to serve as the quantity to be minimized or maximized.

Once declared, variables must be given restrictions. It is not, for example, possible to have negative areas. The options are:

POSITIVE	The variable must have a positive value
NEGATIVE	The variable must have a negative value
INTEGER	The variable must have an integer value, 1,2,..,100
BINARY	The variable can only be 0 or 1
FREE	No constraint (the default type)

The variable to be optimized must be of the **FREE** type. In the farm model Y is kept free by default but X(J), LOFF(T) and LHIRED(T) can not be negative. They are constrained to nonnegativity by the following statement:

```
POSITIVE VARIABLE X, LOFF, LHIRED;
```

Note that the domain of the variables is not repeated. As before, a semicolon is required to terminate the VARIABLE statement.

Boundaries on variables can be changed by using the suffixes **.LO**, **.UP**, **.FX**. By using the suffix **.LO** a lower boundary is placed on the variable. The suffix **.Up** places an upper boundary on the variable and **.FX** is a combination of the lower and upper boundary being equal to a fixed value. For example, an upper limit can be placed on the family labour working off-farm as:

```
LOFF.UP(T) = FAMLAB;
```

Now the algebraic relationships within the model can be defined. This will be done in the next section.

Equations: defining the algebraic relationships

Equations is the keyword used in GAMS for the algebraic relationships that will be used to generate the constraints in the model. In GAMS it is necessary to define equations in two steps. Firstly, the equations are declared and secondly the equations are defined. The format is as follows:

```
EQUATIONS
        NAME1(DOMAIN)           TEXT
        NAME2(DOMAIN)           TEXT
                .
                .
        NAMEn(DOMAIN)           TEXT   ;

        NAME1(DOMAIN)..         EQUATION DEFINITION     ;
        NAME2(DOMAIN)..         EQUATION DEFINITION     ;
                .
                .
        NAMEn(DOMAIN)..         EQUATION DEFINITION     ;
```

The format of the declaration is the same as before. First comes the keyword, followed by the name, domain and (optionally) text. At the end of the section, a semicolon is placed. The farm model contains the following equations declarations.

```
EQUATIONS
        LANDBAL(T)      Land balance (hectares)
        LABOURBAL(T)    Labour balance (days)
        INCOME          Total farm-household income (Dollars) ;
```

Equations can refer to one or several relationships. For example INCOME has no domain so it is a single equation. But LABOURBAL refers to a set of inequalities defined over the domain T, i.e. for each month.

Equation definitions are the most complex statements in GAMS. The components of an equation definition are:

i. The name of the equation being defined.
ii. The domain.
iii. The symbol ".."
iv. The left-hand-side expression.
v. The relational operator: **=L=** less than or equal to
   =**E**= equal to
   =**G**= greater than or equal to
vi. The right-hand-side expression.

It is always necessary to end the equation with a semicolon (;).

The equations in the example are the objective function, and the equations for the land and labour constraints. Before they are defined GAMS's notations for summation and product must be described. In mathematics the Greek letter Σ is used as the symbol for summation. In GAMS the word SUM is used. The format is based on the idea that a summation (product) is an operator with two arguments:

SUM (index of summation, summand)

For example, the land constraints $\sum_j X_j \cdot a_{j,t} \leq 4$, for each month t, is expressed in GAMS language as follows:

```
SUM(J, X(J)*LANDREQ(T,J)) =L= LAND
```

This equation expresses that for each month the summation of the area grown with crop J times the land requirements must be lower than or equal to the total land available.

Products are defined in GAMS using exactly the same format as summation, replacing SUM with PROD. For example,

PROD(J, X(I,J))

is equivalent to $\prod_j X_{ij}$.

Now the three equations of the farm model can be defined.

```
LANDBAL(T)..
        SUM(J, X(J)*LANDREQ(T,J)) =L= LAND ;
LABOURBAL(T)..
        SUM(J, X(J)*LABREQ(T,J)) + LOFF(T) =L= FAMLAB  + LHIRED(T) ;
INCOME..
        SUM(J, X(J)*(Y(J)*P(J)-C(J))) + SUM(T, LOFF(T)*WOFF) - SUM(T, LHIRED(T)*W) =E= Y ;
```

The land and labour balance equations are summated over J, i.e. the cultivation of different crops and not over T, i.e. the time periods. The equations refer thus to a series of relationships, one for every month.

In the last equation the operator =E= was used. It is important to understand the difference between the symbols "=" and "=E=". The symbol "=" is only used in direct assignments to alter or give a desired value to a parameter. This will be discussed later. Notice that variables can be placed on the left-hand-side as well as on the right-hand-side.

The farm model is now finished and ready to be solved. For this, the **MODEL** and **SOLVE** statements are used.

The model and solve statements

The model statement is used to identify the model i.e. to give it a name and to group and label the equations so that they can be solved. The format is as follows:

MODEL MODEL-NAMETEXT / NAMES OF EQUATIONS/ ;

In the farm model the syntax is:

```
MODEL       FARM  Farm household model          /LANDBAL,  LABOURBAL, INCOME/;
```

The equations to be included in the model can be listed. In this way sub-models can be created. To include all equations previously defined, simply use the keyword **/ALL/**.

```
MODEL       FARM  Farm household model    /ALL/    ;
```

After the model has been declared and the equations assigned, the model is ready to be solved. For this use the **SOLVE** statement. The format of the SOLVE statement is as follows:

i. The keyword "SOLVE"
ii. The name of the model to be solved
iii. The keyword "USING"
iv. The type of solution procedure. GAMS allows many options, for example:
"LP" for Linear Programming
"NLP" for Non-Linear Programming
"MIP" for Mix Integer Programming
v. The type of solution, using the keyword "MINIMIZING" or "MAXIMIZING"
vi. The name of the variable to be optimized.

For the example the statement is:

```
SOLVE FARM USING LP MAXIMIZING Y;
```

The execution of the solve statement will cause several things to happen. Among others the solver's output will be printed to a file. The optimal values can be seen in the solver's output, or a display of these results can be requested from GAMS. For the latter, use the **DISPLAY** statement.

The display statement

To see the optimal number of hectares cultivated with crop J, use the following DISPLAY statement.

```
DISPLAY  X.L ;
```

After the keyword DISPLAY, type the variable followed by the suffix "**.L**". GAMS will automatically format a printout in two dimensional tables with appropriate headings.

Of course, the optimal level of hired labour and how much family labour should be allocated to off-farm work are also of interest. The DISLAY statement of the our farm model looks then as follows:

```
DISPLAY  X.L, LHIRED.L, LOFF.L ;
```

Notice that the domain of the variable is not repeated. The formulation of the optimization model has now been completed above. The complete model should look as follows:

```
SETS
        J           crops                   /Rice, Maize, Cassava, Soybean/
        T           months of the year      /Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec/ ;
 SCALARS
        LAND        Farm size (hectares)                                    /4./
        FAMLAB      Total family labour supply (days per month)             /25/
        Woff        Wage rate for off-farm labour (dollars per day)         /3./
        W           Wage rate for hired labour (dollars per day)            /4./ ;
```

```
PARAMETERS
        YIELD(J)        Crop yield in tons per hectare          /Rice 1.5, Maize 2, Cassava 3, Soybean 1/
        PRICE(J)        Crop price in dollars per ton           /Rice 350, Maize 70, Cassava 125, Soybean
200/
        COST(J)         Cash costs in dollars per hectare       /Rice 80, Maize 5, Cassava 50, Soybean 5/ ;

TABLE
            LANDREQ(T,J)        Months of land occupation by crop (hectares)

          Rice          Maize         Cassava       Soybean
  Jan                                 1.            1.
  Feb                                 1.            1.
  Mar     1.                          1.            1.
  Apr     1.                          1.            1.
  May     1.            1.
  Jun     1.            1.
  Jul     1.            1.
  Aug     1.            1.
  Sep     1.            1.
  Oct     1.            1.
  Nov     1.                          1.            1.
  Dec                                 1.            1.   ;

TABLE
            LABREQ(T,J)         Labour requirement in man-days per hectare

          Rice          Maize         Cassava       Soybean
  Jan                                 5.16          .75
  Feb                                 5.            .75
  Mar     5.                          19.6          .75
  Apr     5.                          2.42          16.
  May     9.            4.3
  Jun     2.            5.04
  Jul     1.5           7.16
  Aug     2.            7.97
  Sep     1.            4.41
  Oct     26.           1.12
  Nov     12.                         11.2          7.5
  Dec                                 4.68           .75    ;

VARIABLES
        Y               Farm-household income (the objective)
        X(J)            Number of hectares cultivated with crop J
        LHIRED(T)       Temporary labour hired
        LOFF(T)         Family labour working off-farm   ;

POSITIVE VARIABLE X, LHIRED, LOFF ;

EQUATIONS
          LANDBAL(T)    Land balance (hectares)
          LABOURBAL(T)      Labour balance (days)
          INCOME            Total farm-household income (dollars) ;

LANDBAL(T)..
        SUM(J, X(J)*LANDREQ(T,J)) =L= LAND       ;

LABOURBAL(T)..
        SUM(J, X(J)*LABREQ(T,J)) + LOFF(T) =L= FAMLAB + LHIRED(T)  ;
```

```
INCOME..
        SUM(J, X(J)*(YIELD(J)*PRICE(J)-COST(J))) + SUM(T, LOFF(T)*WOFF) - SUM(T,
        LHIRED(T)*W) =E= Y ;

MODEL       FARM  Farm household model   /ALL/   ;

SOLVE  FARM  USING  LP  MAXIMIZING  Y;

DISPLAY  X.L, LHIRED.L, LOFF.L ;
```

GAMS system initialization

The simplest way to start GAMS is to enter the command

```
C:\GAMS> GAMS "name of the file"
```

from the system prompt. For example, if the file above had been saved in the A:\ directory under the name **expl1.gms**, the command would be:

```
C:\GAMS> gams a:\expl1
```

If there are no errors, GAMS will compile and execute the model with the file name expl1.gms. A new file will be created (regardless of whether there are errors or not) with the extension **.LST**. If mistakes have been made, an error message will appear on the computer screen. In the output file **expl1.lst** the errors are marked with four asterisks **** which indicate the place of the error. At the end of the reprinted original model is a section labeled ERROR MESSAGES. Here a brief explanation is given of the probable cause of the error.

If an extension other than .gms has been used, type in the complete name. GAMS still creates an output file with the extension .LST.

Another way to start GAMS is by using the "option" call. The most commonly-used options are **PW** and **PS**, which are used to alter the page width and the page length respectively. For example,

```
C:\GAMS>gams expl1 PW=40 PS=40
```

will alter the default value of the page width (usually 79) to 40 and the length of the page to 40 lines instead of the default of 60.

After GAMS is finished saving the output file, the message

--- All done
--- Erasing scratch files

will appear on the screen. Using an editor, the output file **expl1.lst** can now be opened.

GAMS output

The GAMS output file consists of two parts: the compilation part and the solution part. The compilation part is the output produced during the initial check of the program. It contains the **Echo print** and the **Reference Maps**. The solution part lists all the output produced by a solve statement. It consists of the following parts: **Equation listing**, **Column listing**, **Model statistics**, and **Solution report**. Data from the DISPLAY statement are shown in the **Execution Output**.

Echo print

The Echo print is a copy of the optimization problem. The only difference being that, for future reference, GAMS puts line numbers on the left-hand side of the echo. If errors are made, they are marked with **** in the left-hand column. A "$" followed by a numerical error code can be found directly below the point at which the compiler thinks the error is made. The codes are explained after the echo print. The echo print for the farm model is as follows:

```
General Algebraic Modeling System
Compilation

1       SETS
2               J       crops                   /Rice, Maize, Cassava, Soybean /
3               T       months of the year      /Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec /;
4
5       SCALARS
6               LAND                    Farm size (hectares)                            /4./
7               FAMLAB                  Total family labour supply (days per month)     /25/
.
.
.
12      PARAMETERS
13              YIELD(J)  Crop yield in tons per hectare  /Rice 1.5, Maize  2, Cassava 3, Soybean 1/
.
.
.
74
75      MODEL    FARM         Farm household model    /ALL/  ;
76
77      SOLVE  FARM  USING  LP  MAXIMIZING  Y ;
78
79      DISPLAY X.L, LHIRED.L, LOFF.L ;
```

If for example, "maize" in the parameter statement were spelled differently from the way it was introduced in the set declaration, the following error message would appear.

```
.
.
12      PARAMETERS
13              YIELD(J)        Crop yield in tons per hectare      /Rice 1.5, Mazie  2, Cassava 3, Soybean 1/
****                                                                  $170
.
.
Error Messages
170     Domain violation for element.
```

Reference maps

Reference maps contain summaries and analyses of the input file for the purpose of debugging and documentation. It consists of a **cross-reference map** and a **list of model entities** (sets, parameters, variables, and equations) including the associated documentary text. A cross-reference map is an alphabetical, cross-referenced list of all entities of the model. The cross-reference map for the farm model example is as follows.

General Algebraic Modeling System
Symbol Listing

SYMBOL	TYPE	REFERENCES					
COST	PARAM	DECLARED	17	DEFINED	18	REF	73
FAMLAB	PARAM	DECLARED	8	DEFINED	8	REF	70
FARM	MODEL	DECLARED	75	DEFINED	75	IMPL-ASN	77
		REF	77				
INCOME	EQU	DECLARED	66	DEFINED	72	IMPL-ASN	77
		REF	75				
J	SET	DECLARED	2	DEFINED	2	REF	13
		15	17	22	39	57	2*68
		2*70	3*7	2*73	CONTROL	68	70
			2				
		72	73				
LABOURBAL	EQU	DECLARED	65	DEFINED	70	IMPL-ASN	77
		REF	75				
LABREQ	PARAM	DECLARED	39	DEFINED	39	REF	70
LAND	PARAM	DECLARED	7	DEFINED	7	REF	68
LANDBAL	EQU	DECLARED	64	DEFINED	68	IMPL-ASN	77
		REF	75				
LANDREQ	PARAM	DECLARED	22	DEFINED	22	REF	68
LHIRED	VAR	DECLARED	58	IMPL-ASN	77	REF	61
		70	73	79			
LOFF	VAR	DECLARED	59	IMPL-ASN	77	REF	61
		70	72	79			
PRICE	PARAM	DECLARED	15	DEFINED	16	REF	72
T	SET	DECLARED	3	DEFINED	3	REF	22
		39	58	59	64	65	68
		3*70	72	73	CONTROL	68	70
		72	73				
W	PARAM	DECLARED	10	DEFINED	10	REF	73
WOFF	PARAM	DECLARED	9	DEFINED	9	REF	72
X	VAR	DECLARED	57	IMPL-ASN	77	REF	61
		68	70	72	73	79	
Y	VAR	DECLARED	56	IMPL-ASN	77	REF	73
		77					
YIELD	PARAM	DECLARED	13	DEFINED	14	REF	72

It indicates, for example, that the symbol COST is a parameter that was declared in line 17, defined, i.e. assigned a value, in line 18, and referenced in line 73 where it is used to calculate the costs of cultivation. The symbol J has a more complicated entry. It is a set which has been declared and defined in line 2. It is referenced once in line 13, 15, 17, 22, 39, 57, twice in line 68 and 70, and three times in line 73. It also services as a controlling index in a summation in lines 68, 70, 72, and 73. The variables Y, X, LOFF and LHIRED are implicitly assigned (**IMPL-ASN**) in line 77. They will be determined as a result of a solve statement.

The second part of the reference map is a list of model entities, grouped by type and listed with the associated documentary text. This list is as follows:

SETS

J	**Crops**
T	**Months of the year**

PARAMETERS

COST	**Cash costs in dollars per hectare**

```
FAMLAB      Total family labour supply (days per month)
LABREQ      Labour requirement in man-days per hectare
LAND        Farm size (hectares)
LANDREQ     Months of land occupation by crop (hectares)
PRICE       Crop prices in dollars per ton
W           Wage rate for hired labour (dollars per day)
WOFF        Wage rate for off-farm labour (dollars per day)
YIELD       Crop yield in tons per hectare

VARIABLES

LHIRED      Temporary labour hired
LOFF        Family labour working off-farm
X           Number of hectares cultivated with crop J
Y           Farm-household income (the objective)

EQUATIONS

INCOME      Total farm-household income (dollars)
LABOURBAL   Labour balance (days)
LANDBAL     Land balance (hectares)

MODELS

FARM        Farm household model

COMPILATION TIME  =        0.160 SECONDS                VERID  MW2-25-087
```

Equation listings

The equation listing shows the general algebraic form of the model. It is extremely useful to verify if the model actually does what it is intended to do. For example the labour constraint as given in the input file is:

```
LABOURBAL(T)..
        SUM(J,X(J)*LABREQ(T,J)) + LOFF(T) =L= FAMLAB + LHIRED(T);
```

while the equation listing of this specific constraint is:

```
---- LABOURBAL       =L=      Labour balance (days)

LABOURBAL(JAN)..             5.16*X(CASSAVA)  + 0.75*X(SOYBEAN) + LOFF(JAN) - LHIRED(JAN)
                         =L=  25  ;  (LHS = 0)

LABOURBAL(FEB)..             5*X(CASSAVA) + 0.75*X(SOYBEAN) + LOFF(FEB) - LHIRED(FEB)
                         =L=  25  ;  (LHS = 0)

LABOURBAL(MAR)..             5*X(RICE) + 19.6*X(CASSAVA) + 0.75*X(SOYBEAN) + LOFF(MAR) -
                     LHIRED(MAR)      =L=  25  ;  (LHS = 0)

REMAINING 9  ENTRIES SKIPPED
```

The name, text and type of constraint are shown. The four dashes can be useful for mechanical searching. All the terms depending on the VARIABLES are collected on the left, while all the constant terms are combined into one number on the right. Terms with a zero coefficient are not shown. By default only the first three equations for each generic equation are listed. The default can be changed by inserting the command **OPTION LIMROW = n ;** before the SOLVE statement. To see the labour constraint for every month, n would be 12.

Column listings

The column listing shows the coefficients of each variable, and their boundaries and level values. The variable Y, for example, is presented in the column listing as follows:

```
---- Y           Farm-household income (the objective)

Y
                 (.LO, .L, .UP = -INF, 0, +INF)
        -1       INCOME
```

It is an independent endogenous variable used to express the objective function. It has a lower limit (**.LO**) of minus infinity (**-INF**), an upper limit of plus infinity (**+INF**), and, if no initial value is imposed, the current level (**.L**) is zero. In a similar way the variable LHIRED, which is part of the income equation and the labour constraint equation, is presented as:

```
---- LHIRED      Temporary labour hired


LHIRED(JAN)
           (.LO, .L, .UP = 0,  0, +INF)
        -1       LABOURBAL(JAN)
        -4       INCOME

LHIRED(FEB)
                 (.LO, .L, .UP = 0,  0, +INF)
          -1     LABOURBAL(FEB)
          -4     INCOME

LHIRED(MAR)
                 (.LO, .L, .UP = 0,  0, +INF)
          -1     LABOURBAL(MAR)
          -4     INCOME
REMAINING 9 ENTRIES SKIPPED
```

The variable was restricted to a positive value. The lower limit therefore is equal to zero (**.LO = 0**). Once again the default is the first three entries for each variable. To change the default, use the command **OPTION LIMCOL = n ;** placed before the solve statement. To repress both the equation and column listing, the following statement can be given:

```
OPTION LIMROW = 0, LIMCOL = 0 ;
```

Model statistics

The next section gives the model statistics shown below. It quickly gives an overview of how large the model is. In our example there are 25 equations: one objective function and 2 times 12 constraints (**SINGLE EQUATIONS**). In the way we defined them, we get 3 blocks: one for the objective and 2 for the constraints (**BLOCKS OF EQUATIONS**). The variables are formulated in four ways (**BLOCK OF VARIABLES**) with 29 individual columns (**SINGLE VARIABLES**). **NON ZERO ELEMENTS** shows the number of coefficients which are not equal to zero.

```
MODEL STATISTICS

BLOCKS OF EQUATIONS     3     SINGLE EQUATIONS        25
BLOCKS OF VARIABLES     4     SINGLE VARIABLES        29
```

```
NON ZERO ELEMENTS       107

GENERATION TIME     =       0.160 SECONDS

EXECUTION TIME      =       0.660 SECONDS      VERID MW2-25-087
```

Solution report

Finally we arrive at the solution of the problem. It consists of three parts:

i. The solve summary.
ii. The results for the equations and the variables.
iii. The report summary.

```
                    S O L V E    S U M M A R Y

MODEL               FARM                OBJECTIVE    Y
TYPE                LP                  DIRECTION    MAXIMIZE
SOLVER              BDMLP               FROM LINE    77

**** SOLVER STATUS                 1 NORMAL COMPLETION
**** MODEL STATUS                  1 OPTIMAL
**** OBJECTIVE VALUE               1805.5822

RESOURCE USAGE, LIMIT              0.379         1000.000
ITERATION COUNT, LIMIT             17            1000

GAMS/BDMLP 1.1      Feb 10, 1995   003.048.026-032.000    386/486 DOS-W

A. Brooke, A. Drud, and A. Meeraus,
Analytic Support Unit,
Development Research Department,
World Bank,
Washington, D.C. 20433, U.S.A.

Work space allocated               --       0.05 Mb

EXIT -- OPTIMAL SOLUTION FOUND.
```

The solve summary is shown above. The model solved is FARM. The solve system used is BDMLP. The objective is Y, which is maximized. The **SOLVER STATUS** characterizes the outcome for the solver. Some possible outcomes are:

1 **NORMAL COMPLETION**.
The solver has not been interrupted by limits or internal difficulties.

2 **ITERATION INTERRUPT**.
GAMS has a limit of 1000 iterations. This message will appear if more iterations are necessary. Use **OPTION ITERLIM = n ;** to increase the iteration limit.

3 **RESOURCE INTERRUPT**.
GAMS has a time limit of 1000 units. If too much time is needed this message will appear. Use **OPTION RELIM = n ;** to increase the time limit.

The **MODEL STATUS** indicates what the solution looks like. Some possible messages are:

1 **OPTIMAL** The solution is optimal

3 **UNBOUNDED** The solution is unbounded
4 **INFEASIBLE** The problem is infeasible.

The value of the objective function is 1805.5822. To attain this solution, GAMS needed 17 iterations and only 0.379 time units. After the message **EXIT – OPTIMAL SOLUTION FOUND**, the second part of the solution, the results, is presented.

---- EQU LANDBAL Land balance (hectares)

	LOWER	LEVEL	UPPER	MARGINAL
JAN	-INF	0.342	4.000	.
FEB	-INF	0.342	4.000	.
MAR	-INF	4.000	4.000	167.338
APR	-INF	4.000	4.000	.
MAY	-INF	4.000	4.000	.
JUN	-INF	4.000	4.000	.
JUL	-INF	4.000	4.000	.
AUG	-INF	4.000	4.000	.
SEP	-INF	4.000	4.000	.
OCT	-INF	4.000	4.000	39.580
NOV	-INF	4.000	4.000	.
DEC	-INF	0.342	4.000	.

---- EQU LABOURBAL Labour balance (days)

	LOWER	LEVEL	UPPER	MARGINAL
JAN	-INF	25.000	25.000	3.000
FEB	-INF	25.000	25.000	3.000
MAR	-INF	25.000	25.000	3.116
APR	-INF	25.000	25.000	3.000
MAY	-INF	25.000	25.000	4.000
JUN	-INF	25.000	25.000	3.000
JUL	-INF	25.000	25.000	3.000
AUG	-INF	25.000	25.000	3.000
SEP	-INF	25.000	25.000	3.000
OCT	-INF	25.000	25.000	4.000
NOV	-INF	25.000	25.000	4.000
DEC	-INF	25.000	25.000	3.000

	LOWER	LEVEL	UPPER	MARGINAL
---- EQU INCOME	.	.	.	-1.000

INCOME total farm-household income (dollars)

	LOWER	LEVEL	UPPER	MARGINAL
---- VAR Y	-INF	1805.582	+INF	.

Y Farm-household income (the objective)

---- VAR X Number of hectares cultivated with crop J

	LOWER	LEVEL	UPPER	MARGINAL
RICE	**.**	**3.658**	**+INF**	**.**
MAIZE	**.**	**0.342**	**+INF**	**.**
CASSAVA	**.**	**0.342**	**+INF**	**.**
SOYBEAN	**.**	**.**	**+INF**	**-59.425**

---- VAR LHIRED Temporary labour hired

	LOWER	LEVEL	UPPER	MARGINAL
JAN	**.**	**.**	**+INF**	**-1.000**
FEB	**.**	**.**	**+INF**	**-1.000**
MAR	**.**	**.**	**+INF**	**-0.884**
APR	**.**	**.**	**+INF**	**-1.000**
MAY	**.**	**9.390**	**+INF**	**.**
JUN	**.**	**.**	**+INF**	**-1.000**
JUL	**.**	**.**	**+INF**	**-1.000**
AUG	**.**	**.**	**+INF**	**-1.000**
SEP	**.**	**.**	**+INF**	**-1.000**
OCT	**.**	**70.479**	**+INF**	**.**
NOV	**.**	**22.726**	**+INF**	**.**
DEC	**.**	**.**	**+INF**	**-1.000**

---- VAR LOFF Family labour working off-farm

	LOWER	LEVEL	UPPER	MARGINAL
JAN	**.**	**23.233**	**+INF**	**.**
FEB	**.**	**23.288**	**+INF**	**.**
MAR	**.**	**.**	**+INF**	**-0.116**
APR	**.**	**5.884**	**+INF**	**.**
MAY	**.**	**.**	**+INF**	**-1.000**
JUN	**.**	**15.959**	**+INF**	**.**
JUL	**.**	**17.062**	**+INF**	**.**
AUG	**.**	**14.955**	**+INF**	**.**
SEP	**.**	**19.832**	**+INF**	**.**
OCT	**.**	**.**	**+INF**	**-1.000**
NOV	**.**	**.**	**+INF**	**-1.000**
DEC	**.**	**23.397**	**+INF**	**.**

The single dots represent zeroes. Sometimes the entry “EPS” meaning very small but nonzero values is found. The limits (**LOWER/UPPER**) and the level (**LEVEL**) of the equations and variables are presented. The values of the equations indicate that from the month March to November all the available land is in use. In every month complete use of the available labour is made. The column **MARGINAL** gives the increase in value of the objective function if the constraints were relieved with one unit. This value is normally called **shadow prices** or **dual values**. It is the maximum price the farmer is willing to pay for one extra unit of the resource.

Next the values of the variables are presented. The variable Y is a free variable with a lower limit of minus infinity (**-INF**) and an upper limit of plus infinity(**+INF**). The variables X, LHIRED and LOFF are obviously positive. The column **LEVEL** gives the optimum use of the land, and the optimum use of labour. The column **MARGINAL** gives the values with which the objective coefficient has to decrease to insure that the activity takes part in the optimal solution.

For example if the price of soybean increased to US$ 260, it would be optimal for the farmer to also grow soybean. In linear programming this is called **reduced cost**.

The last part of the solution report is the report summary. It gives the total number of nonoptimal, infeasible, and unboundaried rows and columns. The example shows the desired outcome of all 0 tallies.

```
**** REPORT SUMMARY :        0     NONOPT
                             0     INFEASIBLE
                             0     UNBOUNDED
```

Execution output

Output from the **DISPLAY** statement is shown in the execution output. The input statement **DISPLAY X.L, LHIRED.L, LOFF.L ;** results in the following output, found at the end of the output file.

```
General Algebraic Modeling System
Execution

----     79 VARIABLE X.L      Number of hectares cultivated with crop J

RICE                3.658,      MAIZE          0.342,      CASSAVA        0.342

----     79 VARIABLE LHIRED.L  Temporary labour hired

MAY                 9.390,      OCT           70.479,      NOV            22.726

----     79 VARIABLE LOFF.L              Family labour working off-farm

JAN                23.233,      FEB           23.288,      APR             5.884
JUN                15.959,      JUL           17.062,      AUG            14.955
SEP                19.832,      DEC           23.397

EXECUTION TIME       =        0.210 SECONDS      VERID MW2-25-087
```

GAMS automatically formats and labels an appropriate array. Zero entries are not shown. Marginal values can also be displayed by using the suffix **.M**. In displaying parameters, a suffix is naturally not allowed.

Advanced topics

More complicated applications are possible with GAMS. The different applications are presented in the framework of a linear programming problem, but of course they also apply just as well to non-linear programming and dynamic programming. These applications will be introduced by extending the original model. The changes and additions made and will be examined and discussed accordingly. The extended model is as follows:

```
$TITLEFARM LEVEL MODEL - example 2
$ONTEXT
This is an extended version of the original model, taking farm risk  into account.
$OFFTEXT

$STITLE         CROP DATA
```

```
SETS
     J        crops                  /Rice, Maize, Cassava, Soybean/
     T        months of the year     /Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec/
     S        seasons                /Summer, Winter/
     SJ(S,J)  season crop mapping    /Summer . (Rice, Maize)
                                      Winter . (Cassava, Soybean)/
     YE       years                  /1991 * 1995/ ;

SCALARS
     LAND     Farm size (hectares)                                  /4./
     FAMLAB   Total family labour supply (days per month)           /25/
     Woff     Wage rate for off-farm labour (dollars per day)       /3./
     W        Wage rate for hired labour (dollars per day)          /4./
     PRENT    Plough rental cost (dollars per hectare)              /1./
     PHI      Risk factor                                           /1/ ;

PARAMETERS
     YIELD(J)     Crop yield in tons per hectare          /Rice 1.5, Maize 2, Cassava 3, Soybean 1/
     COST(J)      Cash costs in dollars per hectare       /Rice 80, Maize 5, Cassava 50, Soybean  5/
     AVPR(J)      Average price (dollars per ton)
     DEVPR(J,YE)  Price deviation for crops ;

TABLE
     LANDREQ(T,J)  Months of land occupation by crop (hectares)

          Rice      Maize     Cassava   Soybean
Jan                           1.        1.
Feb                           1.        1.
Mar       1.                  1.        1.
Apr       1.                  1.        1.
May       1.        1.
Jun       1.        1.
Jul       1.        1.
Aug       1.        1.
Sep       1.        1.
Oct       1.        1.
Nov       1.                  1.        1.
Dec                           1.        1.   ;

TABLE
     LABREQ(T,J)        Labour requirement in man-days per hectare

          Rice      Maize     Cassava   Soybean
Jan                           5.16      .75
Feb                           5.        .75
Mar       5.                  19.6      .75
Apr       5.                  2.42      16.
May       9.        4.3
Jun       2.        5.04
Jul       1.5       7.16
Aug       2.        7.97
Sep       1.        4.41
Oct       26.       1.12
Nov       12.                 11.2      7.5
Dec                           4.68       .75   ;
```

```
TABLE
        PRICE(J,YE)             Price time series (dollars per ton)


              1991        1992        1993        1994        1995
  Rice        300         320         310         360         380
  Maize       70          80          65          50          75
  Cassava     120         100         140         130         125
  Soybean     230         280         240         210         250   ;

AVPR(J) = SUM(YE, PRICE(J,YE))/CARD(YE) ;
DEVPR(J,YE) = PRICE(J,YE) - AVPR(J) ;

DISPLAY AVPR, DEVPR ;

$STITLE ENDOGENOUS VARIABLES AND EQUATIONS

VARIABLES
        Y                       Farm-household income (the objective)
        X(J)                    Number of hectares cultivated with crop J
        LHIRED(T)               Temporary labour hired
        LOFF(T)                 Family labour working off-farm
        PHIRE(S)                Ox-drawn plough rental (hectares ploughed)
        PDEV(YE)                Positive income deviation (dollars)
        NDEV(YE)                Negative income deviation (dollars) ;

POSITIVE VARIABLE X, LHIRED, LOFF, PHIRE, PDEV, NDEV  ;

EQUATIONS
          LANDBAL(T)            Land balance (hectares)
          LABOURBAL(T)          Labour balance (days)
          PLOUGH(S)             Land ploughed
          DDEV(YE)              Income deviation definition
          INCOME                Total farm-household income (dollars) ;

LANDBAL(T)..                    SUM(J, X(J)*LANDREQ(T,J)) =L= LAND          ;

LABOURBAL(T)..                  SUM(J, X(J)*LABREQ(T,J)) + LOFF(T) - LHIRED(T) =L= FAMLAB  ;

PLOUGH(S)..                     SUM(J $ SJ(S,J), X(J)) =L= PHIRE(S) ;
DDEV(YE)..                      SUM(J, DEVPR(J,YE) * X(J) * YIELD(J)) =E= PDEV(YE) - NDEV(YE)  ;

INCOME..                        SUM(J, X(J)*(YIELD(J)*AVPR(J)) + SUM(T, LOFF(T)*WOFF) -
                                SUM(T, LHIRED(T)*W) - SUM(J, X(J)*COST(J)) - SUM(S,HIRE(S)
                                *PRENT) - PHI * SUM(YE, PDEV(YE) + NDEV(YE))/CARD(YE) =E= Y;
MODEL          FARM  Farm household model    /ALL/    ;

SOLVE  FARM  USING  LP  MAXIMIZING  Y;

PHI = 0 ;

*  Risk on revenue is not considered.

SOLVE  FARM  USING  LP  MAXIMIZING  Y;
```

Several changes and additions have been made. Titles, subtitles, and some commentary text have been added. Furthermore, it is assumed that there are two seasons: summer and winter. In every season the land has to be ploughed once before any crops can be grown. To do this the farmer can rent an ox-drawn plough for US$ 1 per hectare ploughed. Also the prices received

for the crops are not directly given but consist of a time series for prices received for the crops in the years 1991 to 1995. Two models will be considered: one which takes risk on revenue into consideration and one without.

First additional text is introduced, then set operations, followed by data handling and manipulation. Finally, the functions ORD and CARD are examined; the possibility to have several solves within one model, the loop statement, the include option and the put statement are discussed.

Titles, additional text and block comments

(Sub) titles and block comments are added by using a **Dollar Control Directive**. The directive **$TITLE** causes every page of output to have the title you have specified. It can contain a maximum of 80 characters of text. **$STITLE** creates subtitles. Titles and subtitles can be placed anywhere and can also be redefined. The **$ONTEXT-$OFFTEXT** is used to create block comments. These comments are ignored by GAMS but will appear in the output file without line numbers. Comments can also be added using an asterisk ***** in the first column. Text of the latter will have a line number in the output file. For other useful Dollar Control Directives see Appendix 4.

Set operations

A lot of operations with sets are possible. For example creating multi-dimensional sets and dynamic sets, or unions, intersections and complements operations. Here, multi-dimensional sets, dynamic sets and lag and lead operations are considered.

Multi-dimensional sets

In GAMS it is possible to create complex sets, in other words sets that have elements that are pairs or even n-tuples. For example, in the example there are two seasons: summer and winter. In every season the land has to be ploughed once before any crops can be grown. To enable GAMS to take account of this fact, a seasonal mapping (i.e. which crop is grown in which season) should be created. For this reason two additional sets have been included.

```
S       Season                  /Summer, Winter/
SJ(S,J) Season crop mapping     /Summer . (Rice, Maize)
                                 Winter . (Cassava, Soybean)/ ;
```

The dot between the seasons and the crops is used to establish a relationship between elements in a different set (the relation between season and crops grown). The notation **(S,J)** after the set **SJ** indicates that the first member of each pair must be a member of the set S (season), and that the second must be in the set **J** (crop). In the equation **PLOUGH(S)** this set will be used to calculate for how many hectares an ox-drawn plough has to be rented. How this is done will be explained later.

Dynamic sets

Dynamic sets are most commonly used as a “controlling index” in an assignment or in an equation definition or as the control entity in a dollar controlled index operation. For example consider the following:

```
SET     T       months                  /Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec/
        S1(T)   Summer (subset of T)    /Jun, Jul, Aug, Sep, Oct, Nov /
        S2(T)   Winter (subset of T)    ;
```

```
S1("May") = YES ;
S1("Nov")= NO ;
S2(T) = YES;
S2(S1) = NO ;
DISPLAY S1, S2 ;
```

The first statement adds one member (the month May) to the subset summer and the second statement removes one (the month November). In the third statement the subset winter is assigned all the members of set T. In the following statement the summer months are subtracted. The words YES and NO have thus the effect of adding or removing one or more members from a set. The output of the above statements is:

```
----        12  SET  S1   Summer (subset of T)
MAY,            JUN,      JUL,      AUG,      SEP,      OCT

----        12  SET  S2   Winter (subset of T)
JAN,            FEB,      MAR,      APR,      NOV,      DEC
```

Lag and lead operations

With lag and lead operators we can relate the current member of a set with the next or previous member. GAMS provides two forms: the Linear Lag and Lead Operators (+, -) and the Circular Lag and Lead Operators (++, --), the difference being that the first is an "end-off" operator while the Circular operator is used for modeling time periods that repeat, such as the months of a year. A simple example, in which the yield and amount of investment depends on the rainfall in the previous and next, respectively, season, for both operators is as follows:

```
SET                 S     seasons      /summer, winter/;
PARAMETER           R(S)  Rainfall     /summer  1000, winter  50/
                    Y(S)  Yield
                    I(S)  Investment ;
LINEAR:
        Y(S) = R(S-1) *2 ;
        I(S) = R(S+1) / 5 ;
CIRCULAR:
        Y(S) = R(S—1) *2 ;
        I(S) = R(S++1) / 5 ;
```

The output in case of the linear operator is:

```
----      9  PARAMETER      Y      Yield
WINTER          2000.000

----      9  PARAMETER      I      Investment
SUMMER          10.000
```

The output in case of the circular operator is:

```
----     12  PARAMETER      Y      Yield
SUMMER          100.000,        WINTER          2000.000

----     12 PARAMETER       I      Investment
SUMMER          10.000          WINTER          200.000
```

Data handling and manipulation

This section includes deal with aspects of data handling and manipulation. The assignment statement and the dollar operator will be emphasized.

The assignment statement

The assignment statement is a very handy and simple way to define or change values associated with sets, parameters, variables and equations. It allows the user to enter data in the simplest way, and these data can later be elaborated with the assignment statement. In the previous section some examples have already been show.

GAMS uses the traditional arithmetic symbols for addition, subtraction, multiplication and division **(+, -, *,** /). For exponentiation GAMS uses the symbol ******. Every direct assignment should end with a semicolon.

In the second example above, the scalar PHI, and the parameters AVPR(J) and DEVPR(J,YE) have initially no value assigned to them; they are declared only. The values of AVPR and DEVPR are later determined by a function. Before the first solve statement PHI is first assigned a value of zero and is given a value of one afterwards. The term CARD in calculation of the average price is an internal variable in GAMS. It returns the number of elements in a set. With the DISPLAY statement one can check if the calculated values are actually correct. For example the statement DISPLAY AVPR(J), displays the following values:

```
---      71  PARAMETER AVPR            Average price in dollars per ton

RICE  334.000,          MAIZE  68.000,          CASSAVA  123.000,          SOYBEAN  242.000
```

In the next three paragraphs, the various possibilities on the right-hand of the = sign are examined in greater detail. Indexed operations and functions can also be used in the formulation of the equation.

Indexed operations

In GAMS there are four indexed operations: **SUM**, **PROD**, **SMIN**, and **SMAX**. **SUM** is the most commonly used operator. It is used to calculate the total over a set and is equivalent to the conventionally used symbol Σ. It is possible to SUM simultaneously over two or more sets. In this case the controlling sets are placed between brackets. If for example, price observations were available not only per year but also in different regions (I), the average price could be expressed as:

```
AVPR(J) = SUM( (I,YE) , PRICE(J,I,YE) / CARD(YE) ;
```

The **PROD** operator represents the conventional symbol Π and takes the product over the domain of the controlling sets. **SMIN** and **SMAX** are used to find the smallest and largest values respectively of the elements of the index set or sets. For example from the farm model the minimum and maximum price over the five-year period can be obtained.

```
PARAMETERS
        MINPRICE(J)             Minimum price for crop J
        MAXPRICE                Maximum price obtained for a crop ;

MINPRICE(J) = SMIN(YE, PRICE(J,YE));
```

```
MAXPRICE = SMAX( (J,YE) , PRICE(J,YE));
DISPLAY MINPRICE, MAXPRICE ;
```

Functions

In GAMS a whole series of functions can be used. A list of commonly used functions is provided in Appendix 4. For example:

```
L(J) = LOG(Y(J)) ;
```

assigns the value of the natural logarithm of Y to L over the domain J.

For exponentiation there are two possibilities, use either POWER(X,Y) or X**Y. The next table shows possible the results:

Value		Operation	
X	**Y**	**X**Y**	**POWER(X,Y)**
2	**2**	**4**	**4**
-2	**2**	**UNDF**	**4**
2	**2.1**	**4.287**	**UNDF**
INF	**2**	**INF**	**INF**
2	**INF**	**UNDF**	**UNDF**
NA	**2**	**NA**	**NA**

* INF = infinite, NA = not available, UNDF = undefined.

Matrix operations

The sum and subtraction with matrices can be realized quite simply in GAMS. To construct a matrix, which is the sum of the matrices A and B, use the following assignment:

```
PARAMETER   C(I,J)    Sum of A and B ;
C(I,J) = A(I,J) + B(I,J);
```

The new matrix C has I rows and J columns. Each element has a value equal to the sum of the correspondent elements of A and B.

The product of matrices works in a similar way. However, the matrix C is defined only, and only if, the number of columns of the matrix A equals the number of rows of matrix B. Thus, if A has the domain (I x J) and B the domain (J x K), J is not K, the product AB is possible, but the product BA is not. The inverse of a matrix can be calculated as follows:

```
PARAMETER   AT(J,I)  Inverse of A ;

AT(J,I) = A(I,J) ;
```

The dollar operator $

The dollar operator provides a powerful tool for handling exceptions. The general form is:

```
(expression) $ (condition)
```

The effect is that the expression will only apply if the condition is fulfilled. In general the condition is a relation, for example A has to be greater than B. In GAMS, this would be written:

```
(expression) $ (A GE B) ;
```

The different relational operators possible are:

```
LT      less than
LE      less than or equal to
EQ      equal to
NE      nor equal to
GE      greater than or equal to
GT      greater than
NOT     not equal to
AND     and
OR      or
XOR     either or
```

The $ operator can either be on the left- or the right-hand side of the relationship. With the "dollar on the right" we have the following result:

```
SCALAR      X, Y ;
            Y = 2 ;  X = 1 ;
            X = 2 $ (Y GT 1.5) ;
```

The last statement says: if y is greater than 1.5, assign the value 2 to X, if not assign the value 0. With the "dollar on the left" we have the following result:

```
SCALAR      X, Y ;
      Y = 2 ;  X = 1 ;
      X $ (Y GT 1.5) = 2;
```

which says: if the value of Y is greater than 1.5, assign the value 2 to X, if not do not make an assignment. X, thus, keeps its original value of 1.

The condition **$(*identifier* NE 0)** can also simply be expressed as **$(*identifier*)**. For example, the condition: **$(B(J) NE 0)** is equal to **$ B(J)**. In the example above, there is one such dollar operator,

```
PLOUGH(S)..   SUM(J $ SJ(S,J), X(J)) =L= PHIRE(S) ;
```

which states that we only summate over the domain J if J is also an element of SJ(S,J). If the dollar operator hadn't been included, all crop areas would have been summated whether the crops occupied the land in winter or summer.

The dollar operator is often useful to define a parameter which is the result of a division. This is because if the denominator is equal to zero, the result is undefined and GAMS will give an error message. To avoid this problem, the relation can be defined as:

```
PARAMETER   RATIO(J) ;
      RATIO(J) = A(J)/B(J) $ B(J) ;
```

The parameter RATIO only has a value if the value of B(J) is not equal to zero.

ORD and CARD

The function ORD returns the ordinal position of an element in a set, whereas CARD returns the number of elements in a set. For example:

```
SET     YE      Years    /1991*1995/;
PARAMETER   P number of years, N(YE) sequence ;
P = CARD(YE);
N(YE) = ORD(YE);
```

As a result of the assignments P will have a value of 5, while the value of N("1991") will be 1, N("1992") will be 2 and so on. In the farm model, the function CARD was used to calculate the average price.

```
AVPR(J) = SUM(YE,PRICE(J,YE))/CARD(YE) ;
```

In dynamic programming an important example with the function ORD is calculating the present value. For example:

```
SET                 T           periods          /1990*1995/ ;
SCALAR              R           discount rate    /0.05/ ;
PARAMETER   DISC                Calculation discount factor
                    ALPHA(T)    Discount factor ;
DISC = 1/(1+R) ;
ALPHA(T) = DISC**(ORD(T) - 1) ;
DISPLAY ALPHA ;
```

The objective function can then be formulated as something like:

```
Z =E= SUM(T, ALPHA(T) * (Expression))  ;
```

which gives, for example, the present value of all future income or savings, etc.

The sets created with the **ALIAS** statement can be used with ORD but not with CARD. With CARD it is necessary to use the original set.

Various models and solves

GAMS gives the possibility to create more than one model using the same system. For example, in the farm model problem it is possible to include a version in which it is not possible to rent or rent in out labour. To do this another labour balance equation is included next to the labour balance equation. In this way two versions of basically the same model are created. The changes made are as follows:

```
        .
        .
LABOURBAL1(T)..     SUM(J, X(J)*LABREQ(T,J)) + LOFF(T) - LHIRED(T) =L= FAMLAB  ;

LABOURBAL2(T)..     SUM(J, X(J)*LABREQ(T,J)) =L= FAMLAB ;
        .
        .

MODEL       FARM1       Model with possibility to hire and hire-out labour
                                /LANDBAL, LABOURBAL1, PLOUGH, DDEV, INCOME/;
MODEL       FARM2       Model with no possibility to hire and hire-out labour
                                /LANDBAL, LABOURBAL2, PLOUGH, DDEV, INCOME/;

SOLVE FARM1 USING LP MAXIMIZING Y;
SOLVE FARM2 USING LP MAXIMIZING Y;
```

The models created have the names FARM1 and FARM2. GAMS will solve the model twice: once including hiring labour and once without. Consequently, the results can be compared.

Another possibility to create more than one model is to change values using an assignment statement after the first SOLVE statement and then solve the model with different values again. This is done in the example above. After the first SOLVE the value of PHI was changed to zero. By doing this, a new model was created in which risk is not taken into

consideration. Again the results can be compared with the previous results. This method can be used to see how sensitive the model is to changes in the parameters (sensitivity analysis).

The INCLUDE option

With the option $INCLUDE the model can be separated into several smaller modules. In this way the different modules can be treated in individual ways to facilitate the understanding of the different variables used and to easily modify the parameters. This is especially handy with larger models. For example the model could be split into a definition, a data file, an equation file and a model and solve file. The main file will then look as follows:

```
$TITLE  FARM LEVEL MODEL

* This is the main file including all other files

$INCLUDE DEF.FIL
$INCLUDE DATA.FIL
$INCLUDE EQ.FIL
$INCLUDE SOLVE.FIL

* The DEF.FILE is the definition file, explaining all sets, scalars,
* parameters, tables, variables and equations used.
* The DATA.FILE is the data file assigning values to parameters, scalars
* and tables.
* The EQ.FILE is the equation file, determining all the algebraic
* relationships.
* The SOLVE.FILE includes the model and solve statement for two models,
* one with and one without risk consideration.
```

The LOOP statement

The LOOP statement is a very powerful device. It allows calculations to be made in a recursive rather than a parallel way. It is a very interesting tool especially for simulation. The general form is:

LOOP((***controlling set(s)*** **) ,** ***Statement(s)*** **;);**

The controlling set or sets determine the number of loops and over what set(s) looping takes place. There may be several statements inside the body of the LOOP. Aside from its applications in modeling, the LOOP statement is also used in the PUT statement.

The PUT statement

The purpose of the PUT writing facility is to organize the results of a solution under format control onto different files. In this way the results can be shown in an organized report or can be made ready for export/import into other programs, for example EXCEL. Although it takes more programming than the DISPLAY statement, the flexibility and control over the different items is much greater. Here the concept will be introduced by writing a .XLS file. This permits making graphs in EXCEL to visualize the output.

The syntax of the PUT statement is a bit complex. The basic structure is:

```
FILE    file name ;
PUT     file name ;
PUT     items     ;
```

where file name is the name used inside the GAMS model to refer to an external file. The second line activates the file. Items are any type of output, such as text, labels, parameters, etc., which are actually written to the active file.

To create a graph in EXCEL, which shows the use of family labour and hired labour in the different months, the following table could be created:

LABOUR REPORT				
	LABOUR DEMAND	**FAMILY LABOUR ON-FARM**	**LABOUR HIRED**	**FAMILY LABOUR OFF-FARM**
JAN **FEB** **MAR** . . .				

To create this table the PUT statement would be written as follows:

```
FILE    RESULTS        /FARM.XLS/ ;
PUT     RESULTS ;
RESULTS.PC = 5 ;
PUT     "LABOUR REPORT" //
PUT     " ", "LABOUR DEMAND", "FAMILY LABOUR ON-FARM", "LABOUR HIRED", "FAMILY
LABOUR OFF-
          FARM" // ;
LOOP (T, PUT T.TL, SUM (J, X.L(J) * LABREQ(T,J)), (FAMLAB - LOFF.L(T)), LHIRED.L(T),
LOFF.L(T)/) ;
```

At the first line the internal file name, RESULTS, is defined. This internal file name refers to the external EXCEL file FARM.XLS. The second line activates the file. The suffix **.PC** is an output format control. GAMS provides a variety of formats for page formatting and output style. The option PC = 5, stands for a comma delimited file and quoted text. Thus, in order for the output to appear in the correct cells in EXCEL, a comma has to be used to separate the different output items and quoted text has to be used. The next line writes the text "LABOUR REPORT" to the file. Forward slashes "/" are used to represent returns. The labels used for the table are written in the next put statement. They are separated by a comma. Each PUT statement is terminated with a semicolon. To present all the results of interest, simply write the PUT statement inside a LOOP statement which iterates over the set T (months). The names of the months are identified by using the suffix **.TL** which returns the set element labels (Jan, Feb,…). When the file FARM.XLS is opened in EXCEL after GAMS was executed, the results will look like:

	A	**B**	**C**	**D**	**E**
1	LABOUR REPORT				
2		LABOUR DEMAND	FAMILY LABOUR ON-FARM	LABOUR HIRED	FAMILY LABOUR OFF-FARM
3					
4	JAN	7.74	7.74	0	17.26
5	FEB	7.53	7.53	0	17.47
6	MAR	33.06	25	8.06	0
7	APR	31.91	25	6.91	0
8	MAY	23.38	23.38	0	1.62
9	JUN	16.16	16.16	0	8.84

10	JUL	21.2	21.2	0	3.8
11	AUG	24.03	24.03	0	0.97
12	SEP	13.16	13.16	0	11.84
13	OCT	37.18	25	12.18	0
14	NOV	40.72	25	15.72	0
15	DEC	7.12	7.12	0	17.88

This example just gives a short introduction to the PUT writing facility. There are many more possibilities. For these the reader is referred to the "Guide to the 'PUT' Writing Facility" found in the GAMS - Installation and System Notes.

Conclusion

GAMS is a powerful system designed to solve large and complex problems. It is possible to model complex agricultural problems in a comprehensive way. Its basic structure is very similar to the general algebraic formulation. The reader should now be ready to start modeling. In the last section there are several exercises, which will help the beginner. Answers can be found in Annex 3.

Exercises

Exercise I - a farm model

A farmer owns 12 hectares of land on which he can grow the following crops: sorghum, millet and cassava. He wants to maximize the gross margin. Next to land, the following resources are available: 80 man-days of labour, 8 days of ox-drawing power and 400 dollars of working capital. The gross revenue per hectare (price times yield) for the different crops is given in the table below.

Crop	Gross revenue ($/ha)
Sorghum	108.3
Millet	66.36
Cassava	127.58

It takes five days of labour to grow one hectare of sorghum and millet, and eight days to grow one hectare of cassava. Millet does not require any ploughing while sorghum and cassava each take one day per hectare. The capital requirements are 30 dollars per hectare for sorghum, 20 dollars for millet and 40 dollars for cassava (Deybe 1995).

Model the above problem in GAMS. Hint: Declare two SETS, one for crops and one for resources. For data entry use the PARAMETER statement and the TABLE statement.

Exercise II - a simple dynamic crop-irrigation model

Consider the highly simplified problem of a farmer who grows three horticultural crops in successive seasons over one year (taken from Kennedy 1986). Each crop takes four months, or one season to reach maturity from the time of planting. The crops are planted on 100 hectares of land and the yield (in hundred tons per 100 ha) in each season is determined by:

$$Y_t = w_t - 0.1 \cdot w_t^2$$

where, w_t is the depth of water in centimeters received by the crop grown in the *t*-th season. The depth of the water received depends on the amount of rainfall during each season (r_t in centimeters) and the height of water released from a dam (u_t in meters). If the area of the dam is 1 hectare, we can state that:

$$w_t = r_t + u_t$$

At the beginning of the first season the dam is full with a water height of 3 meter. The water level of the dam (x_t in meters) is depleted by releasing water for irrigation (u_t). Of course, the amount of water which can be released is constrained by the amount of water in storage.

$$0 \le u_t \le x_t$$

Conversely, rainfall increases the water level of the dam. The catchment area is 100 hectares, so 1 cm of rainfall raises the water level by 1 meter, provided that the dam is not full. If the latter is the case, water will be lost due to overflow. Thus,

$$x_{t+1} = MIN\left(\left(x_t - u_t + r_t\right), 3\right)$$

which states that the water level in the next period is either 3 meter or less.

The farmer's objective is to determine the level of u_t in each season, so that the present value of receipts from sales of the crops is maximized (*Z* in dollars).

$$Z = MAX \sum_{t=1}^{3} \alpha^{t-1} \cdot p_t \cdot Y_t$$

where, α is the discount factor (0.95) and p_t is the price in dollars per ton received for the *t*-th season crop. Data on rainfall and prices are given.

	Rainfall (cm)	Price ($/ton)
Season 1	2	50
Season 2	1	100
Season 3	1	150

Model the crop-irrigation problem in GAMS. Use DNLP to solve the model. This model contains non-linear relationships; consequently, LP can not be used. Furthermore, the variable *x* contains a discontinuity. That is why a special form of non-linear programming (NLP) is necessary, namely DNLP which is used for non-linear programming with discontinuous derivatives. It is basically the same as NLP, except that discontinuous functions can appear as well. The drawback is that they are more difficult to solve than normal NLP problems. Hint: The water levels in the second and third season are determined in an equation which does not include the first season. The initial water level at the beginning of the first season is determined in a separate equation, i.e.

```
EQUATIONS
INIT          initial water level
WATER(t)      water level in second and third season ;

INIT..        X("S1") =E= …….
WATER(t+1)    X(t+1) =E= ……
```

Exercise III - a Pakistani farmer

A) Basic model

A farmer in Pakistan wants to maximize his net return for farm activities. He can choose between four different crops. The inputs required are labour and land. In addition to family labour, temporary labour can be hired. We can distinguish two growing seasons: kharif and rabi.

In GAMS the SETS, SCALARS, PARAMETERS, TABLES and VARIABLES are defined as follows:

```
$TITLE      A PAKISTAN FARMER'S MODEL

SETS
            C       Crops                       /Wheat, Rice, Cotton, Sugarcane/
            I       Inputs                      /Land, Labour/
            S       Seasons                     /Kharif, Rabi/ ;

SCALARS
            FS      Farm Size (acres)                       /12.5/
            LC      Cost of hired labour (Rs per man-day)   /10/

PARAMETERS
            PRICE(C)            Crop prices (Rs per maund)
                                /wheat 41.15, rice 89.38, cotton 127.27, sugarcane 5.96/
            MAXFLAB(S)          Maximum seasonal family labour (man-days)
                                /kharif 390, rabi 530/
            YIELD(C)            Yield per acre (maund per acre per year)
                                /wheat 16, rice 11.4, cotton 10, sugarcane 375/
            REV(C)              Crop revenue (Rs per acre)  ;

TABLE INPUT(I, S, C) Input-output matrix for crops

*   This is a multi-dimensional table, giving the required inputs (I) in each season (S) for
*   each crop (C).

                    wheat       rice        cotton      sugarcane
land.kharif         .1          1.          1.          1.
land.rabi           1.          .1          .4          1.
labour.kharif       7.          21.         11.         15.
labour.rabi         16.         9.          10.           62.   ;

REV(C) = PRICE(C)*YIELD(C);

DISPLAY REV;

VARIABLES
        XCROP(C)        Cropping activities (acres)
        XLAB(S)         Hired labour (man-days)
        LCOST           Cost of hired labour (Rs)
        REVENUE         Gross revenue
        RETURN          Net return ;

POSITIVE VARIABLES XCROP, XLAB ;
```

In defining the land constraints, labour constraints and the objective function, we declare the following equations:

```
EQUATIONS
        LANDUSE(S)      Land use by season (acre)
```

```
LABUSE(S)      Labour use by season (man-days)
COSTLAB        Cost of hired labour (Rs)
 TREV          Total gross revenue (Rs)
OBJ            Net return (Rs) ;
```

The land constraint can be defined as:

```
LANDUSE(S)..  SUM(C, INPUT("LAND", S, C)*XCROP(C)) =L= FS ;
```

which states that, in every season the total land use has to be lower than the land available. Define the other four equations and solve the model.

B) Adding two inputs: capital and water

In addition to land and labour, two other inputs, capital and water, are also required. The maximum capital available per year is Rs 20,000. The average rainfall in kharif season is 700 inches and during rabi season 380 inches. In addition to this, water can be purchased at a rate of Rs 20 per inch. The input requirements per hectare for rice and cotton during kharif season are 65 and 25 inches, respectively. In rabi season the water requirements per hectare are 20 inches for wheat, 10 inches for cotton and 45 inches for sugarcane. The annual capital requirement per hectare for wheat is 150, rice 180, cotton 145, and sugarcane 500 Rupees.

As capital availability is given on a yearly basis, an extra set T, expressing the different time periods must be added. The set S will then be a subset of T. The set statement will then look as follows:

```
SETS
     C     Crops          /Wheat, Rice, Cotton, Sugarcane/
     I     Inputs         /Land, Labour, Capital, Water/
     T     Time periods   /Annual, Kharif, Rabi/
     S(T)  Seasons        /Kharif, Rabi/ ;
```

Complete the model incorporating the additional inputs.

C) Possibility of livestock production

Now the possibility of livestock production will be included. Assume the farmer keeps buffalo and cattle. The question is what is his optimal livestock production? For simplicity we assume that for livestock feed only fodder can be used. Fodder has a yield of 1 maund per acre. To produce fodder, 12 man-days of labour and 5 inches of water per hectare are needed in each season. The annual capital requirement is 120 Rupees per hectare and for ploughing 6 workdays of draftpower are necessary during rabi season. The draftpower requirements per hectare for the other crops are given below.

Draftpower requirement in work days

season/crop	wheat	rice	cotton	sugarcane
kharif	12	8		
rabi	6	1	7	30

Only buffalos provide draftpower. One buffalo can provide a maximum of 125 workdays of draftpower per season. For livestock production, the input requirements are given. In the GAMS model the table for input requirements is as follows:

TABLE LINPUT(*,T,H) Livestock input output table

	Buffalo	Cattle
Labour.kharif	**19.**	**14.**
Labour.rabi	**29.**	**21.5**
Capital.annual	**250.**	**180.**
Fodder.annual	**1.**	**.8 ;**

The asterisk indicates that any label can be used in the corresponding index position. Labour and Capital are elements of the set I (inputs for crop production) and Fodder is an element of the set C (crops). Consequently, set I can not be used, as fodder is not an element of set I. Therefore, we replace it with an asterisk. Buffalo and cattle have a gross revenue of Rs 1,750 and Rs 1,500, respectively.

Incorporate the possibility of livestock production and solve the model.

Solutions to these exercises can be found in Appendix 5.

References

Brooke, A.; Kendrick, D.; and Meeraus, A. 1992. GAMS , A User's Guide, Release 2.25. The Scientific Press Series, Massachusetts.

Deybe, D. 1995. L'écriture d'un modèle mathématique à l'aide du logiciel GAMS, Support de Cours, Notes et Documents URPA no 53, CIRAD, Paris.

GAMS Development Corporation. 1995. GAMS - Installation and System Notes. Washington.

Hazell, P.B.R.; and Norton, R.D. 1986. Mathematical Programming for Economic Analysis in Agriculture, MacMillan, New York.

Kutcher, G.P.; Meeraus, A.; and O'Mara, G.T. 1988. Agricultural Sector and Policy Models. The World Bank, Washington.

Kutcher, G.; and Scandizzo, P. 1981. The Agricultural Economy of Northeast Brazil. Johns Hopkins University Press, Baltimore.

Kennedy, John O.S. 1986. Dynamic Programming: Applications to Agriculture and Natural Resources. Elsevier Applied Science Publishers, London.

Schweigman, C. 1985. Operations Research Problems in Agriculture in Developing Countries. Khartoum University Press, Khartoum.

Appendix 1: Statistical Tables

Table 1 Distribution of t.

Degrees of Freedom	Probability								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	.718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
◎	.6745	.8416	1.2816	1.6448	1.9600	2.2414	2.5758	2.8070	3.2905

Table 2 Confidence intervals for binomial distribution.

95% Confidence Interval

Number Observed	Size of Sample, *n*												Fraction Observed	Size of Sample			
f	10		15		20		30		50		100		*f/n*	250		1000	
0	0	31	0	22	0	17	0	12	0	07	0	4	.00	0	1	0	0
1	0	45	0	32	0	25	0	17	0	11	0	5	.01	0	4	0	2
2	3	56	2	40	1	31	1	22	0	14	0	7	.02	1	5	1	3
3	7	65	4	48	3	38	2	27	1	17	1	8	.03	1	6	2	4
4	12	74	8	55	6	44	4	31	2	19	1	10	.04	2	7	3	5
5	19	81	12	62	9	49	6	35	3	22	2	11	.05	3	9	4	7
6	26	88	16	68	12	54	8	39	5	24	2	12	.06	3	10	5	8
7	35	93	21	73	15	59	10	43	6	27	3	14	.07	4	11	6	9
8	44	97	27	79	19	64	12	46	7	29	4	15	.08	5	12	6	10
9	55	100	32	84	23	68	15	50	9	31	4	16	.09	6	13	7	11
10	69	100	38	88	27	73	17	53	10	34	5	18	.10	7	14	8	12
11			45	92	32	77	20	56	12	36	5	19	.11	7	16	9	13
12			52	96	36	81	23	60	13	38	6	20	.12	8	17	10	14
13			60	98	41	85	25	63	15	41	7	21	.13	9	18	11	15
14			68	100	46	88	28	66	16	43	8	22	.14	10	19	12	16
15			78	100	51	91	31	69	18	44	9	24	.15	10	20	13	17
16					56	94	34	72	20	46	9	25	.16	11	21	14	18
17					62	97	37	75	21	48	10	26	.17	12	22	15	19
18					69	99	40	77	23	50	11	27	.18	13	23	16	21
19					75	100	44	80	25	53	12	28	.19	14	24	17	22
20					83	100	47	83	27	55	13	29	.20	15	26	18	23
21							50	85	28	57	14	30	.21	16	27	19	24
22							54	88	30	59	14	31	.22	17	28	19	25
23							57	90	32	61	15	32	.23	18	29	20	26
24							61	92	34	63	16	33	.24	19	30	21	27
25							65	94	36	64	17	35	.25	20	31	22	28
26							69	96	37	66	18	36	.26	20	32	23	29
27							73	98	39	68	19	37	.27	21	33	24	30
28							78	99	41	70	19	38	.28	22	34	25	31
29							83	100	43	72	20	39	.29	23	35	26	32
30							88	100	45	73	21	40	.30	24	36	27	33
31									47	75	22	41	.31	25	37	28	34
32									50	77	23	42	.32	26	38	29	35
33									52	79	24	43	.33	27	39	30	36
34									54	80	25	44	.34	28	40	31	37
35									56	82	26	45	.35	29	41	32	38
36									57	84	27	46	.36	30	42	33	39
37									59	85	28	47	.37	31	43	34	40
38									62	87	28	48	.38	32	44	35	41
39									64	88	29	49	.39	33	45	36	42
40									66	90	30	50	.40	34	46	37	43
41									69	91	31	51	.41	35	47	38	44
42									71	93	32	52	.42	36	48	39	45
43									73	94	33	53	.43	37	49	40	46
44									76	95	34	54	.44	38	50	41	47
45									78	97	35	55	.45	39	51	42	48
46									81	98	36	56	.46	40	52	43	49
47									83	99	37	57	.47	41	53	44	50
48									86	100	38	58	.48	42	54	45	51
49									89	100	39	59	.49	43	55	46	52
50									93	100	40	60	.50	44	56	47	53
											*			†		†	

Table 2 Confidence intervals for binomial distribution (continued).

99% Confidence Interval

Number Observed	Size of Sample, *n*												Fraction Observed	Size of Sample			
f	10		15		20		30		50		100		*f/n*	250		1000	
0	0	41	0	30	0	23	0	16	0	10	0	5	.00	0	2	0	1
1	0	54	0	40	0	32	0	22	0	14	0	7	.01	0	5	0	2
2	1	65	1	49	1	39	0	28	0	17	0	9	.02	1	6	1	3
3	4	74	2	56	2	45	1	32	1	20	0	10	.03	1	7	2	4
4	8	81	5	63	4	51	3	36	1	23	1	12	.04	2	9	3	6
5	13	87	8	69	6	56	4	40	2	26	1	13	.05	2	10	3	7
6	19	92	12	74	8	61	6	44	3	29	2	14	.06	3	11	4	8
7	26	96	16	79	11	66	8	48	4	31	2	16	.07	3	13	5	9
8	35	99	21	84	15	70	10	52	6	33	3	17	.08	4	14	6	10
9	46	100	26	88	18	74	12	55	7	36	3	18	.09	5	15	7	12
10	59	100	31	92	22	78	14	58	8	38	4	19	.10	6	16	8	13
11			37	95	26	82	16	62	10	40	4	20	.11	6	17	9	14
12			44	98	30	85	18	65	11	43	5	21	.12	7	18	9	15
13			51	99	34	89	21	68	12	45	6	23	.13	8	19	10	16
14			60	100	39	92	24	71	14	47	6	24	.14	9	20	11	17
15			70	100	44	94	26	74	15	49	7	26	.15	9	22	12	18
16					49	96	29	76	17	51	8	27	.16	10	23	13	19
17					55	98	32	79	18	53	9	29	.17	11	24	14	20
18					61	99	35	82	20	55	9	30	.18	12	25	15	21
19					68	100	38	84	21	57	10	31	.19	13	26	16	22
20					77	100	42	86	23	59	11	32	.20	14	27	17	23
21							45	88	24	61	12	33	.21	15	28	18	24
22							48	90	26	63	12	34	.22	16	30	19	26
23							52	92	28	65	13	35	.23	17	31	20	27
24							56	94	29	67	14	36	.24	18	32	21	28
25							60	96	31	69	15	38	.25	18	33	22	29
26							64	97	33	71	16	39	.26	19	34	22	30
27							68	99	35	72	16	40	.27	20	35	23	31
28							72	100	37	74	17	41	.28	21	36	24	32
29							78	100	39	76	18	42	.29	22	37	25	33
30							84	100	41	77	19	43	.30	23	38	26	34
31									43	79	20	44	.31	24	39	27	35
32									45	80	21	45	.32	25	40	28	36
33									47	82	21	46	.33	26	41	29	37
34									49	83	22	47	.34	26	42	30	38
35									51	85	23	48	.35	27	43	31	39
36									53	86	24	49	.36	28	44	32	40
37									55	88	25	50	.37	29	45	33	41
38									57	89	26	51	.38	30	46	34	42
39									60	90	27	52	.39	31	47	35	43
40									62	92	28	53	.40	32	48	36	44
41									64	93	29	54	.41	33	50	37	45
42									67	94	29	55	.42	34	51	38	46
43									69	96	30	56	.43	35	52	39	47
44									71	97	31	57	.44	36	53	40	48
45									74	98	32	58	.45	37	54	41	49
46									77	99	33	59	.46	38	55	42	50
47									80	99	34	60	.47	39	55	43	51
48									83	100	35	61	.48	40	56	44	52
49									86	100	36	62	.49	41	57	45	53
50									90	100	37	63	.50	42	58	46	54
											*		†	†			

Table 3 Arc sine transformation.

%	0	1	2	3	4	5	6	7	8	9
0.0	0	0.57	0.81	0.99	1.15	1.28	1.40	1.52	1.62	1.72

%	0	1	2	3	4	5	6	7	8	9
0.1	1.81	1.90	1.99	2.07	2.14	2.22	2.29	2.36	2.43	2.50
0.2	2.56	2.63	2.69	2.75	2.81	2.87	2.92	2.98	3.03	3.09
0.3	3.14	3.19	3.24	3.29	3.34	3.39	3.44	3.49	3.53	3.58
0.4	3.63	3.67	3.72	3.76	3.80	3.85	3.89	3.93	3.97	4.01
0.5	4.05	4.09	4.13	4.17	4.21	4.25	4.29	4.33	4.37	4.40
0.6	4.44	4.48	4.52	4.55	4.59	4.62	4.66	4.69	4.73	4.76
0.7	4.80	4.83	4.87	4.90	4.93	4.97	5.00	5.03	5.07	5.10
0.8	5.13	5.16	5.20	5.23	5.26	5.29	5.32	5.35	5.38	5.41
0.9	5.44	5.47	5.50	5.53	5.56	5.59	5.62	5.65	5.68	5.71
1		6.02	6.29	6.55	6.80	7.04	7.27	7.49	7.71	7.92
2		8.33	8.53	8.72	8.91	9.10	9.28	9.46	9.63	9.81
3		10.14	10.31	10.47	10.63	10.78	10.94	11.09	11.24	11.39
4	11.54	11.68	11.83	11.97	12.11	12.25	12.39	12.52	12.66	12.79
5	12.92	13.05	13.18	13.31	13.44	13.56	13.69	13.81	13.94	14.06
6	14.18	14.30	14.42	14.54	14.65	14.77	14.89	15.00	15.12	15.23
7	15.34	15.45	15.56	15.68	15.79	15.89	16.00	16.11	16.22	16.32
8	16.43	16.54	16.64	16.74	16.85	16.95	17.05	17.16	17.26	17.36
9	17.46	17.56	17.66	17.76	17.85	17.95	18.05	18.15	18.24	18.34
10	18.44	18.53	18.63	18.72	18.81	18.91	19.00	19.09	19.19	19.28
11	19.37	19.46	19.55	19.64	19.73	19.82	19.91	20.00	20.09	20.18
12	20.27	20.36	20.44	20.53	20.62	20.70	20.79	20.88	20.96	21.05
13	21.13	21.22	21.30	21.39	21.47	21.56	21.64	21.72	21.81	21.89
14	21.97	22.06	22.14	22.22	22.30	22.38	22.46	22.55	22.63	22.71
15	22.79	22.87	22.95	23.03	23.11	23.19	23.26	23.34	23.42	23.50
16	23.58	23.66	23.73	23.81	23.89	23.97	24.04	24.12	24.20	24.27
17	24.35	24.43	24.50	24.58	24.65	24.73	24.80	24.88	24.95	25.03
18	25.10	25.18	25.25	25.33	25.40	25.48	25.55	25.62	25.70	25.77
19	25.84	25.92	25.99	26.06	26.13	26.21	26.28	26.35	26.42	26.49
20	26.56	26.64	26.71	26.78	26.85	26.92	26.99	27.06	27.13	27.20
21	27.28	27.35	27.42	27.49	27.56	27.63	27.69	27.76	27.83	27.90
22	27.97	28.04	28.11	28.18	28.25	28.32	28.38	28.45	28.52	28.59
23	28.66	28.73	28.79	28.86	29.00	29.00	29.06	29.13	29.20	29.27
24	29.33	29.40	29.47	29.53	29.60	29.67	29.73	29.80	29.87	29.93
25	30.00	30.07	30.13	30.20	30.26	30.33	30.40	30.46	30.53	30.59
26	30.66	30.72	30.79	30.85	30.92	30.98	31.05	31.11	31.18	31.24
27	31.31	31.37	31.44	31.50	31.56	31.63	31.69	31.76	31.82	31.88
28	31.95	32.01	32.08	32.14	32.20	32.27	32.33	32.39	32.46	32.52
29	32.58	32.65	32.71	32.77	32.83	32.90	32.96	33.02	33.09	33.15
30	33.21	33.27	33.34	33.40	33.46	33.52	33.58	33.65	33.71	33.77
31	33.83	33.89	33.96	34.02	34.08	34.14	34.20	34.27	34.33	34.39
32	34.45	34.51	34.57	34.63	34.70	34.76	34.82	34.88	34.94	35.00
33	35.06	35.12	35.18	35.24	35.30	35.37	35.43	35.49	35.55	35.61
34	35.67	35.73	35.79	35.85	35.91	35.97	36.03	36.09	36.15	36.21
35	36.27	36.33	36.39	36.45	36.51	36.57	36.63	36.69	36.75	36.81
36	36.87	36.93	36.99	37.05	37.11	37.17	37.23	37.29	37.35	37.41
37	37.47	37.52	37.58	37.64	37.70	37.76	37.82	37.88	37.94	38.00
38	38.06	38.12	38.17	38.23	38.29	38.35	38.41	38.47	38.53	38.59
39	38.65	38.70	38.76	38.82	38.88	38.94	39.00	39.06	39.11	39.17
40	39.23	39.29	39.35	39.41	39.47	39.52	39.58	39.64	39.70	39.76
41	39.82	39.87	39.93	39.99	40.05	40.11	40.16	40.22	40.28	40.34
42	40.40	40.46	40.51	40.57	40.63	40.69	40.74	40.80	40.86	40.92
43	40.98	41.03	41.09	41.15	41.21	41.27	41.32	41.38	41.44	41.50
44	41.55	41.61	41.67	41.73	41.78	41.84	41.90	41.96	42.02	42.07
45	42.13	42.19	42.25	42.30	42.36	42.42	42.48	42.53	42.59	42.65
46	42.71	42.76	42.82	42.88	42.94	42.99	43.05	43.11	43.17	43.22
47	43.28	43.34	43.39	43.45	43.51	43.57	43.62	43.68	43.74	43.80
48	43.85	43.91	43.97	44.03	44.08	44.14	44.20	44.25	44.31	44.37
49	44.43	44.48	44.54	44.60	44.66	44.71	44.77	44.83	44.89	44.94
50	45.00	45.06	45.11	45.17	45.23	45.29	45.34	45.40	45.46	45.52
51	45.57	45.63	45.69	45.75	45.80	45.86	45.92	45.97	46.03	46.09
52	46.15	46.20	46.26	46.32	46.38	46.43	46.49	46.55	46.61	46.66
53	46.72	46.78	46.83	46.89	46.95	47.01	47.06	47.12	47.18	47.24
54	47.29	47.35	47.41	47.47	47.52	47.58	47.64	47.70	47.75	47.81

Table 3 Arc sine transformation (continued).

%	0	1	2	3	4	5	6	7	8	9
55	47.87	47.93	47.98	48.04	48.10	48.16	48.22	48.27	48.33	48.39
56	48.45	48.50	48.56	48.62	48.68	48.73	48.79	48.85	48.91	48.97
57	49.02	49.08	49.14	49.20	49.26	49.31	49.37	49.43	49.49	49.54
58	49.60	49.66	49.72	49.78	49.84	49.89	49.95	50.01	50.07	50.13
59	50.18	50.24	50.30	50.36	50.42	50.48	50.53	50.59	50.65	50.71
60	50.77	50.83	50.89	50.94	51.00	51.06	51.12	51.18	51.24	51.30
61	51.35	51.41	51.47	51.53	51.59	51.65	51.71	51.77	51.83	51.88
62	51.94	52.00	52.06	52.12	52.18	52.24	52.30	52.36	52.42	52.48
63	52.53	52.59	52.65	52.71	52.77	52.83	52.89	52.95	53.01	53.07
64	53.13	53.19	53.25	53.31	53.37	53.43	53.49	53.55	53.61	53.67
65	53.73	53.79	53.85	53.91	53.97	54.03	54.09	54.15	54.21	54.27
66	54.33	54.39	54.45	54.51	54.57	54.63	54.70	54.76	54.82	54.88
67	54.94	55.00	55.06	55.12	55.18	55.24	55.30	55.37	55.43	55.49
68	55.55	55.61	55.67	55.73	55.80	55.86	55.92	55.98	56.04	56.11
69	56.17	56.23	56.29	56.35	56.42	56.48	56.54	56.60	56.66	56.73
70	56.79	56.85	56.91	56.98	57.04	57.10	57.17	57.23	57.29	57.35
71	57.42	57.48	57.54	57.61	57.67	57.73	57.80	57.86	57.92	57.99
72	58.05	58.12	58.18	58.24	58.31	58.37	58.44	58.50	58.56	58.63
73	58.69	58.76	58.82	58.89	58.95	59.02	59.08	59.15	59.21	59.28
74	59.34	59.41	59.47	59.54	59.60	59.67	59.74	59.80	59.87	59.93
75	60.00	60.07	60.13	60.20	60.27	60.33	60.40	60.47	60.53	60.60
76	60.67	60.73	60.80	60.87	60.94	61.00	61.07	61.14	61.21	61.27
77	61.34	61.41	61.48	61.55	61.62	61.68	61.75	61.82	61.89	61.96
78	62.03	62.10	62.17	62.24	62.31	62.37	62.44	62.51	62.58	62.65
79	62.72	62.80	62.87	62.94	63.01	63.08	63.15	63.22	63.29	63.36
80	63.44	63.51	63.58	63.65	63.72	63.79	63.87	63.94	64.01	64.08
81	64.16	64.23	64.30	64.38	64.45	64.52	64.60	64.67	64.75	64.82
82	64.90	64.97	65.05	65.12	65.20	65.27	65.35	65.42	65.50	65.57
83	65.65	65.73	65.80	65.88	65.96	66.03	66.11	66.19	66.27	66.34
84	66.42	66.50	66.58	66.66	66.74	66.81	66.89	66.97	67.05	67.13
85	67.21	67.29	67.37	67.45	67.54	67.62	67.70	67.78	67.86	67.94
86	68.03	68.11	68.19	68.28	68.36	68.44	68.53	68.61	68.70	68.78
87	68.87	68.95	69.04	69.12	69.21	69.30	69.38	69.47	69.56	69.64
88	69.73	69.82	69.91	70.00	70.09	70.18	70.27	70.36	70.45	70.54
89	70.63	70.72	70.81	70.91	71.00	71.09	71.19	71.28	71.37	71.47
90	71.56	71.66	71.76	71.85	71.95	72.05	72.15	72.24	72.34	72.44
91	72.54	72.64	72.78	72.84	72.95	73.05	73.15	73.26	73.36	73.46
92	73.57	73.68	73.74	73.89	74.00	74.11	74.21	74.32	74.44	74.55
93	74.66	74.77	74.88	75.00	75.11	75.23	75.35	75.46	75.58	75.70
94	75.82	75.94	76.06	76.19	76.31	76.44	76.56	76.69	76.82	76.95
95	77.08	77.21	77.34	77.48	77.61	77.75	77.89	78.03	78.17	78.32
96	78.46	78.61	78.76	78.91	79.06	79.22	79.37	79.53	79.69	79.86
97	80.02	80.19	80.37	80.54	80.72	80.90	81.09	81.28	81.47	81.67
98	81.87	82.08	82.29	82.51	82.73	82.96	83.20	83.45	83.71	83.98
99.0	84.26	84.29	84.32	84.35	84.38	84.41	84.44	84.47	84.50	84.53
99.1	84.56	84.59	84.62	84.65	84.68	84.71	84.74	84.77	84.80	84.84
99.2	84.87	84.90	84.93	84.97	85.00	85.03	85.07	85.10	85.13	85.17
99.3	85.20	85.24	85.27	85.31	85.34	85.38	85.41	85.45	85.48	85.52
99.4	85.56	85.60	85.63	85.67	85.71	85.75	85.79	85.83	85.87	85.91
99.5	85.95	85.99	86.03	86.07	86.11	86.15	86.20	86.24	86.28	86.33
99.6	86.37	86.42	86.47	86.51	86.56	86.61	86.66	86.71	86.76	86.81
99.7	86.86	86.91	86.97	87.02	87.08	87.13	87.19	87.25	87.31	87.37
99.8	87.44	87.50	87.57	87.64	87.71	87.78	87.86	87.93	88.01	88.10
99.9	88.19	88.28	88.38	88.48	88.60	88.72	88.85	89.01	89.19	89.43
100.0	90.00	-	-	-	-	-	-	-	-	-

Table 4 Accumulative distribution of chi-square.

Degrees of Freedom	Probability												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1	...	...	...	...	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5..01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17

Table 5 Standard normal cumulative distribution.

$$F(z) = P[Z \le z]$$

z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$
.00	.5000						
.01	.5040	.51	.6950	1.01	.8438	1.51	.9345
.02	.5080	.52	.6985	1.02	.8461	1.52	.9357
.03	.5120	.53	.7019	1.03	.8485	1.53	.9380
.04	.5160	.54	.7054	1.04	.8508	1.54	.9382
.05	.5199	.55	.7088	1.05	.8531	1.55	.9394
.06	.5239	.56	.7123	1.06	.8554	1.56	.9406
.07	.5279	.57	.7157	1.07	.8577	1.57	.9418
.08	.5319	.58	.7190	1.08	.8599	1.58	.9429
.09	.5359	.59	.7224	1.09	.8621	1.59	.9441
.10	.5398	.60	.7257	1.10	.8643	1.60	.9452
.11	.5438	.61	.7291	1.11	.8665	1.61	.9463
.12	.5478	.62	.7324	1.12	.8686	1.62	.9474
.13	.5517	.63	.7357	1.13	.8708	1.63	.9484
.14	.5557	.64	.7389	1.14	.8729	1.64	.9495
.15	.5596	.65	.7422	1.15	.8749	1.65	.9505
.16	.5636	.66	.7454	1.16	.8770	1.66	.9515
.17	.5675	.67	.7486	1.17	.8790	1.67	.9525
.18	.5714	.68	.7517	1.18	.8810	1.68	.9535
.19	.5753	.69	.7549	1.19	.8830	1.69	.9545
.20	.5793	.70	.7580	1.20	.8849	1.70	.9554
.21	.5832	.71	.7611	1.21	.8869	1.71	.9564
.22	.5871	.72	.7642	1.22	.8888	1.72	.9573
.23	.5910	.73	.7673	1.23	.8907	1.73	.9582
.24	.5948	.74	.7704	1.24	.8925	1.74	.9591
.25	.5987	.75	.7734	1.25	.8944	1.75	.9599
.26	.6026	.76	.7764	1.26	.8962	1.76	.9608
.27	.6064	.77	.7794	1.27	.8980	1.77	.9616
.28	.6103	.78	.7823	1.28	.8997	1.78	.9625
.29	.6141	.79	.7852	1.29	.9015	1.79	.9633
.30	.6179	.80	.7881	1.30	.9032	1.80	.9641
.31	.6217	.81	.7910	1.31	.9049	1.81	.9649
.32	.6255	.82	.7939	1.32	.9066	1.82	.9656
.33	.6293	.83	.7967	1.33	.9082	1.83	.9664
.34	.6331	.84	.7995	1.34	.9099	1.84	.9671
.35	.6368	.85	.8023	1.35	.9115	1.85	.9678
.36	.6406	.86	.8051	1.36	.9131	1.86	.9686
.37	.6443	.87	.8078	1.37	.9147	1.87	.9693
.38	.6480	.88	.8106	1.38	.9162	1.88	.9699
.39	.6517	.89	.8133	1.39	.9177	1.89	.9706
.40	.6554	.90	.8159	1.40	.9192	1.90	.9713
.41	.6591	.91	.8186	1.41	.9207	1.91	.9719
.42	.6628	.92	.8212	1.42	.9222	1.92	.9726
.43	.6664	.93	.8238	1.43	.9236	1.93	.9732
.44	.6700	.94	.8264	1.44	.9251	1.94	.9738
.45	.6736	.95	.8289	1.45	.9265	1.95	.9744
.46	.6772	.96	.8315	1.46	.9279	1.96	.9750
.47	.6803	.97	.8340	1.47	.9292	1.97	.9756
.48	.6844	.98	.8365	1.48	.9306	1.98	.9761
.49	.6879	.99	.8389	1.49	.9319	1.99	.9767
.50	.6915	1.00	.8413	1.50	.9332	2.00	.9772

Table 5 Standard normal cumulative distribution (continued).

z	*F (z)*	*z*	*F (z)*	*z*	*F (z)*	*z*	*F (z)*
2.01	.9778	2.51	.9940	3.01	.9987	3.51	.9998
2.02	.9783	2.52	.9941	3.02	.9987	3.52	.9998
2.03	.9788	2.53	.9943	3.03	.9988	3.53	.9998
2.04	.9793	2.54	.9945	3.04	.9988	3.54	.9998
2.05	.9798	2.55	.9946	3.05	.9989	3.55	.9998
2.06	.9803	2.56	.9948	3.06	.9989	3.56	.9998
2.07	.9808	2.57	.9949	3.07	.9989	3.57	.9998
2.08	.9812	2.58	.9951	3.08	.9990	3.58	.9998
2.09	.9817	2.59	.9952	3.09	.9990	3.59	.9998
2.10	.9821	2.60	.9953	3.10	.9990	3.60	.9998
2.11	.9826	2.61	.9955	3.11	.9991	3.61	.9998
2.12	.9830	2.62	.9956	3.12	.9991	3.62	.9999
2.13	.9834	2.63	.9957	3.13	.9991	3.63	.9999
2.14	.9838	2.64	.9959	3.14	.9992	3.64	.9999
2.15	.9840	2.65	.9960	3.15	.9992	3.65	.9999
2.16	.9846	2.66	.9961	3.16	.9992	3.66	.9999
2.17	.9850	2.67	.9962	3.17	.9992	3.67	.9999
2.18	.9854	2.68	.9963	3.18	.9993	3.68	.9999
2.19	.9857	2.69	.9964	3.19	.9993	3.69	.9999
2.20	.9861	2.70	.9965	3.20	.9993	3.70	.9999
2.21	.9864	2.71	.9966	3.21	.9993	3.71	.9999
2.22	.9868	2.72	.9967	3.22	.9994	3.72	.9999
2.23	.9871	2.73	.9968	3.23	.9994	3.73	.9999
2.24	.9875	2.74	.9969	3.24	.9994	3.74	.9999
2.25	.9878	2.75	.9970	3.25	.9994	3.75	.9999
2.26	.9881	2.76	.9971	3.26	.9994	3.76	.9999
2.27	.9884	2.77	.9972	3.27	.9995	3.77	.9999
2.28	.9887	2.78	.9973	3.28	.9995	3.78	.9999
2.29	.9890	2.79	.9974	3.29	.9995	3.79	.9999
2.30	.9893	2.80	.9974	3.30	.9995	3.80	.9999
2.31	.9896	2.81	.9975	3.31	.9995	3.81	.9999
2.32	.9998	2.82	.9976	3.32	.9996	3.82	.9999
2.33	.9901	2.83	.9977	3.33	.9996	3.83	.9999
2.34	.9904	2.84	.9977	3.34	.9996	3.84	.9999
2.35	.9906	2.85	.9978	3.35	.9996	3.85	.9999
2.36	.9909	2.86	.9979	3.36	.9996	3.86	.9999
2.37	.9911	2.87	.9979	3.37	.9996	3.87	.9999
2.38	.9913	2.88	.9980	3.38	.9996	3.88	.9999
2.39	.9916	2.89	.9981	3.39	.9997		
2.40	.9918	2.90	.9981	3.40	.9997		
2.41	.9920	2.91	.9982	3.41	.9997		
2.42	.9922	2.92	.9982	3.42	.9997		
2.43	.9925	2.93	.9983	3.43	.9997		
2.44	.9927	2.94	.9984	3.44	.9997		
2.45	.9929	2.95	.9984	3.45	.9997		
2.46	.9931	2.96	.9985	3.46	.9997		
2.47	.9932	2.97	.9985	3.47	.9997		
2.48	.9934	2.98	.9986	3.48	.9997		
2.49	.9936	2.99	.9986	3.49	.9998		
2.50	.9938	3.00	.9986	3.50	.9998		

Table 6 Distribution of F.

5% (Roman Type) and 1% (Bold Face Type) Points for the Distribution of F

f_2	f_1 degrees of freedom (for greater mean square)																								f_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	8	
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254	1
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,258**	**6,286**	**6,302**	**6,323**	**6,334**	**6,352**	**6,361**	**6,366**	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.38	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50	2
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53	3
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63	4
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	5
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67	6
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23	7
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93	8
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71	9
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54	10
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**	
11	4.84	3.98	3.59	3.36	3..20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40	11
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**	
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30	12
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**	
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21	13
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	14
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07	15
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01	16
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	17
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**	

Table 6 Distribution of F (continued).

	f_1 degrees of freedom (for greater mean square)																								
$f2$	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	8	$f2$
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92	18
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88	19
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84	20
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81	21
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78	22
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	23
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73	24
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**	
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71	25
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**	
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69	26
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67	27
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**	
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65	28
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**	
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.60	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64	29
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**	
30	4.17	3.32	2.92	2.68	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62	30
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**	
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.68	1.67	1.64	1.61	1.59	32
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**	
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57	34
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**	
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.68	1.65	1.62	1.59	1.56	1.55	36
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**	
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53	38
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51	40
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**	

Table 6 Distribution of F (continued).

	f_1 degrees of freedom (for greater mean square)																								
f_2	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	f_2
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49	42
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**	
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48	44
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**	
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46	46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**	
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45	48
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**	
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44	50
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**	
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41	55
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**	
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39	60
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**	
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37	65
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**	
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35	70
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**	
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32	80
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**	
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28	100
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**	
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25	125
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**	
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22	150
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**	
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19	200
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**	
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13	400
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**	
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08	1000
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**	
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00	∞
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**	

Appendix 2: Solution to Linear Programming Exercises

A) Solutions using Excel

Exercise I - Linear programming model

Using *Excel*, the spreadsheet appears as follows after solving the problem:

	A	B	C	D	E	F
1						
2		**Farm Activity**	**Production of Maize**	**Production of Wheat**	**LHS**	**RHS**
3		Hectare	0.4285714	3.5714286		**Maximize**
4						
5		Selling price (in 1000 Rp/kg)	3	5		
6		Costs (in 1,000,000 Rp/ha)	4	2.5		
7		Yield (in 1000 kg/ha)	1.8	1		
8		**Gross Margin per ha**	1.4	2.5	*((Selling price * Yield) - Costs)*	
9						
10		**Resource constraints:**				
11		1. Land (ha)	1	1	4	4
12		2. Labour (hours/two months)	150	200	778.5714	800
13		3. Labour (hours/two months)	200	150	621.4286	800
14		4. Ox-Plough (hours/two months)	56	0	24	200
15		5. Ox-Plough (hours/two months)	0	56	200	200
16		Non-negative:	1	0	0.428571	0
17			0	1	3.571429	0
18						
19		**Objective**				9.528571

The optimal solution is to cultivate 0.43 hectares with maize and 3.57 hectares with wheat. The availability of land and hours of ox-plough are the binding constraints. The maximum revenue is Rp 9.5 million.

Exercise II - Soils with different fertility

In this problem there are four types of farm activity:

X_1 = Growing maize on soil type I
X_2 = Growing wheat on soil type I
X_3 = Growing maize on soil type II
X_4 = Growing wheat on soil type II

After filling in the data and solving the problem, the spreadsheet for this problem is appears follows:

	A	B	C	D	E	F	G	H	I
1									
2		**Farm Activity**		**X_1**	**X_2**	**X_3**	**X_4**	**LHS**	**RHS**
3		Hectare		0.41904762	1.28571429	0	2		**Maximize**
4									
5		Selling price (in 1000 Rp/kg)		3	5	3	5		
6		Costs (in 1,000,000 Rp/ha)		4	2.5	4.5	3		

7		Yield (in 1000 kg/ha)	1.8	1	1.8	1.2		
8		**Gross Margin per ha**	1.4	2.5	0.9	3		
9								
10		**Resource constraints:**						
11		1. Land (ha) Soil Type 1	1	1			1.704762	2
12		2. Land (ha) Soil Type 2			1	1	2	2
13		3. Labour (hours/two months)	200	150	240	165	606.6667	800
14		4. Labour (hours/two months)	150	200	180	240	800	800
15		5. Ox-Plough (hours/two months)	56		64		23.46667	200
16		6. Ox-Plough (hours/two months)		56		64	200	200
17		Non-negative	1				0.419048	0
18				1			1.285714	0
19					1		0	0
20						1	2	0
		Objective						9.800952

The maximum revenue obtainable is Rp 9.8 million. This will be realized if the land is devided as shown in row 3. The binding constraints are land availability of soil type 2, labour in the period November-December, and the availability of ox-drawing power.

Exercise III - Fertilizer or manure

Now three farm activities are considered. The choice will be between producing maize making use of fertilizer or manure and producing wheat. The spreadsheet is as follows:

	A	B	C	D	E	F	G	H
1								
2		**Farm Activity**		**Production of Maize without Fertilizer**	**Production of Maize with Fertilizer**	**Production of Wheat**	**LHS**	**RHS**
3		Hectare		0.214285714	0.214285714	3.571428571		**Maximize**
4								
5		Selling price (in 1000 Rp/kg)		3	3	5		
6		Costs (in 1000 Rp/ha)		4	5	2.5		
7		Yield (in 1000 kg/ha)		1.8	2.34	1		
8		**Gross Margin per ha**		1.4	2.02	2.5		
9								
10		**Resource Constraints:**						
11		1. Land (ha)		1	1	1	4	4
12		2. Labour (hours/two months)		200	200	150	621.4286	800
13		3. Labour (hours/two months)		150	250	200	800	800
14		4. Ox-Plough (hours/two months)		56	56		24	200
15		5. Ox-Plough (hours/two months)				56	200	200
16		Non-negative:C33		1			0.214286	0
17					1		0.214286	0
18						1	3.571429	0
19								
20		**Objective**						9.661429

Given the availability of land, labour and ox-plough, the best the farmer can do is to divide the land as indicated in row 3. On half of the land allocated to maize production, fertilizer is used and on the other half, manure is used. The maximum revenue will then be Rp 9.7 million.

B) Solutions using GAMS

For the three exercises only the setup of the input file will be provided. After running GAMS, the output file created can easily be read by using an editor.

Exercise I - Linear programming model

The LP model can be set up in GAMS using two sets, one for the different crops and one for the time periods. For the input coefficients, the labour and ox-plough requirements, a table statement is used. Tables in general have two dimensions, in this case the different time periods as rows and the crops as columns. Scalars are used for the availability of land, labour and ox and with a parameter statement the data on yield, prices and costs are entered. The decision variable is the number of hectares planted with each crop and the objective variable is the farmer's income.

```
SETS
c                crops              / maize, wheat /
t                time period        / jan-feb, apr-may, sep-oct, nov-dec/;

TABLE
Labreq(t,c)      labour requirements (man hours per ha)

           maize   wheat
jan-feb    200.00  150.00
apr-may    180.00
sep-oct            100.00
nov-dec    150.00  200.00 ;

TABLE
 Oxreq(t,c)      ox-plough requirement (hours per ha)

           maize   wheat
sep-oct            56
nov-dec    56              ;

PARAMETERS
Yield(c)  crop yield (tons per hectare)          / maize = 1800, wheat = 1000 /
Price(c)         crop product prices (Rp per kg)          / maize = 3000, wheat = 5000 /
Pcost(c)         production costs (Rp per ha)             / maize = 4000000, wheat = 2500000 /
Margin(c)        net margin per hectare;

Margin(c)=Yield(c)*Price(c)-Pcost(c);

SCALARS
labour           family labour available                  / 800 /
ox               ox power available                       / 200 /
land             land availability                        / 4 /  ;



VARIABLES
```

```
xcrop(c)          the level at which each cropping activity is chosen
inc               the net income ;

POSITIVE VARIABLE  xcrop ;

EQUATIONS
landbal           land balance (hectares)
labavail(t)               labour balance (hours)
Oxavail(t)                availability of Ox (hours)
income            the net income;

landbal..         sum(c, xcrop(c)) =l= land;
labavail(t)..     sum(c, xcrop(c)*labreq(t,c)) =l= labour;
Oxavail(t)..      sum(c, xcrop(c)*oxreq(t,c)) =l= ox;
income..          inc =e= sum(c, xcrop(c)*margin(c));


MODEL farm1       farm  model        / all / ;

SOLVE  farm1      USING LP MAXIMIZING inc ;
```

Exercise II - Soils with different fertility

Now set c is extended to distinguish four different farm activities. A set for the different soil types is also included, as is an extra table for the land requirement. Instead of a scalar, a parameter is used to define the land available for both soil types.

```
SETS
c                 crops                     / maize1, wheat1, maize2, wheat2 /
t                 time periods              / jan-feb, apr-may, sep-oct, nov-dec/
l                 land type                         /soil-I, soil-II/;

TABLE
Labreq(t,c)       labour requirements (man hours per ha)

          maize1  wheat1  maize2  wheat2
jan-feb   200.00  150.00  240.00  165.00
apr-may   180.00          198.00
sep-oct           100.00          120.00
nov-dec   150.00  200.00  180.00  240.00  ;


TABLE
Oxreq(t,c)                ox-plough requirement (hours per ha)

          maize1  wheat1  maize2  wheat2
sep-oct           56              64
nov-dec   56              64              ;

TABLE
LANDREQ(l,c)      landrequirement

          maize1  wheat1  maize2  wheat2
soil-I    1       1
soil-II                   1       1       ;


PARAMETERS
land(l)   land availability
```

```
                    /soil-I = 2, soil-II = 2/
yield(c)  crop yield (tons per hectare)
                    / maize1 = 1800, wheat1 = 1000, maize2 = 1800, wheat2 = 1200 /
price(c)  crop product prices (Rp per kg)
                    / maize1 = 3000, wheat1 = 5000, maize2 = 3000, wheat2 = 5000 /
pcost(c)  production costs (Rp per ha)
                    / maize1 = 4000000, wheat1 = 2500000, maize2 = 4500000, wheat2 = 3000000 /
Margin(c)           net margin per hectare;

Margin(c)=yield(c)*price(c)-pcost(c);

SCALARS
labour    family labour available     / 800 /
ox        Ox power available          / 200 / ;

VARIABLES
xcrop(c)            the level at which each cropping activity is chosen
inc       the net income ;

POSITIVE VARIABLE xcrop;

EQUATIONS
landbal(l)          landbalance (hectares)
labavail(t)         labour balance (hours)
Oxavail(t)          availability of Ox (hours)
income              the net income;

landbal(l)..        sum(c, xcrop(c)*landreq(l,c)) =l= land(l);
labavail(t)..       sum(c, xcrop(c)*labreq(t,c)) =l= labour;
Oxavail(t)..        sum(c, xcrop(c)*oxreq(t,c)) =l= ox;
income..            inc =e= sum(c, xcrop(c)*margin(c));

MODEL farm2    farm  model        / all / ;

SOLVE  farm2    USING LP MAXIMIZING           inc;
```

Exercise III - Fertilizer or manure

In this last exercise three farm activities can be distinguished: growing maize using manure, growing maize using fertilizer and growing wheat. Set c, consequently, is extended to these three activities. Further on, the tables will get an extra column.

```
SETS
c         crops             / maizem, maizef, wheat /
t         periods           / jan-feb, apr-may, sep-oct, nov-dec/;

TABLE
Labreq(t,c)         Labour requirements (mandays per ha)

          maizem  maizef   wheat
jan-feb   200.00  200.00   150.00
apr-may   180.00  200.00
sep-oct                    100.00
nov-dec   150.00  250.00   200.00 ;


TABLE
Oxreq(t,c)                    ox-plough requirement (hours per hectare)
```

```
            maizem  maizef   wheat
sep-oct                     56
nov-dec 56          56                  ;

PARAMETERS
yield(c)            crop yield (tons per hectare)                   / maizem = 1800, maizef = 2340, wheat = 1000 /
price(c)            crop product prices (Rp per kg)     / maizem = 3000, maizef = 3000, wheat = 5000 /
pcost(c)            production costs (Rp per ha)                    / maizem = 4000000, maizef = 5000000,
                                                                      wheat = 2500000 /
Margin(c)           net margin per hectare;

margin(c)=yield(c)*price(c)-pcost(c);

SCALARS
labour              family labour available     / 800 /
ox                  ox power available                  / 200 /
land                land availability           / 4 / ;

VARIABLES
xcrop(c)            the level at which each cropping activity is chosen
inc                 the net income;

POSITIVE VARIABLE xcrop;

EQUATIONS
landbal             land balance (hectares)
labavail(t)                 labour balance (hours)
Oxavail(t)                  availability of Ox (hours)
income              the net income;

landbal..           sum(c, xcrop(c)) =l= land;
labavail(t)..       sum(c, xcrop(c)*labreq(t,c)) =l= labour;
Oxavail(t)..        sum(c, xcrop(c)*oxreq(t,c)) =l= ox;
income..            inc =e= sum(c, xcrop(c)*margin(c));

MODEL farm3         farm  model       / all / ;

SOLVE  farm3        USING LP MAXIMIZING  inc;
```

Appendix 3: MGLP Solution to Case Study

A) The MGLP model

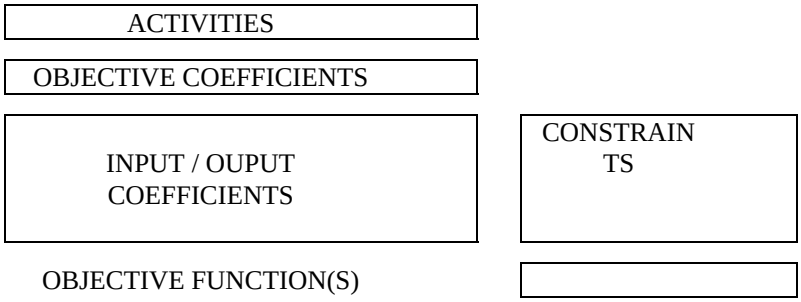
There are seven different land use types. In Zone I, rice, maize or cassava can be grouwn or the land can be used for a conservation area. In Zone II, rice production is not possible. Both zones face land constraints and, as labour is assumed perfectly mobile, the labour contraints for each month can be defined for the region as a whole. There is also a restriction on the capital used in the region of 38 million Rp.

The two policy views can be operationalized by formulating the following two objective functions:

Objective I Maximization of income

Objective II Maximization of natural area or minimization of agricultural area

In *Excel* the problem can easily be solved by using the spreadsheet format discussed before and schematically presented below.



The complete spreadsheet solved for the first objective function is shown below. The maximum income possible is 4.8 billion rupiah. The consequent area for conservation is then 148 ha. Of course, if the second objective is maximized, the area for conservation will be the total area (1,800 ha). The income earned from this land use type is 0.9 billion rupiah.

	A	B	C	D	E	F	G	H	I	J	K
1			Zone I				Zone II			LHS	RHS
2		**Land Use Type**	**Rice**	**Maize**	**Cassava**	**Nature**	**Maize**	**Cassava**	**Nature**		
3		Area (ha)	215,0	546,8	238,2	0,0	175,4	476,7	147,8		
4											
5		Price ('000 Rp/kg)	1	0,65	0,25		0,65	0,25			
6		Yield (kg/ha)	6000	4200	9000		3900	8500			
7		**Gross Margin ('000 Rp/ha)**	6000	2730	2250	500	2535	2125	500		
8											
9		**Constraints:**									
10		**A. Land (ha)**									
11		1. Area suitable for rice in zone I	1							215,0	500
12		2. Area suitable for maize in zone I		1						546,8	700
13		3. Area suitable for cassava in zone I			1					238,2	800
14		4. Total area in zone I	1	1	1	1				1000	1000
15		5. Area suitable for maize in zone II					1			175,4	480
16		6. Area suitable for cassava in zone II						1		476,7	560

17	7. Total area in zone II					1	1	1	800	800
18	**B. Labour (mandays/month)**									
19	8. Labour in January						17		8104,1	31250
20	9. Labour in February						2		953,4	31250
21	10. Labour in March			18,5		10,6	3		7696,9	31250
22	11. Labour in April			2		4,3	17,5		9573,3	31250
23	12. Labour in May			3		6,5	3,5		3523,5	31250
24	13. Labour in June	22,9	10,8	18,7		44	17,3		31250	31250
25	14. Labour in July	122,3	4,5	4		2	2,5		31250	31250
26	15. Labour in August	35,9	5	18		25,5	1		19690,7	31250
27	16. Labour in September	8,4	44,2	3		26			31250	31250
28	17. Labour in October	6,4	2,5	1					2981,2	31250
29	18. Labour in November	43,9	26,8						24092,1	31250
30	19. Labour in December	20,4	27						19149,1	31250
31	**C. Costs ('000 Rp/ha):**									
32	Equipment	7	4	4,5		4	4,5			
33	Fertilizer	31	18	15		18	15			
34	20. Total costs	38	22	19,5		22	19,5		38000	38000
35	**D. Non-negative:**	1							215,0	0
36			1						546,8	0
37				1					238,2	0
38					1				0,0	0
39						1			175,4	0
40							1		476,7	0
41								1	147,8	0
42										
43	**OBJECTIVE I - MAXIMIZATION OF INCOME**				*(Specify a minimum acceptable income level and add as a*				**4812372**	…
44					*constraint.)*					
45	**OBJECTIVE II - MAXIMIZATION OF NATURAL AREA**				*(Specify a minimum acceptable area for conservation and add as*				**147,8**	…
46					*a constraint.)*					

In GAMS the set up of the MGLP model is presented in the box below. Four sets are used, one for the different land use types, one for the costs of equipment and fertilizer, one expressing the months of the years and the last one expressing the two zones. For data entry use has been made of scalar, parameter and table statements. Also, direct assignments are used to calculate the maximum land available for each land use activity and the gross revenue per hectare. The tables for land and labour requirement have three dimensions: land use activity, zones and months. Two objective variables are declared, one for the maximization of income and the other for the maximization of conservation area. The model is solved twice. First income is maximized and in the second solve statement the conservation area is maximized.

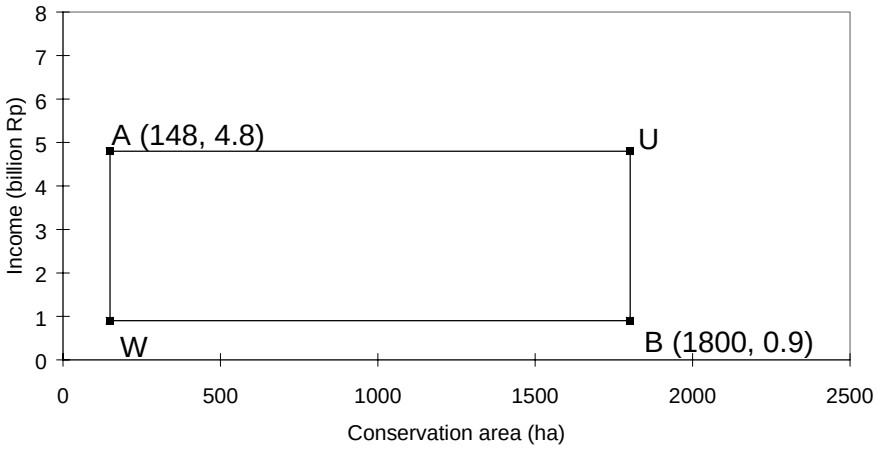
```
SETS
c          land use activity   /rice, maize, cassava, nature/
i          costs               /equipment, fertilizer/
z          zones               /zone-I, zone-II/
m          months              /Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec/;

SCALARS
MAXLAB                labour available   /31250/
MAXCAP                capital available  /38000/;
```

```
PARAMETERS
P(c)            price           /rice 1, maize 0.65, cassava 0.25, nature 1/
lAND(z)         land availability  /zone-I  1000, zone-II 800/
MAXLAND(c,z)    max land available for different crops in the two zones
GREV(c,z)       gross revenue ;

TABLE
Y(z,c)          yield
            rice    maize   cassava  nature
zone-I      6000    4200    9000     500
zone-II             3900    8500     500 ;

GREV(c,z)=P(c)*Y(z,c);

TABLE
LS(c,z)         land suitability
            zone-I  zone-II
rice        .5
maize       .7      .6
cassava     .8      .7
nature      1.      1. ;

MAXLAND(c,z)=LS(c,z)*LAND(z);

TABLE
LABREQ(c,z,m)   labour requirements
                    Jan     Feb     Mar     Apr     May     Jun     Jul     Aug     Sep
rice.zone-I                                                 22.9    122.3   35.9    8.4
maize.zone-I                                                10.8    4.5     5.0     44.2
cassava.zone-I                      18.5    2.      3.      18.7    4.      18.     3.
maize.zone-II                       10.6    4.3     6.5     44.     2.      25.5    26
cassava.zone-II     17.     2.      3.      17.5    3.5     17.3    2.5     1.

+                   Oct     Nov     Dec
rice.zone-I         6.4     43.9    20.4
maize.zone-I        2.5     26.8    27.
cassava.zone-I      3.      1.              ;

TABLE
LANDREQ(c,z,m)      land requirements
                    Jan     Feb     Mar     Apr     May     Jun     Jul     Aug     Sep
rice.zone-I                                                 1.      1.      1.      1.
maize.zone-I                                                1.      1.      1.      1.
cassava.zone-I                      1.      1.      1.      1.      1.      1.      1.
nature.zone-I       1.      1.      1.      1.      1.      1.      1.      1.      1.
maize.zone-II                       1.      1.      1.      1.      1.      1.      1.
cassava.zone-II     1.      1.      1.      1.      1.      1.      1.      1.
nature.zone-II      1.      1.      1.      1.      1.      1.      1.      1.      1.

+                   Oct     Nov     Dec
rice.zone-I         1.      1.      1.
maize.zone-I        1.      1.      1.
cassava.zone-I      1.      1.
nature.zone-I       1.      1.      1.
nature.zone-II      1.      1.      1.      ;
```

```
TABLE
COST(i,c)                 equipment and fertilizer

                  rice    maize   cassava
equipment         7       4       4.5
fertilizer        31      18      15  ;

VARIABLES
XCROP(c,z)        cropping activities
N                 conservation area
INC               net income ;

POSITIVE VARIABLE    XCROP ;

EQUATIONS
lABOURBAL(m)      labour constraint for each month
LANDBAL(m,z)      land constraint in each zone for each month
CAPBAL                  capital constraint
OBJ_I             maximization of  income
OBJ_II            maximization of  conservation area ;

LABOURBAL(m)..    SUM ((c,z), LABREQ(c,z,m) * XCROP(c,z))          =L=    MAXLAB ;
LANDBAL(m,z)..    SUM (c, LANDREQ(c,z,m) * XCROP(c,z)) =L=    LAND(z) ;
CAPBAL..          SUM ((c,i,z), COST(i,c) * XCROP(c,z))            =L=    MAXCAP ;
OBJ_I..           SUM ((c,z), GREV(c,z) * XCROP(c,z)) -
                  SUM ((c,i,z), COST(i,c) * XCROP(c,z))            =E=    INC ;
OBJ_II..          SUM (z, XCROP("nature",z))                       =E=    N ;

XCROP.UP(c,z)=MAXLAND(c,z);

MODELMGLP   explorative land use study /ALL/ ;

SOLVE  MGLP   USING LP MAXIMIZING          INC ;

SOLVE  MGLP   USING LP MAXIMIZING          N ;
```

Because there are only two objectives, the playing field can be presented in a graph as shown in Figure 11.

Figure 11 The playing field.



The playing field shows the extreme values of both objective functions. The stakeholders do not have to accept any worse values. This means that a satisfying solution for all stakeholders has to be found somewhere in this feasible space.

B) Scenarios

Different scenarios can be developed depending on the weight given to each objective function. Here four scenarios are considered.

- Scenario I - To alleviate poverty, absolute priority will be given to generating income.
- Scenario II - The unique ecosystem is placed on the world heritage list. Absolute priority is given to conservation of the area for natural development.
- Scenario III - Priority is given to conservation of the natural area provided a minimum income of 3 billion rupiah.
- Scenario IV - In the last scenario, priority is given to generation of income. For the support of donor agencies, a minimum of 800 hectares of conservation area is needed.

To solve scenario III in *Excel*, the minimum income level of 3 billion rupiah must be specified in the RHS cell of objective function I (indicated by three dots). Next this objective function is added as a constraint and the model is solved by maximizing objective II. For scenario IV, the minimum conservation area of 800 hectares is specified in the RHS cell of the objective function II and added as a constraint. This time the model is solved by maximizing income.

Creating scenarios in GAMS is relatively easy. Upper and lower limits can be placed on variables using a suffix **.UP** for the upper boundary and **.LO** for the lower boundary. After the minimum or maximum levels for the variables are specified, the model is solved again. The next box shows how the four scenarios are generated in GAMS. The scenarios should be added to the model system after the model statement.

```
* SCENARIO I
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

SOLVE MGLP USING LP MAXIMIZING INC;

* SCENARIO II
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

SOLVE MGLP USING LP MAXIZING N;

* SCENARIO III ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

INC.LO = 3000000;

SOLVE MGLP USING LP MAXIZING N;

* SCENARIO IV ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

N.LO = 800 ;

SOLVE MGLP USING LP MAXIZING INC;
```

C) Results

The first two scenarios have already been established in the zero round of the optimization. To develop Scenario III a minimum boundary of 3 billion rupiah is placed on the goal constraint income. To establish the maximum conservation area, a new optimization run is needed in which the level of income should be at least 3 billion rupiah. The outcome is an area of 1,206 hectares of which 406 hectares are in zone I and 800 hectares in zone II.

In scenario IV a minimum boundary is placed on the conservation area. The resulting maximum income which can be achieved is 3.8 billion rupiah. Figure 12 shows that both the outcomes of scenario III and IV fall within the established playing field. The results for each scenario for the different land use types are presented in Figure 13.

Table 12 Position of the scenarios in the playing field.

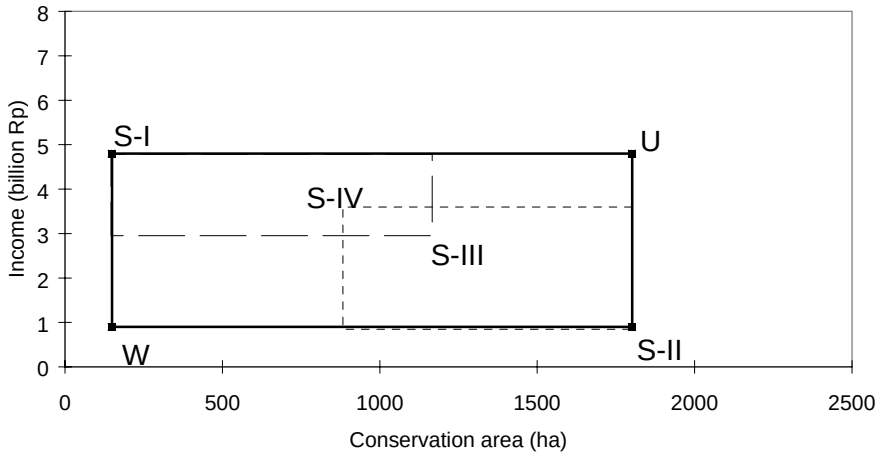
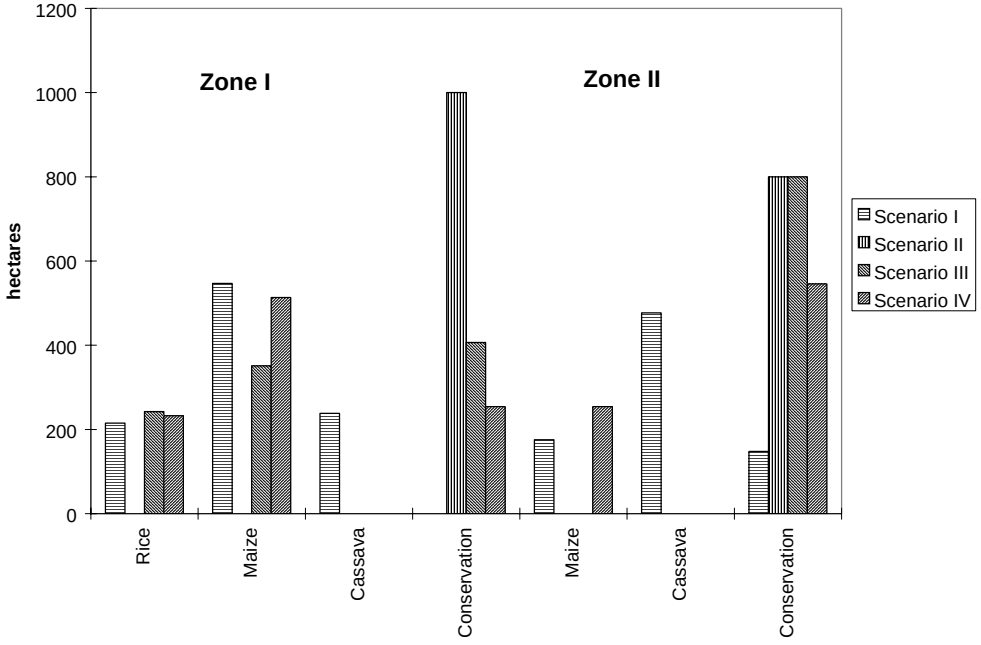


Figure 13 Land use types under different scenarios.



Depending on the priorities given to each objective function, various scenarios can be developed. The outcomes can then be compared and used to make the different policy views explicit.

Appendix 4: GAMS Functions and Dollar Control Options

Functions in GAMS

Name and arguments	Description
ABS(x)	Absolute value of x
CEIL(x)	Round x off to nearest higher integer
ERRORF(x)	Integral of standard normal distribution, from -∞ to x
EXP(x)	Exponential, e to the power x
FLOOR(x)	Round x off to nearest lower integer
LOG(x)	Natural logarithm of x
LOG10(x)	Common logarithm of x
MAX(x1, x2, x3,…, xn)	Largest value of {x1, x2, x3…,xn}
MIN(x1, x2, x3,…, xn)	Smallest value of {x1, x2, x3,…, xn}
NORMAL(x,y)	Normal random, random number normally distributed with mean x and standard deviation y.
POWER(x,y)	x to the power y, y has to be an integer
ROUND(x,y)	x is rounded of to y decimals
SIGN(x)	The sign of x (-1, 0, +1)
SQR(x)	x times x
SQRT(x)	The square root of x
UNIFORM(x,y)	Uniform random, random number with uniform distribution between x and y

Dollar Control Options

Option	Description
DOUBLE	Double spaced listing follows
EJECT	Advance to the next page
HIDDEN <text>	Ignore <text> and do not list
LINES <integer>	Next <integer> lines have to fit on the page
OFFMARGIN - ONMARGIN	Off margin marking - On margin marking
OFFSYMLIST - ONSYMLIST	Off symbol listing - On symbol listing
OFFSYMXREF - ONSYMXREF	Off symbol cross reference listing - On symbol cross reference listing
OFFTEXT - ONTEXT	Off text mode - On text mode
OFFUPPER - ONUPPER	Off upper case printing - On upper case printing
STITLE <text>	Set subtitle and reset page
TITLE <text>	Set title, reset subtitle and page

Appendix 5: Solutions to GAMS Exercises

Exercise I - a farm model

Using the two sets, the problem can be modeled as follows:

```
SETS
J       Crops                   /Sorghum, Millet, Cassava/
RE      Resources               /Land, Labour, Ox-plough, Capital/;

PARAMETER
R(J)    Revenue                 /Sorghum 108.3, Millet 66.36, Cassava 127.58/
B(RE)   Resources available     /Land 12, Labour 80, Ox-plough 8, Capital 400/;

TABLE  A(RE,J)          Resource requirement
                        Sorghum         Millet  Cassava
Land                    1               1       1
Labour                  5               5       8
Ox-plough               1               1
Capital                 30              20      40  ;

VARIABLES
Z       Gross margin
X(J)    Number of hectares with crop J;

POSITIVE VARIABLES X ;

EQUATIONS
REVENUE                 Objective function
RESOURCES(RE)           Resource constraints ;

REVENUE..               SUM(J, X(J)*R(J))    =E= Z     ;
RESOURCES(RE)..         SUM(J, X(J)*A(RE,J)) =L= B(RE)  ;

MODEL           EX1  Solution first  exercise       /ALL/ ;
SOLVE EX1  USING LP MAXIMIZING  Z ;
```

The optimal solution is found with sorghum 1.33 hectares, millet at 4.0 hectares and cassava at 6.67 hectares. With this yield, the maximum attainable income is 1,260.37 dollars. The shadow prices for land, labour, ox-plough and capital are 34.23, 6.43, 41.9 and zero, respectively. Since all three crops are grown, the reduced costs are zero for all of them.

Exercise II - a simple dynamic crop-irrigation model

This dynamic problem can be modeled as follows over three seasons:

```
SET
t               Seasons         /S1*S3/;

SCALAR
ALPHA           discount rate   /.95/
X0              initial water level         /3./ ;

PARAMETER
P(t)            Revenue by season (dollars per ton)     /S1=50, S2=100, S3=150/
R (t)           Rainfall (cm)                           /S1=2, S2=1, S3=1/ ;
VARIABLES
Y(t)    Yield (hundred tons per 100 ha)
```

```
W(t)     Depth of water received by the crop in season t (cm)
U(t)     Height of water released from reservoir  (meters)
X(t)     Water level in reservoir  (meters)
Z        Present value of receipts from sale of the crops ;
POSITIVE VARIABLES Y, W, X, U;

EQUATIONS
OBJ                 Objective function
YIELD(t)            Calculation of the yield (hundred tons per 100 ha)
WATERBAL(t)         Water received by the crops (cm)
INIT                Initial water level in reservoir (m)
WATER(t)            Water level in second and third season (m)
OUT(t)              Water release constraint ;

OBJ..               SUM(t, (ALPHA**(ORD(t)-1)) * P(t) * Y(t)) =E= Z ;
YIELD(t)..          Y(t) =E= W(t) - 0.1 * POWER(W(t), 2) ;
WATERBAL(t)..       W(t) =E= U(t) + R (t) ;
INIT..              X("S1") =E= X0 ;
WATER(t+1)..        X(t+1) =E= MIN ((X(t)-U(t)+RAIN(t)), X0);
OUT(t)..            U(t) =L= X(t) ;

MODEL       IRR     dynamic crop-irrigation model    /ALL/ ;

X.UP(t)=X0 ;

SOLVE       IRR     USING DNLP MAXIMIZING    Z ;
```

The optimal solution for each season is as follows:.

Season	Dam water level (m)	Water released (m)	Water received by each crop (cm)	Return by season ($000)	Discounted total return ($000)
1	3.00	2.00	4.00	12.000	60.661
2	3.00	1.65	2.65	19.475	48.661
3	2.35	2.35	3.35	33.419	30.160

The three crops receive a total of 10 units of water. Six are supplied from the dam and four are received as rainfall. At the beginning of the third season, the dam is completely emptied. However, rainfall leaves the dam at the end of the third season with a water level of 1 meter. The discounted total return is 60,661 dollars (from Kennedy 1986).

Exercise III - a Pakistani farmer

A) Basic model

The remaining four equations are defined as follows:

```
LABUSE(S)..   SUM(C, INPUT("LABOUR",S,C)*XCROP(C)) =L= MAXFLAB(S) + XLAB(S);
COSTLAB..     LCOST =E= LC * SUM(S, XLAB(S)) ;
TREV..        REVENUE =E= SUM(C, REV(C)*XCROP(C));
OBJ..         RETURN =E= REVENUE-LCOST;
```

The equation ‘labuse(s)..’ constrains the labour use in every season by the total labour available. The ‘costlab..’ and ‘trev..’ equations are accounting relations and are not really

necessary. The last equation expresses the objective function. The rest only add to the model and solve statement.

```
MODEL PAKISTAN  a Pakistani farm model  /ALL/;
SOLVE PAKISTAN USING LP MAXIMIZING RETURN;
```

The optimal solution is to plant all the land with sugarcane. The return is then Rs 25,487.50. In rabi season 245 man-days of labour are hired.

B) Adding two inputs: capital and water

Two scalars are added for the cost of purchasing water and the annual available capital.

```
SCALARS
WC          Cost of purchased water (Rs per inche)    /20/
MAXCAP      Maximum available capital (Rs)          / 20000 /;
```

The data on average rainfall per season is entered by adding a parameter.

```
PARAMETER
RAIN(S)     Water available at no cost (inches)         /kharif  700, rabi 80/;
```

The domain of the input table changes from S to T as for the annual capital requirements are accounted. The correct form of the input table is as follows:

```
TABLE INPUT(I, T, C) Input-output matrix for crops
                  wheat     rice      cotton    sugarcane
land.kharif       .1        1.        1.        1.
land.rabi         1.        .1        .4        1.
labour.kharif     7.        21.       11.       15.
labour.rabi       16.       9.        10.        62.
water.kharif                65.       25.
water.rabi        20.                 10.       45.
capital.annual    150.      180.      145.      500. ;
```

To account for the quantity of water purchased in each season and the capital used, two variables and three equations and added.

```
VARIABLES
WP(S)       Water purchased (inches)
WCOST       Cost of purchased water (Rs) ;

EQUATIONS
WATUSE(S)   Water used by season (inches)
CAPUSE      Capital required (Rs)
COSTWAT     Cost of purchased water (Rs)

WATUSE(S).. SUM(C, INPUT("WATER",S,C)*XCROP(C)) =L= RAIN(S) + WP (S) ;
CAPUSE..    SUM(C, INPUT("CAPITAL","ANNUAL",C)*XCROP(C)) + LCOST + WCOST =L=
            MAXCAP ;
COSTWAT..   WCOST =E= WC * SUM(S, WP (S)) ;
```

Of course, the cost of purchased water has to be extracted from the gross revenue to calculate the net return.

Incorporating these changes and additions leads to an optimal solution of growing 3.6 hectares of wheat, 5.4 hectares of cotton and 6.8 hectares of sugarcane. Rice is not grown. No seasonal labour is hired and in the rabi season an extra amount of 49.6 inches of water is purchased. The maximal attainable return is then Rs 23,322.77.

C) Possibility of livestock production

To feed livestock some of the land has to be allocated to the production of fodder. We therefore have to add fodder as a possible cropping activity. Furthermore, to produce the different crops draftpower is necessary as an input. The relevant part of the SET statement changes as follows:

```
SETS
C       Crops           /Wheat, Rice, Cotton, Sugarcane, Fodder/
I       Inputs          /Land, Labour, Capital, Water, Draftpower/
H       Livestock       /Buffalo, Cattle/
```

In the parameter statement the following additions have to be made.

```
PARAMETER
YIELD(C)        Yield per acre (maund per acre per year)
                                        /wheat 16, rice 11.4, cotton 10, sugarcane 375, fodder 1./
GREV(H)         Gross revenue from livestock    /buffalo 1750, cattle 1500/
DP(H)           Draftpower                      /buffalo 125/
```

Adding fodder as an additional crop to the input table and adding the draftpower requirements gives the following table statement.

```
TABLE INPUT(I, T, C)  Input-output matrix for crops
                      wheat   rice    cotton  sugarcane       fodder
land.kharif           .1      1.      1.      1.              1.
land.rabi             1.      .1      .4      1.              1.
labour.kharif         7.      21.     11.     15.             12.
labour.rabi           16.     9.      10.     62.             12.
water.kharif          65.     25.                     5.
water.rabi            20.             10.     45.             5.
capital.annual        150.    180.    145.    500.            120.
draftpower.kharif     12.     8.
draftpower.rabi 6.    1.      7.      30.             6.  ;
```

As we want to know the optimal production of buffalo and cattle we obviously have to include a decision variable.

```
VARIABLE
XLIVEST(H)      Livestock production ;
```

This variable is, of course, a positive variable.

In the equation statement the draftpower use and fodder use have to be added as a constraint. Furthermore, production of livestock requires labour and capital which should be reflected in the labour and capital constraints. Lastly, the revenue of livestock production is added to the total revenue accounting equation.

```
EQUATIONS
DRAFTUSE(S)     Draftpower constraint (work days)
FODUSE          Fodder needed (maund) ;
.
```

```
.
.
LABUSE(S)..     SUM(C, INPUT("LABOUR",S,C)*XCROP(C)) + SUM(H, LINPUT("LABOUR", S,
                H) *
                XLIVEST(H)) =L= MAXFLAB(S) + XLAB(S) ;
CAPUSE..        SUM(C, INPUT("CAPITAL","ANNUAL",C) * XCROP(C)) + SUM(H, LINPUT
                ("CAPITAL", "ANNUAL", H) * XLIVEST(H)) + LCOST + WCOST =L= MAXCR ;
DRAFTUSE(S)..   SUM(C, INPUT("DRAFTPOWER",S,C)*XCROP(C)) =L= SUM(H, DP(H) *
                XLIVEST(H)) ;
FODUSE..        SUM(H, LINPUT("FODDER", "ANNUAL", H)*XLIVEST(H)) =L=
                YIELD("FODDER")
                *XCROP("FODDER") ;
TREV..          REVENUE =E= SUM(C, REV(C)*XCROP(C)) + SUM(H, GREV(H)*XLIVEST(H)) ;
.
.
.
```

The optimal solution is to allocate 7.9 hectares of land to the production of sugarcane and 4.6 hectares to fodder production. The farmer will have 2 buffalo and 3 cattle. No water is purchased and during rabi season 144 man-days of labour are hired. This will lead to a return of Rs 24,589.44.