Part 4
Tools and data sources

You may be surprised that up to this point in the book there has not been much about data collection and statistical analysis. That changes now! But you will still find little in the way of mathematics or technical details. There are two reasons. First, many of the *ideas* you need to design and analyse good studies can be explained and understood without using mathematics. Secondly, there are many books around that describe the mathematics, and many of the courses in 'research methods', or 'statistics' that you will have followed will have used a mathematical, rather than an intuitive, approach. We want to provide an alternative.

These chapters can only be introductions to important ideas and methods. Maybe they will be all you need. It is more likely that they will raise all sorts of questions that are important in your research, and prompt you to seek out further understanding. They may even help you make sense of that statistics course you took and hated so!

The more technical aspects of a research project are important, and sadly many students have failed, or had to redo parts of their project, through failing to understand them early enough. If the material here raises any questions or uncertainties then you should get help. Biometricians and statisticians are experts in this stuff, so consult them! And any successful researcher must also have a sound grasp, and should be able to help you.

The topics covered in Part 4 are only a selection from those that could have been included. We have used two criteria to select them:

1. The topic is essential to most projects, yet commonly misunderstood. (Included here are the chapters on design of experiments, planning surveys, measurement, managing and analysing data).
2. The topic is important but often omitted from research methods courses. (We have included chapters on finding and using secondary and spatial data, and modelling).

The chapters are roughly arranged in the logical order in which they might occur in a project. But please don't wait until the end of your project to look at later chapters! The design of the study depends on how you will analyse it, so you must be aware of the later steps during early stages.

Richard Coe


TOOLS and DATA SOURCES
SURVEY
NIKON
ALEYA

4.1 Using secondary data sources

Jayne Stack

- **Primary data are observations you collect yourself to meet a specific research objective**
- **Secondary data are observations collected by others for other purposes**
- **Every research study should start with a review of relevant secondary data before planning collection of primary data**
- **Secondary data can help define the scope and extent of a problem**
- **The value of secondary data may be limited by availability, accuracy, timeliness and the definitions used**
- **There are very many sources of secondary data. International data are increasingly available free on the Internet**
- **Techniques are available to help you use secondary data effectively**

Overview

At the broadest level information sources that are available to you can be classified as primary or secondary. Primary data are those which you collect yourself for a specific research purpose. Secondary data is information that has been previously collected by individuals or agencies, usually for purposes other than your own particular research study. Secondary data may be qualitative or quantitative. Qualitative data is generally thought of as subjective, verbal and descriptive and includes information captured by a wide range of media. It includes photographs and maps, case studies, reported happenings, in-place observation, and tape or video recordings of conversations and/or activities. In contrast, quantitative data is generally numerical data, collected using some form of measurement and amenable to mathematical analysis. Quantitative data includes information captured by direct measurement through field observation (rainfall, temperature, crops yields, price-monitoring) and by direct measurement within structured questioning (household income and expenditure data collected by governments as part of national households surveys). Although the nature of secondary data influences the selection of tools that can be used to manage or interrogate a data set it is easy to exaggerate the differences. Both types of information are collated, sifted and organised into some sort of meaningful form by looking for connections or relationships in the data. The guidelines for reviewing secondary material outlined in this chapter can be adapted to a wide range of secondary sources compiled by other people.

The aims of this chapter are to:

- Show that no study should begin without first reviewing existing knowledge, regardless of the data-collection techniques to be followed later
- Identify the main sources of secondary information
- Point out and illustrate the need to critically examine the concepts and definitions used in secondary information
- Illustrate some of the conceptual and analytical tools that can be used to 'interrogate' data from secondary sources
- Secondary data analysis is not just a first-stage activity but can and should contribute to every stage of the research cycle.

Why is secondary data useful?

All too often inadequate attention is given to reviewing existing knowledge before embarking on primary data collection. No re-

search study should be undertaken and without a prior search of secondary sources (also termed desk research). There are several grounds that give us confidence in making such a bold statement. (The following material was adapted from Crawford and Wycoff, 1990):

- Secondary data helps you to: define a research problem, formulate research questions and hypotheses, and select a research design. The assembly and analysis of secondary data almost invariably makes an important contribution to the research process. A review of existing knowledge will improve your understanding of the research problem, including the key issues, core concepts, and on-going debates. It will reveal approaches to data collection (e.g., useful conceptual models, variables for concepts of interest, appropriate analysis techniques) that may improve or complement your own initial research design. In sifting purposefully through secondary data, you may find something else that sends you exploring new regions or ideas you may not even have thought of before. And, you might find evidence that will actually change the shape of your ideas.
- Secondary data may be sufficient to answer the research question. Occasionally you may find the available data are so adequate that primary data collection is unnecessary. If useful secondary data are available, they can be used to substitute for primary data collection at any stage during your research. It is not always necessary for you to collect all the information required for the analysis yourself. For example, daily rainfall records for the last 10 years obtained from the Meteorological Office allow you to draw conclusions about the adequacy of the growing season and the problem of dry spells, or agricultural data from a national sample survey can provide good information on the major characteristics of a farming system.
- Data costs are substantially lower for secondary data than for primary data. A thorough review of secondary sources can be completed at a fraction of the cost and time it takes to complete even a modest primary data collection exercise. Finding a 'ready made' solution in existing sources is unlikely, but even partial solutions help primary data collection needs, and therefore save time and money. For example, the current livestock situation in a country in terms of stocking densities, grazing pressure, herd structure, and management practices could be studied using a combination of secondary livestock data from the Ministry of Agriculture, the veterinary services, and reports of past research studies.
- Secondary sources of information can yield more accurate data than that obtained through primary data. This is not always true, but when a government or international agency has undertaken a large-scale survey or even a census, their results are likely to be far more accurate than your own surveys when these are based on relatively small sample sizes. For example, a national income and expenditure sample survey is likely to yield more accurate results than an income and expenditure study of 200 sample households in a single area. However, it should be remembered that all secondary data was once someone else's primary data. Some people who work with official statistics wrongly conclude that their own analysis is more objective than analyses of primary data, which is 'soft' data.
- Secondary sources help define the population. They can be extremely useful both in defining the population and in structuring the sample you wish to take. For instance, government statistics on a country's agriculture will help to stratify a sample and, once you have calculated your sample statistics, the stratified sample can be used to project those estimates from the sample to the population.

- Secondary data can be used to make comparisons. Within and between nations and societies, comparisons can enlarge the scope for generalisations and insights. Global and regional data sets (e.g., those of the Food and Agriculture Organization of the United Nations (FAO), World Resources Institute (WRI), or the World Bank) are a valuable source of secondary data for between-country comparisons on a vast range of topics including poverty issues, food security, trade patterns, growth rates, and technical change. Within-country comparisons can be made using national data sets disaggregated by administrative or natural regions.
- The availability of secondary data over time enables the employment of a longitudinal research design. One can find baseline measurements in studies made in the past and locate similar data collected more recently. With an increasing emphasis on understanding patterns of change, the use of secondary sources can also be critical to single point surveys, which lack a time dimension.
- Secondary data can be used to increase the credibility of research findings obtained from primary data. The comparative use of other research together with a comparison of data collected during your study with official statistics on the same topic can be very valuable when you reach the analysis stage. **Research results are more credible when supported by other studies.**

Limitations of secondary data sources

The following material was adapted from Crawford and Wycoff (1990). Whilst the benefits of secondary information can be considerable, like any other data collection method, the validity of the data must be carefully assessed. The main problems include:

- **Access.** Once potential sources of useful secondary data have been located there may be difficulties in accessing variables of interest if the data are not in the public domain or are unpublished. If this is the case, you will need to approach the organisation or individual holding the data to seek permission to use the information it contains. Unpublished data sets residing with government or non-governmental institutions are usually made available once permission has been sought in writing, clearly explaining the purpose for which the data will be used and the user's willingness to adhere to specific conditions of use. A supporting letter from the institution sponsoring your research may be helpful. Sometimes the original investigator will not make data available, particularly if they are still using it to pursue their own research. This can be frustrating, especially if data analysis is taking a long time. Sometimes researchers may be willing to provide some data in aggregate form ahead of publication. This can be used providing its source is acknowledged.
- **Relevance.** There is an inevitable gap between primary data collected personally by an investigator with specific research questions and hypotheses in mind and data collected by others for different purposes. It often happens that there is an abundance of secondary data, but much of it is not of direct relevance to your specific research problem. During the early stages of examining secondary data, you explore and gather anything and everything that you think might be of interest or use. However, as you begin to organise the material you have collected to support one or another of your ideas some secondary data will not be relevant. Every bit of evidence that you include must justify its existence; it can only do so in support of an idea. Use the list above to ask yourself on what grounds the secondary data you have collected is useful.
- **Reliability.** The reliability of secondary sources may vary substantially and it is difficult to ascertain if insufficient information is available about how the data were collected and

potential sources of bias and errors. It helps considerably if you are able to speak to individuals involved in the collection of the data to gain some guidance on the level of its accuracy and limitations.

- **Definitions.** A common problem in using secondary data is how various terms were defined by those responsible for its preparation. Terms such as family size, income, credit, farm size, output sales, and price need very careful handling. For example, a family size may refer to only the nuclear family or include the extended family. In census data a household is often a group of people who stayed the census night in the dwelling unit, irrespective of whether they are part of the nuclear family or not. Income data often exclude the value of own-produced goods. Credit and sales statistics often ignore transactions that pass through the informal sector. Even apparently simple terms like the year for which the data apply may need care in interpretation. For instance, in Zimbabwe, the marketing year 2002/2003 refers to the period 1 April 2002–31 March 2003. Any crop sales data recorded against 2002/2003 refers to sales from the 2002 harvest. Sales from the 2003 harvest are recorded under the 2003/2004 marketing year! Special care in interpreting definitions and years is necessary in combining secondary data from several sources to produce a derived data set.
- **Timescale.** Most secondary data has been collected in the past so it may be out-of-date when you want to use it. If the data source includes estimates of growth rates this information may be used to extrapolate figures for subsequent years. For example, population censuses usually include an estimate of population growth that can be used to estimate inter-census population data.
- **Source bias.** You should be aware of vested interests when you consult secondary sources. The objectivity of officials may be affected when it comes to reporting situations for which they themselves are partly responsible. Similarly respondents may provide biased information depending on their perceptions of the purpose of data collection (e.g., planning drought relief, forced destocking,). Further, official economic data may be a very inaccurate source of statistics in situations where the informal economy and/or black market account for a significant share of economic transactions.

Sources of secondary data

Secondary data sources can be divided into two categories, internal and external.

Internal information sources

All organisations collect a range of information during their daily operations. For instance, a marketing board records deliveries of crops, payments made to farmers, stocks, and orders from buyers that are dispatched, and invoices sent out. Such information may be available in a more disaggregated form than is reported in the organisation's internal reports. Much of this internal information is of potential use to researchers, but surprisingly little of it is actually used.

You may be unaware of some of the data collected and the regular reports submitted by the organisation for which you work. Begin your secondary data search with an internal audit. Familiarise yourself with available internal information whether you are a researcher in a government body, non-governmental organisation (NGO), or a business organisation.

External information sources

The primary sources of official and semi-official statistics are:

- Government statistics. These may include population censuses, national income data, agricultural statistics, poverty surveys, trade data, cost of living surveys, nutritional surveys, the results of commissions of enquiry into particular issues (e.g., land tenure) and possibly data on market prices.

Secondary sources can include:

- Marketing boards, which are likely to have information on quantities purchased of different commodities, imports and exports, buying and selling prices, and stocks
- Extension organisations who will have crop area and production estimates for various crops and probably farm budget data for different enterprises
- Agricultural research institutes that are an important source of information on such agronomic issues as soil fertility studies, crop and livestock breeding programmes and technology
- Veterinary departments who may have data on livestock numbers and disease control measures, e.g., dip tank records
- Hospitals and clinics might have data on incidence of malnutrition, particular diseases and causes of death
- Local administration offices often have lists of households which could be useful in the construction of sampling frames. They might also provide information on project activities in the district, e.g., active NGOs, or registered cooperatives
- Archives are a useful source of information to help you understand patterns of change
- International organisations may have country studies available at their local information centres or offices
- Websites. With the rapid development of information technology and computerised databases, the scope for you to carry out a search of secondary sources and to use secondary data sets compiled by other organisations and posted on websites, has increased dramatically. The following is a selection of key websites providing access to statistical data of particular interest to African agricultural, environmental and rural development researchers.

Main collections of wide-ranging development statistics

- World Bank website offers on-line access to country statistics and prepared tables for 207 countries and 18 country groups. 54 time series indicators on people, economy, environment, spanning 5 years are available and you can choose several ways of displaying the data: index, percentage change and graphs and can export the results to other documents. http://www.worldbank.org/data/
- UNDP Human Development Report and Indicators provides statistics on human development indicators including poverty, health, education, food security, employment, urban development, population, environmental degradation and national income accounts. http://www.hdr.undp.org/
- International Development Statistics (OECD/DAC) on-line database covering debt and aid. http://www.oecd.org/department/0,2668.en_2649_34447_1_1_1_1,00.html
- United Nations Statistical Organization (UNSO) has a wide range of statistical databases on-line on trade, national account, demography, population, gender, industry, energy, environment, human settlements and disability. http://www.un.org/Depts/unsd/stadiv.htm
- World Health Organization (WHO) website for health related statistical information. Also provides supporting materials including analysis software. http://www.who.int/whosis/

- Millennium Development Goal Indicators has 48 social and economic indicators for 1985–2000 used to monitor the implementation of the goals and targets of the United Nations Millennium Declaration. http://milleniumindicators.un.org/

Subject-focused development statistics websites

There are numerous subject-focused websites that provide data on specific topics (see http://www.eldis.org/statistics/index.htm). Those of particular interest include:

- United States Department of Agriculture (USDA) provides global and US agriculture data. http://www.ers.usda.gov
- Food and Nutrition and Crop Forecast. 'Food Outlook' is a report produced by FAO five times a year. It provides a global perspective on the production, stocks and trade of cereals and other basic commodities. Food Outlook can be downloaded from http://www.fao.org/giews/english/fo/fotoc.htm
- A helpful guide to other sources of food security statistics is the ELDIS Food Security Resource Guide at http://www.eldis.org/food/statistics.htm.
- HIV/AIDS. A resource guide for information and data is available at http://www.eldis.org/hivaids/aidsstats.htm.

Country-focused development statistics websites

A good starting point is the ELDIS country profile service. http://www:eldis.org/statistics/

Non-official sources

- Consultants reports (which may be gathering dust on the shelves of the body sponsoring the research!)
- Records of NGO activities including drought relief and supplementary feedings schemes
- Baseline surveys and project documents.

As you can see, secondary information can come from a bewilderingly large number of sources. Perhaps the most efficient and effective way to begin is to talk to people. Find the authorities in the field; search out the researchers working in your areas of interest. Conversations with them can get you further faster than almost any other search method. Researchers outside your own country can usually be contacted by e-mail and many are happy to forward copies of their own publications. Develop a network of contacts in key positions and cultivate them over time. Such contacts are particularly useful sources of semi-official and unpublished reports from research institutions and universities. In addition, experienced researchers have usually built up their own list of favourite websites that provide material on key research themes in development.

Recording details of secondary data material

Mountains of information can grow alarmingly quickly and it is imperative that you keep a record of the material that you have consulted in the course of your research so that you can acknowledge all the sources. The most important aspect of collating secondary data is to establish a 'trail' so that you or anyone who wishes to check your sources can easily find them again. Note the source of every piece of information you find useful. There is no single universally accepted format for referencing but a common order for the required information is:

- Author(s)
- Date of publication
- Title of the work cited

- Publisher
- Place of publication.

Evaluating secondary information

Information obtained from secondary sources is not equally reliable or equally useful. As mentioned earlier, just because data is published it does not mean that it is accurate. Just because data is available it does not mean it is useful for your particular study. If you are using secondary data, be it quantitative or qualitative, you should routinely ask the following questions, according to Dillon *et al.* (1990).

- What was the purpose of the study? Data are usually collected for some specific purpose, that ultimately determines the study variables of main importance, the reporting domains and the degree of precision
- Who collected the information? Because you are not collecting the data yourself, a natural question concerns the expertise and credibility of the source. Find out how the data were obtained and what sort of training and expertise is present in the organisation providing the data
- What information was collected? It is important to check this exactly. For example, in a study on household income were all income sources included, or only cash income?
- When was the information collected? The time the data were collected plays a role in its interpretation. For example, information on the nature of the season should be examined when interpreting household information on incomes or food security
- Which geographical area does the data represent? Not all data is collected for the same spatial area. Administrative boundaries often differ from geographical boundaries and may also vary depending on the organisation collecting the data. Boundaries also change over time as new administrative districts are formed by splitting or amalgamating existing units. The boundary issue is generally most problematic if data from different secondary sources are being combined and/or information from various points in time is being compared.
- How was the information obtained? The method used to collect data is an essential ingredient in evaluating the quality of secondary information: for example, the size and nature of the target sample, whether it is based on observation or recall, how it is collected (key informant interviews, household surveys, focus group, satellite imagery, etc.) and if surveys were from single or multiple visits. Some methods or combinations of methods are better than others at providing specific types information. Familiarise yourself with the alternative ways of obtaining information so that you can make an informed assessment of secondary data quality.
- Is the information consistent with other information? A valuable principal in data collection is that of triangulation where information is collected from multiple sources. If similar conclusions can be drawn from different sources of data this lends credibility to the findings. If differences exist, you should try to find out why, and which source is more reliable. The consistency of information is frequently a problem with agricultural production statistics from different sources.

Working with secondary data

Research studies use secondary data in several ways, the following are three broad types:

1. Research which uses aggregated secondary data to inform a study that will generate its own primary data as a major source of information.

For example:

- Study of consumption and marketing decisions of smallholders where the major source of information is a household sample survey, but where secondary data on grain production and marketed surplus by region are combined with official population data to examine past trends in agricultural production and marketed output. Here the analysis of secondary data provides a context for the analysis of the primary data.
- Investigation into the feasibility of edible insect farming using an experimental farm, where secondary information on artificial feeding is used to identify alternative feeding methods for field trials.

2. Research which uses aggregated secondary data as a major source of information, when interpreting this information. For example:
   - International comparison of various development indicators using a World Bank's global data set.
   - Regional human poverty comparisons made by the United Nations Development Programme (UNDP) for Zimbabwe using a poverty assessment study survey undertaken the Ministry of Public Service, Labour and Social Welfare and other secondary data sets (UNDP, 1998).
3. Research which uses disaggregated secondary data, perhaps in raw form, as a major source of information, with a new analysis of the same data. For example:
   - Modelling agricultural supply response using a data set derived from secondary data found in official statistics
   - Construction of a food balance sheet using official statistics
   - Lenin's' famous analysis of peasant differentiation using Zemstov house-to-house census data as his major source of data (Lenin, 1961).

The conceptual and analytical tools used to interrogate secondary data will vary depending on the role that secondary data play in the study. For instance, if secondary data are the major source of data for your research task, the analytical process, (specification and estimation) is likely to be a central component of your thesis. On the other hand, if you are assembling secondary data to improve your understanding of the socio-economic conditions in a field study area you are more likely to use simple descriptive statistics to highlight important trends and characteristics.

Regardless of the way you intend to use secondary data some general comments can be made about methods of interrogating it.

General

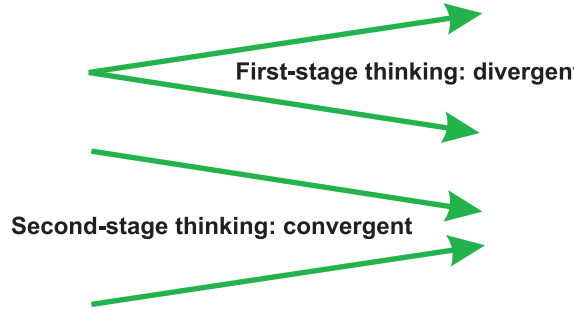
Research, like any other types of thinking, can be thought of as involving two stages. [The distinction between first stage and second stage thinking was first brought to my attention in a highly recommended course called 'Writing for effective change' distributed by an NGO called Fahamu, Learning for Change (Fahamu)]:

- First-stage thinking: exploration, discovery, generating ideas
- Second-stage thinking: collating, sifting, organising the ideas into a robust structure

First-stage thinking. Sometimes called 'divergent' or 'radiant' thinking; during this stage, you explore and gather anything and everything that you think might be of interest or use to your study.

Second-stage thinking. By contrast this is sometimes called 'convergent' or 'focused' thinking. It organises the material you have collected to support one or another of your ideas.

We tend to be much better at second-stage thinking than at first-stage thinking. So much so that we often fail to see first-stage thinking as thinking at all, but we ignore it at our peril. No amount of excellent second-stage thinking can compensate for poor or inadequate ideas. You must spend time generating ideas from secondary information **before** trying to assemble them into a structure.



Two techniques can help you:

- Mindmaps are powerful devices in first-stage thinking, they will help you gather and initially sort ideas
- Grouping and summarising is a second-stage thinking technique that helps you to organise your ideas.

Mindmapping

Mindmapping has been around for a long time, but the person who has done most to explain it and make it popular is Tony Buzan (1993). Mindmapping exploits our mind's extraordinary ability to create meaningful connections between ideas. Mindmapping helps us to see – or make – connections in our thinking, increasing our creativity and making thinking more efficient.

Brainstorming is the first step in mind mapping. Figures 1–3 show how a mindmap was developed to think about how to improve feeding systems using traditional practices (The mindmap example is adapted from 'Writing for effective change' distributed by an NGO called Fahamu, Learning for Change). Begin by writing the main research question or concept in a circle in the centre of a page. Then, jot down any ideas that come to mind when you think of this concept. (Figure 1). As you think of each new idea, new branches are created from the central balloon and the idea is written along the line (Figure 2). The next step is 'free associating' on each idea to build a verbal map of words or images that are connected to it. Sub-sets of

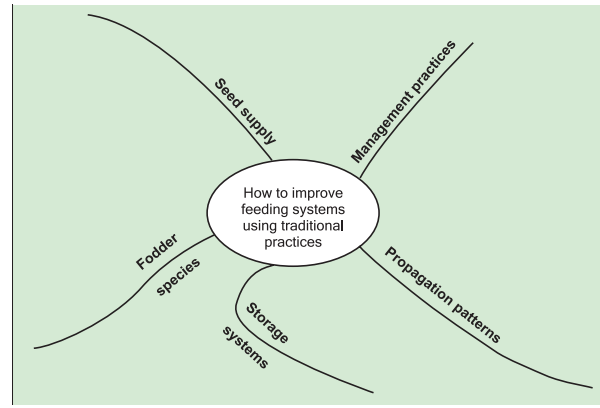


Figure 1. Example mindmap (4)

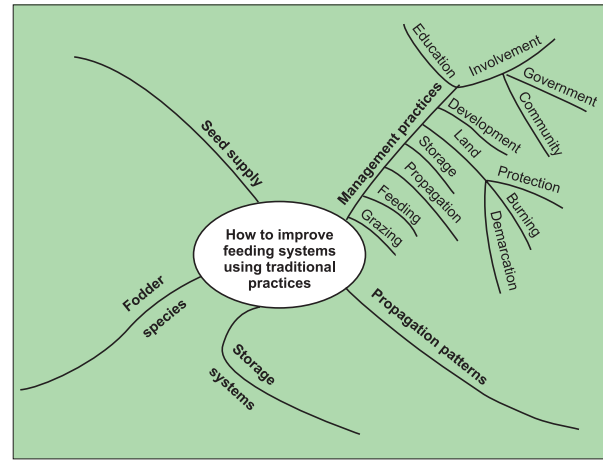


Figure 2. Example mindmap (5)

ideas are drawn as twigs from the appropriate branch line (Figure 3). Gradually a verbal mindmap tree of associations is built up. The final stage is to introduce hierarchies and categories to order or structure your mindmap. In the final mindmap you could use colour to emphasise hierarchies of ideas. The five main ideas for improving feeding systems radiating from the central image are shown in white whilst the sub-sets of ideas radiating from each of these ideas are black (Figure 3).

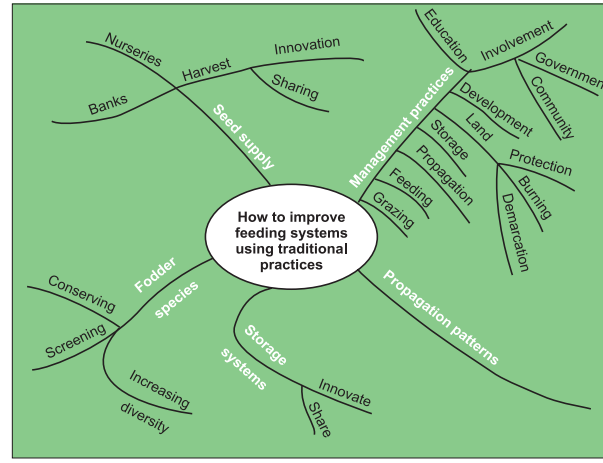


Figure 3. Example mindmap (final)

Grouping, summarising and organising ideas

It is unlikely that secondary data can be presented at the same level of detail in which it was previously collected. In other words you will not be able to report all the secondary material you have reviewed in its raw form and will need to summarise and present your data in a way that reveals patterns and trends in the data set. In order to summarise and present data it must be organised. There are various ways to do this. Some of the most common techniques that enable you to interrogate both qualitative and quantitative secondary data include:

Selecting categories in which the data can be summarised

A simple but effective way of revealing patterns in secondary information is to reorganise the information into new categories. The categories selected will depend on what is relevant to the topic being considered but if you are investigating food security you could use a data set of district-level information on estimates of per capita food availability to identify different categories of districts on the basis of their potential vulnerability to food insecurity. A study concerned with trade could rework information on the value of exports and imports between different countries by country groups to show the relative importance of different groups of trading partners. Putting available information into a tabular format is another way to organise either qualitative or quantitative information. Information from earlier studies was used in one study to compose a table showing different types of natural resource-access systems [individual, regulated common property, and unregulated common property (open access)] cross-tabulated with information about who controls access, who harvests, who benefits, who is included, and management implications. Organising the information in a tabular format highlighted the differences and similarities between each system more clearly than if the same information was just presented as paragraphs of written text (see Table 1).

Production of derived (secondary) data – new data sets

Very often the use of simple descriptive tools such as percentages, means, indices, or rates of growth can highlight patterns and trends in a data set that are not obvious in the original format of the data. For example, if crop production and sales data are available for two different farming sectors (smallholder and large-scale commercial) calculating the percentage share of each sector in total output and sales is a useful way of examining the relative

Table 1. Categorising and characterising different mopane woodland access systems in southern Africa

Land tenure institution	Who controls and how?	Who harvests?	Who benefits?	Who excluded?	Management implications
Individual	Individual	Individual	Individual	All others	Every owner has to exert effort to protect Cost of protection low if resource close to residence
Regulated common property	Rural Councils / Community resource management groups	Community members Licensed outsiders Community members Outsiders Licensed outsiders	Harvesting members Non-harvesting members (from licenses) Licensed harvesters	Unlicensed harvesters	Organised collective action to manage and protect resource. Potential economies of scale in protection activities. Transaction cost of formulating management rules and enforcing them may be too high for resource to be effectively regulated
Unregulated common property (open access)	Traditional control mechanisms		All harvesters	No one	If traditional regulations on extraction of resources break down, tragedy of commons results in overexploitation and deterioration over time
Centralised management of common property (e.g., State land)	Forestry Commission		Licensed harvesters State (from licenses)	Unlicensed harvesters	Organised centralised management system Potential economies of scale but cost of protection high for minor forest products Weaknesses in management may result in resources being ineffectively regulated

importance of each sector. If information on the volume, value and prices of exports of a particular commodity are available over a period of time, then calculating instability indices for each variable will demonstrate the level of export earnings instability and the extent this is due to either export price instability or instability in quantity exported.

Combining secondary data sources to form a derived secondary data set

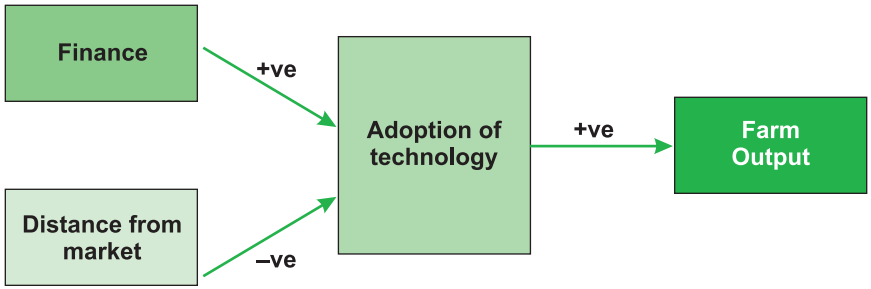
Combining secondary data sources may create a new data set that is more informative than each separate data set. A common calculation that illustrates this technique is the use of population data to express data sets of such variables as production, income, or cultivatable land in per capita terms.

Conceptual models and diagramming

Diagrams are very powerful tools for organising qualitative data. At its very simplest this could be diagramming a hypothesis about the main factors affecting a variable of interest using descriptive information from earlier studies (see **Chapter 4.3**).

Diagramming hypotheses (Dixon *et al.* 1995)

For example, in the following diagram, two concepts, finance and distance from market are hypothesised to be related as independent concepts to the dependent concept, adoption of technology. One of the independent concepts is seen to be positively related and the other negatively related to the dependent concept. Technology adoption is in turn hypothesised to positively affect farm output. Diagramming hypotheses promotes clear thinking and it is a useful way to summarise information from earlier studies. Use secondary data to diagram what you plan to study and even beyond the immediate research issue to show where your research fits in to the larger frame of reference.



However, diagrams can also be a useful way of conceptualising links and feedbacks within a system. For example, the livelihoods framework illustrated in Figure 4 is useful for thinking through livelihood circumstances of individuals, households, villages, and even communities and districts. The limitations of any such 2-dimensional representation of a process as complex as livelihood formation are recognised from the outset. The purpose of such a diagram is to organise ideas into manageable categories and identify the main components (assets, mediating processes, activities) and the critical links and dynamic processes between them (Ellis, 2000).

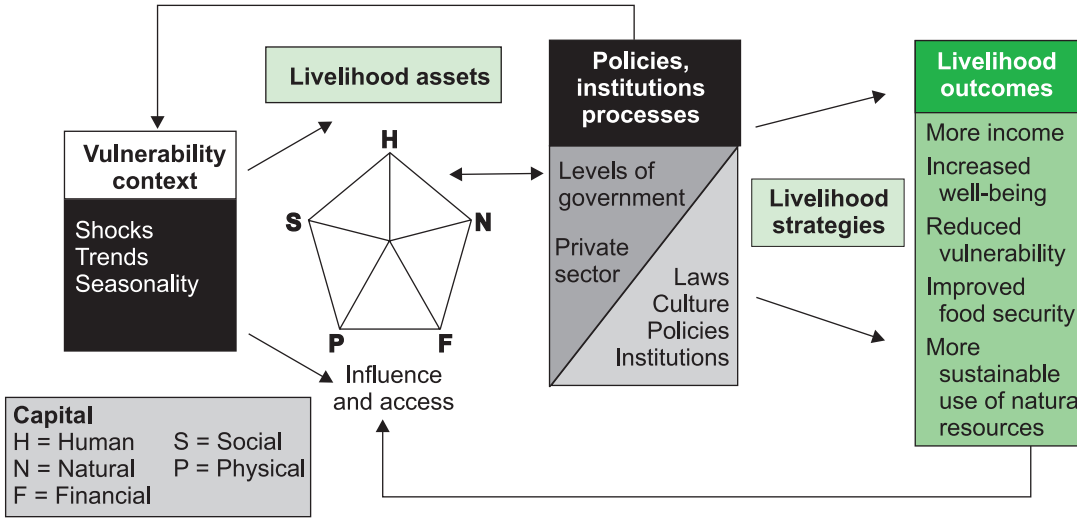


Figure 4. Sustainable livelihoods analytical framework

Analytical frameworks

Analytical frameworks are useful tools that enable you to concentrate on the broader picture. For example, food policy analysts may compute a food balance sheet using secondary information to examine food availability and identify key characteristics of domestic food consumption. Economic statisticians use accounting frameworks to prepare a country's national accounts and balance of payments based on secondary data.

Conclusions

A lot can be learned from secondary data and you should be prepared to explore various alternative ways of interrogating available information. Data sources should always be acknowledged and some guidance provided on the reliability and limitations of data used. In practice the collection and interrogation of secondary data is not just a first-stage activity but is something that can and should contribute to every stage of the research cycle. As noted in the opening section, secondary data can assist in designing a sampling frame, and in identifying a potentially useful method of analysis or appropriate conceptual framework. Secondary information provides a context for the analysis of primary data. The comparative use of secondary data can be especially valuable at the analysis stage and a good researcher will highlight areas of contrast and similarity between their own data and research findings of earlier studies on similar topics. Whilst findings gain more credibility if they are supported by a number of other studies, you should not be afraid to indicate where findings are different.

Resource material and references

Buzan, T. 1993. *The Mindmap Book*. BBC Books, London, UK.

Crawford, I.M. and Wycoff, J.J. 1990. *Market Research Teaching Notes*. FAO Report no. GCP/RAF/238/JPN (December). pp. 15–16; 16–17. Food and Agriculture Organization of the United Nations, Rome, Italy.

Dillon, W.R., Madden, T.J. and Firtle, N.H. 1990. *Marketing Research in a Marketing Environment*. pp. 98–100. Irwin, Homewood, Illinois, USA.

Dixon, B.R., Bouma, G.D. and Atkinson, J. 1995. *A Handbook of Social Science Research*. Oxford University Press Inc., New York, USA.

Ellis, F. 2000. *Rural Livelihoods and Diversity in Developing Countries*. Oxford University Press, Oxford, UK.

ELDIS guide. http://www.eldis.org/statistics/index.htm.

Fahamu. *Writing for Effective Change*. http://www.fahamu.org.uk

Lenin, V.I. 1961. *The Development of Capitalism in Russia*. 1908 from V.I. Lenin, Collected Works. Fourth English edition. Foreihn Languages Publishing House, Moscow, Russia.

UNDP. 1998. *Zimbabwe Human Development Report*. United Nations Development Programme, Rome, Italy.

Wye College. 1995. *Research Methods and Data Analysis*. Wye College External Programme, University of London, London, UK.

4.2 Spatial data and geographic information systems

Thomas Gumbricht

- **Understanding the spatial context of an agricultural and resource management problem will probably be an important part of solving it, so can not be ignored in your research**
- **Geographical information systems (GIS) allow you to manage and manipulate spatial data**
- **Simple manipulation of data sets that has already been prepared can be learned quickly. However, using data from multiple sources for more complex tasks can be a major undertaking**
- **Many basic spatial data sets are available for Africa but poor Internet connections may limit access to them**
- **Freely available software is now sophisticated enough to be useful in many spatial research projects**

Introduction

The Earth is a sphere with an average distance to the Sun of 150 million km. The Sun radiates energy, which is received by the rotating Earth in diurnal cycles with annual modulation as the Earth completes its annual ellipse. The energy that hence reaches the Earth is mainly dissipated at the Earth's surface. It rotates the hydrological cycle, releases nutrients that feeds the ecosystems, and drives photosynthesis (all which have been largely altered by man since the Industrial Revolution). Thanks to these processes life exists and the Earth's surface has developed a 'natural' logic. In dry areas with poor resources vegetation is sparse, in valleys where water and resources accumulate the vegetation is more luxurious. If there is a trough and enough water a body of water will form. In a similar way the human landscape is also logical, with fields in fertile valleys and dwellings along the ridges. Cities have to be close to large sources of water. These logical landscapes are also evident on a much smaller scale. Most vegetation is bound to specific habitats narrowly defined by conditions of climate, soil and water; that can shift within a scale as small as one metre. At an even finer scale a human thought is also dependent on energy dissipation at interfaces – in a very well described spatial context between the synapses of nerve cells. Image analysis and location information systems are hence very important tools in medicine, sociology, anthropology, biology, ecology, geology, hydrology and many other sciences.

For a particular study the spatial information needed might only be a map – as were the descriptive studies conducted by the first European explorers. In most instances a researcher is probably more interested in extracting more information in order to test a hypothesis. This could be comparing two district-level data sets, perhaps one on poverty and one on incidence of malaria. This is easily done in a geographic information system (GIS), and you still only need a single map, with attributes (databases) on both malaria and poverty. But malaria is a vector-borne disease, and the mosquito carrying the parasite breeds in water, so proximity to water is most probably important. To test that hypothesis an additional data layer of water availability is needed. This step is a major complication that has yet to be fully taken in the case of malaria. Rivers and lakes can easily be found, and their proximity to each population group calculated. But now you ideally also want population and malaria data on village level, not just for districts. Then you realise that mosquitoes can breed in water

tanks, small puddles, or even water trapped in an old bucket or boot. Now the comparison becomes almost impossible, and you need to get data on rainfall and temperature in order to calculate the daily water balance. This calculation is possible; the data are there (as you will see below), but the calculation is not a trivial task.

In general, a thematic study will need more refined data, whereas an interdisciplinary study must probably be satisfied with more generalised data. Often this is because detailed data of different origin are seldom compatible in their spatial resolution. However, the use of GIS and spatial data can be very rewarding. The first level, including a map, is almost always welcomed and very simple, it will only take a few days. The second level, comparing ('overlaying' in GIS jargon) attribute data related to the same spatial context is also quite simple, and will take a week to a month. The third level, analysing spatial relations introduces complexity, but can still be done by most standard GIS packages (and some of the freeware packages listed on page 142), but it will take some months to a year. The fourth level of integrating GIS with dynamic (time-resolved) models is quite complicated. This level will demand in-depth knowledge of both GIS and modelling, and most probably of programming as well. It will take longer than a year.

Mapping and modelling with GIS – A game of chess

Capturing the chess board

It is seldom convenient to have a sphere to portray the Earth's surface; so humans have used two-dimensional (2D) maps for at least 5000 years. Most GIS are also static 2D, even if the processes occurring on the Earth's surface are often 3D and dynamic. The most common GIS spatial data model represents space as 'vectors' (points, lines and polygons). This model is suitable for human-created objects and concepts (wells, roads, states, cities, rivers and lakes). Natural phenomena are better represented as continuous fields (elevation, land cover, vegetation density), which in GIS translate to a 'raster' or grid data model (Figure 1). For simplistic reasons, the raster model is often preferred in modelling. It is also the implicit format of satellite images or photographs.

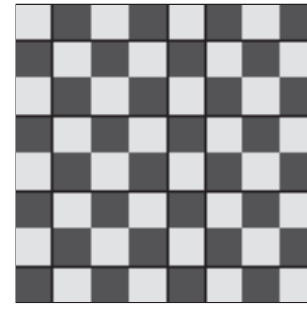


Figure 1. A raster

The raster landscape in Figure 1 is also the landscape where the game of chess takes place. Let us assume that you are unaware of the game of chess, but want to understand it by using GIS. Once you have identified the problem from a GIS perspective you must decide which data model (raster or vector) to use and how to *capture* the data. The chessboard can be captured as primary data (from satellite image or digital photograph) or from an existing analogue (secondary) source (digitising or scanning). Whatever you choose you will, implicitly or explicitly choose a certain grain size (or spatial resolution) when you capture the data (Figure 2). M*eta*-information on capture technique, resolution, and who did it should ideally always follow the GIS data, but is frequently lost on the way to the end-user.

For most spatial phenomena that are studied, you usually have a conceptual idea about the spatial patterns and dynamic processes that are occurring. If you assume that you already have some existing knowledge of chess, you can decide on a stratified sampling of data. For each square of unit distance, take one sample at a randomised point. You can further assume that there is neither an error in position, nor in the obtained value. A point is the simplest kind of vector data, and you can assign attributes to it – you can form a

Figure 2. Two raster data sets with different grain size. The ease with which you will be able to understand chess will obviously be dependent on the resolution or grain size used to capture the chessboard

database describing the properties of this point (in this case colour, but it could also be some other capacity such as depth, elevation, or type). Then you can *manipulate* the point data by *rasterising* it to arrive at something that looks like a chessboard. If you honour the value of the measurement in each cell (or *pic*ture *el*ement – pixel) you will arrive at the correct landscape (that you happen to know in this simple case) (Figure 3).

If instead you manipulate the data by using a geostatistical interpolation function, which do not honour the observed value *per se*, you get more or less erroneous results (Figure 4).

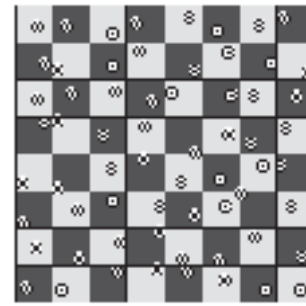


Figure 3. Sample points (vector data) and rasterised pattern

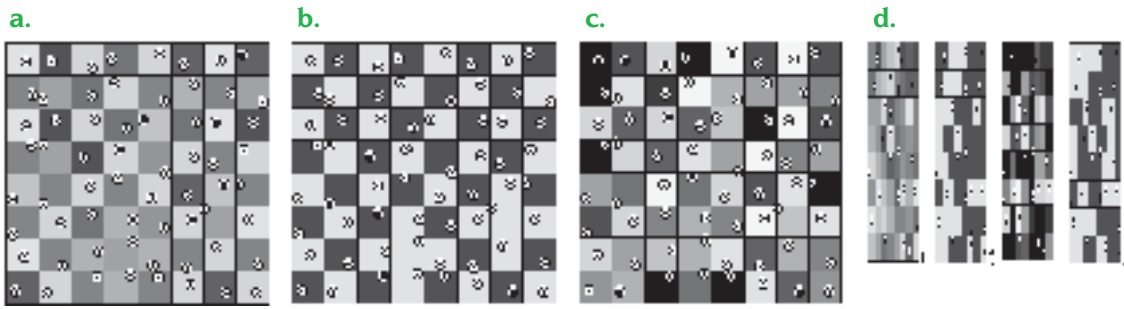


Figure 4. Interpolated 8 x 8 raster image from 64 Boolean sample points, randomly placed in each grid cell: a. Inverse distance weights (IDW) to 8 neighbours, b. Reclassification of a, c. Spline smoothing function to 8 neighbours, d. Reclassification of c. The reclassification is done as a threshold using the value 0.5. Both illustrated interpolation methods can be parameterised to get a true chessboard, that, however, demands iterations and skills, together with knowledge about the pattern of the generated surface

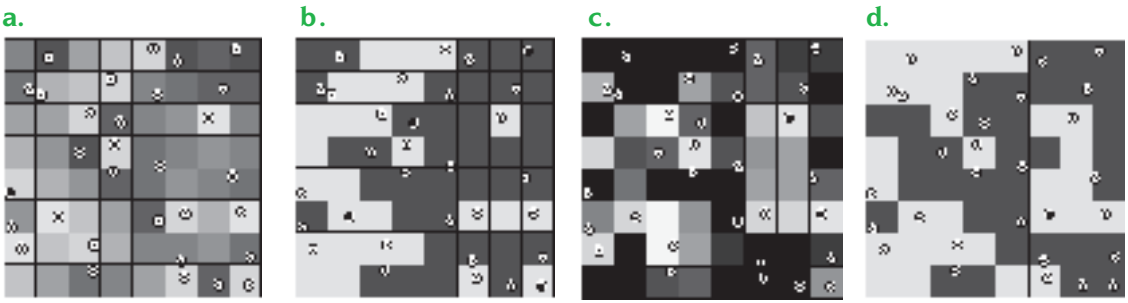


Figure 5. Interpolated 8 x 8 raster image from 31 randomly selected points (see Figure 4) a. IDW to 8 neighbours, b. Reclassification of a, c. Spline smoothing function to 8 neighbours, and d. Reclassification of c. The reclassification is done as a threshold using the value 0.5

The high cost of field and inventory work requires the fullest use of existing data and the application of interpolation methods. Hence, the sampling grid is generally much sparser than the interpolated grid (Figure 5).

Note that the interpolation of the chessboard data are truly 2D, whereas the Earth's surface is a spheroid and interpolation with different geoids and projections render different results. To choose the right projections for a particular purpose is not trivial, but is beyond the scope of this book).

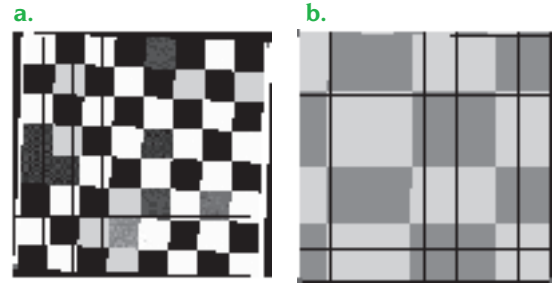


Figure 6. Schematic examples of problems with using remotely sensed data to portray the Earth's surface: a. Georeferencing and spectral properties of the observed phenomena, b. Grain size and geometrical distortions in the sensor

Primary data capture from remotely sensed imagery to GIS is an important part of the integration of GIS and modelling, also in social science for updating or downscaling census data (see below). Remotely sensed data have a definitive grain size and thus resolution. Apart from grain size, problems with sensor quality, spectral properties of the observed phenomena and georeferencing introduce errors when interpreting and classifying remotely sensed data (Figure 6).

Monitoring the dynamic game

Having established the chess-playing arena, a working hypothesis for the processes that are occurring needs to be formulated. For most dynamic phenomena an initial inductive approach is almost inevitable. Only after a set of observations is available is it possible to use coincidental data to formulate a deductive hypothesis.

Observations of natural and human phenomena are often made at regular intervals. Satellite images over an area are usually taken at the same time of day with a given interval (approximately 14 days for Landsat), as are many climate station data, water flow and water quality measurements. Measuring the chess game every morning at 09.30 (cheapest because it is outside the coffee room) always gives the same result (Figure 7).

But if you work late one evening and chance to look at the chessboard, and suddenly see

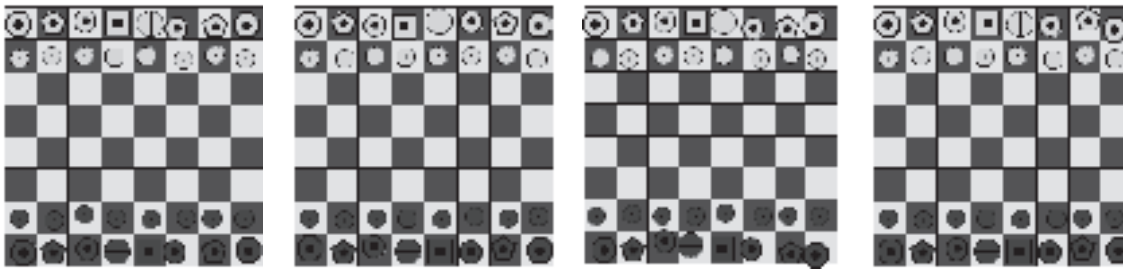


Figure 7. Observations of a chess game on four occasions. At first the game is apparently static. Only with a more detailed scrutiny it is revealed that the players actually are shifted a little between each observation. However, as we have no hypothesis or information of sub-cell pattern or process we neglect this as observation error

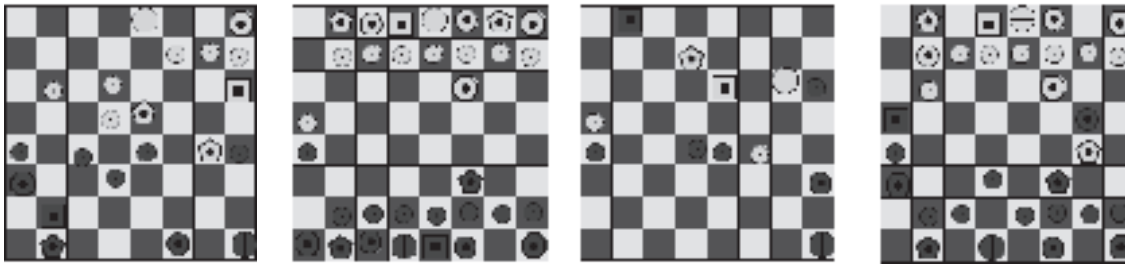


Figure 8. A series of temporally random observations of the chessboard

something has happened, you realise that there is obviously another time scale to the daily one (weekly, monthly or annual). So you start to observe the game regularly when you are working late. A rather erratic series of observations turns up (Figure 8).

Because of the strange observation angle (from above or 'nadir' in remote-sensing jargon) the visualisation of the players is poor, and it is difficult to distinguish the actors. However, a few hypotheses on their roles can be put forward:

1. One species (Bishops) seems to be bound to a certain feature type, or habitat, (namely black or white) in the playing ground.
2. The smallest and most common species (Pawns) seems only to be able to move in one direction like water downhill.
3. The species in the corners (Rooks) seem to be the most home-bound.

After some months of random observations hypotheses 1 and partially 2 are corroborated whereas 3 is falsified. After several years of fund-seeking the observations can be transformed into intense evening campaigns. With observations down to 10-minute intervals some of the players rules crystallise themselves, however the role of the knights escapes a robust formulation. Finally, a sensor connected to a real-time observation can capture the full sequence of activities, and the role of each player can be formulated.

Modelling the full game

The identified role of each player leads to a surge in modelling the game, mostly by using a rule-based (rather than statistical) approach where the roles of each player can be unambiguously defined. The formulation of initial (setting at start) and boundary conditions (edges of the playing arena) are straightforward. The application of an object-oriented approach for each player is favourable; a certain actor can only do a certain action, which cannot be done by another actor.

However, even though the game is spatially defined, it is not possible to use the toolbox of any commercial GIS to play the game. And, only a few softwares have architecture open enough to allow the GIS game to be programmed to them, but with great difficulties. With a customised GIS it is possible to create a graphical user interface (GUI) that can help to set up initial and boundary conditions, and even to allow the set-up of the players' positions in the middle of a game, and the use of that as an initial condition. This leads to the development of an intermediate coupling of the chessboard and the game simulator through their sharing a common file format. It is a bit cumbersome to use and never reaches widespread use to improve the social awareness of the game.

For the game itself, the combination of such advanced machine-learning as artificial neural networks and faster computers, mean more alternative game outcomes can be foreseen after each activity (draw). Finally, one computer (Deep Blue) succeeds in winning the game. This is

now more esoteric interest among the chess community, but the general public, policy- and decision-makers are unaware of this development.

Implications

A game of chess always aims at checkmate – which is unambiguously defined, as is the role of each player. The rules of the game show no evolution, neither in space, nor over time. If you change the extent of the arena, the role of the players or the outcome for checkmate to an unknown event, the computer would have little chance of winning. In a transient social or natural environment that is how the evolutionary game is played. In the simple case of chess there are only two scales that are of importance, that of a cell and the whole board. Furthermore, the game as such has no influence on the arena. In a landscape all discretised scales are arbitrarily chosen, the real landscape is a continuous nested hierarchy: but some scales have dominance-generating spatial architectures and temporal cycles, entrapped by key stone species and related processes. This also leads to the conclusion that the processes are forming the patterns rather than the other way around – and that the systems has feedback loops at various scales. All those aspects can be disregarded in the special (and simple) case of the chess game.

The general conclusion that can be drawn is that modelling in GIS is hampered by several shortcomings, that care must be exercised when using distributed data for modelling, and that the quality of many GIS integrated models is poor. They are also poor because they have poor GUIs, fail to visualise the results, and hence do not reach the intended user community. In order to secure high-quality GIS-integrated models the following issues need to be considered:

- Close co-operation between GIS model researchers in general, and particular among
  - researchers studying the same phenomena but adopting different methods and/or scales
  - researchers, planners and decision-makers
- Up- and down-scaling, and nesting models of different resolution
- Spatial and temporal domain, grain size and sampling intensity when integrating data from various sources
- Strategies for sampling spatial phenomena to get representative data
- Selection of spatial interpolation methods and spatially correlated error tracking and tagging
- Methods for evaluating the influence of error and error propagation on model performance, and error visualisation for communication information on uncertainty
- Integration of remote sensing into GIS models
- Integration of temporal processes into GIS (3D- and 4D-GIS)
- Integrated systems that support a complete digital data flow from data collection with mobile field GIS (Global Positioning Systems, GPS) to visualise and exchange results via networks
- Formulation of versatile criteria for evaluating the prediction power of GIS-related environmental models
- Compilation of high quality, accessible (shared) databases to be used as back-drops to evaluate the predictive power of different GIS-related environmental models
- Establishing baseline and framework data
- Development of guiding GUIs that can lead the user to select the best method for the formulated problem and the available data

- Development of friendly interfaces that promote the dissemination of GIS and integrated models to domain experts, planners and managers.

Using GIS in Africa

Studies involving spatial dependence and GIS in Africa are hampered by lack of data and computer resources, and poor knowledge and communications infrastructure. However, with the growth of geoinformatics over the Internet, global and continental-scale data are becoming increasingly available. Together with more powerful free GIS and remote-sensing software, there is a good chance that the data and software needed for many studies are available, either directly or via map algebraic modelling and other manipulations applied to available GIS data in combination with satellite imagery. The global trend in adopting remote-sensing data for spatial studies is strong in traditionally data-poor regions. Free high-resolution satellite images [Landsat Thematic Mapper (TM) and Enhanced TM (ETM)] are now available for the whole African continent. Access to this data in Africa, however, is often illusive due to poor Internet connections. The global data sets derived from satellite data (including land cover) are seldom adjusted for continental needs, leading to semantic discrepancies and interoperability problems when merging data sets. Local knowledge is mostly disregarded. Further, studies employing global data in Africa are often esoteric, and seldom used for policy or management inside Africa.

GIS and remotely sensed (RS) databases for Africa

In this chapter spatial databases have been divided into framework databases and field databases. **Framework databases** are base maps holding mostly information on anthropogenic-derived features – e.g., political boundaries and infrastructure, but they sometimes also have more object-oriented physical themes like elevation contours and hydrography. These databases are typically object-oriented and in vector format. They can be used to create simple thematic maps. Framework databases available for Africa typically contain data at district level, and hence simple descriptive statistical analyses (population density, travel distances, etc.) can be done at a level based on this data. Framework data can seldom be used directly for advanced analyses and modelling (environmental studies). Environmental studies demand field data, usually in raster format for such parameters as population density, soil classes, drainage, elevation, temperature, and precipitation.

Framework databases for Africa

The foremost baseline framework database for Africa (and other parts of the world) is the Digital Chart of the World (DCW). DCW is a 1:1 million scale thematic map developed by the Defense Mapping Agency (DMA) and compiled by Environmental Systems Research Institute, Inc. (ESRI). For large parts of Africa these base maps are the largest scale maps available, either due to lack of other data or to the larger-scale maps being classified. Themes in DCW include political boundaries, populated places, roads/railroads and other infrastructure, hypsometry, hydrographical data, and rudimentary land coverage.

Based on the DCW, ESRI has assembled a more easily accessible database and has also developed a more field-oriented World Thematic Database. Several other GIS software producers have also established databases based on DCW (and additional sources mentioned below) for bundled delivery. The most comprehensive probably being the Mud-Springs Geographers – AWhere Almanac Characterisation Tool (ACT). This and other software tools (listed on **page 142**) are a very good way to learn GIS using data over Africa. AWhere-ACT is

especially powerful for analysing climate data (supplied with the software) for agriculture and natural resource management applications. In many cases the software and bundled data are free for use in Africa by non-profit organisations.

Several recent efforts in creating more-detailed (large-scale) regional framework databases for Africa have been made. The most comprehensive is probably the Africover project by the Food and Agricultural Organization of the United Nations (FAO). This database also includes detailed land cover derived from combinations of Landsat ETM data and topographic maps. The agencies of the UN have also initiated an attempt to create a common depot for their GIS data – which has led to the Data Exchange Platform for the Horn of Africa (DEPHA) (see **page 140** for a more complete list of framework data sources available).

Field databases for Africa

Elevation

For Africa the elevation data in DCW (contour lines and spot elevation data) together with generalised 3-arcsecond digital terrain elevation data form the primary source for the global 30-arcsecond (approximately 1 km) GTOPO30 elevation database released by the United States Geological Survey (USGS) in 1996. The data in GTOPO30 have been hydrographically corrected and resampled to a 1-km grid, to create the HYDRO1k database. From the hydrologically corrected HYDRO1k Digital Elevation Model (DEM) seven derivative themes have been extracted: flow directions, flow accumulations, slope, aspect, compounded wetness indices, stream-lines and basin areas. Several individual countries have better elevation databases. The next elevation data set covering the whole of Africa will be the Shuttle Radar Topography Mission (SRTM) database (90 m resolution), expected to be released during 2004.

Land use/cover

Two global land cover data sets covering Africa in 1-km resolution are presently available. The latest is derived from TERRA-MODIS (Moderate Resolution Imaging Spectroradiometer) data (2000/2001) and was created by the University of Boston. MODIS has also been used to create a global tree cover database in 500-m resolution available from the University of Maryland. The older land cover is produced by the USGS from NOAA–AVHRR (Advanced Very High Resolution Radiometer) data (1992/1993). It exists in several versions useful for different applications and also includes monthly vegetation data from April 1992 to March 1993. The Africover database mentioned above is superior to these global databases but does not yet cover the whole continent.

Climate and vegetation

The United States Agency for International Development (USAID), as part of the Famine Early Warning System (FEWS), continuously provide 10-day composites of vegetation density (Normalised Difference Vegetation Index – NDVI) derived from NOAA–AVHRR in 8-km resolution covering the whole African continent. The data set goes back to 1981 and is archived and disseminated by the USGS. It can be retrieved from the African Data Dissemination Service (ADDS). Thermal Meteosat images together with 760 ground precipitation stations are used to estimate precipitation over Africa as part of USAID FEWS. Processing is based on 30-minute image intervals for cloud top temperature combined with the ground data and derived fields of humidity, winds and DEM. The data extends from 1995 and are archived and disseminated by USGS (ADDS webpage) as 10-day composites. More coarse resolution databases that cover climate together with scenarios of climate conditions under various assumptions of

human impacts on the climate are available from the Climate Research Unit (CRU), University of East Anglia, UK, either directly via the Internet, or from the Intergovernmental Panel on Climate Change (IPCC) as a CD.

Population

The best and latest population figures are the 1-km resolution Landscan project data for 2000, 2001 and 2002 from the Oak Ridge National Laboratory, USA. These figures are created from census data and downscaled using intelligent interpolation (using relations such as light at night, slope, or elevation, which correlates strongly with population density). The Center of International Earth Science Information Network (CIESIN) hosted by University of California, has compiled global population data for 1990 and 1995. The data has an original resolution of 5 arc-minutes (approximately 10 km), but for Africa the data mostly represent averages for larger regions. United Nations' African population figures for selected countries covering the second half of the 20th century are available from Central African Regional Program for the Environment (CARPE) (see **page 140**).

Soil map

FAO has produced a Digital Soil Map of the World (DSMW) in 1:5 million scale. Soil classes are given as polygons, with derived characteristics attributed. The soil map is only available as a CD. For some regions FAO also has a 1:1 million scale soil map.

Satellite imagery

Remote sensing (RS) data are increasingly important for creating and updating both physical/biological and socio-economic databases. Access to RS data is constantly improving thanks to: lowered prices, declassification of historical high-resolution data, a new generation of multi-sensor satellites (TERRA and ENVISAT) that are now operating, improved computing power and better software–user interfaces.

For national to continental studies NOAA-AVHRR and TERRA-MODIS data and their derivatives are the most easily accessible. Other data of similar resolution that can be easily accessed include the European Space Agency (ESA) ERS-2 satellite and its ATSR 7-band sensor (which can be downloaded from the Internet in near real time), and the SeaWiFS 6-band sensor.

Full coverage, high-resolution Landsat TM and ETM data are now also freely available for the whole of sub-Saharan Africa via the University of Maryland. Landsat E(TM) composites in Mr-SID compressed formats of the whole globe are more easy to download and available from NASA. To find all available Landsat MSS, TM and ETM scenes, and other satellite data sources use the NASA Earth Observing System Data Gateway.

The original TERRA–MODIS and NOAA-AVHRR scenes that were used for the land use/cover classifications (see above) are all freely available as composites from University of Maryland (TERRA–MODIS) and USGS (NOAA–AVHRR). The Africa NOAA–AVHRR tiles for vegetation are also available from the International Centre for Insect Physiology and Ecology (ICIPE). Additional, raw, NOAA–AVHRR data are available via the NOAA Satellite Active Archive on the Internet, or from USGS at the cost of reproduction.

Georeferenced time series point data for Africa

Time series point data on climate (weather station data) and hydrology are available via a variety of web pages. As this is outside the main scope of this chapter we recommend the ICIPE data server as a source of archived weather station data from around Africa.

Data accuracy and merging

Most of the older global and regional databases are not quality labelled, neither for positional error nor attribute accuracy, thus potentially leading to large problems in interpretations and applications. Even if most data that can be downloaded are georeferenced, their accuracy often does not allow mapping at higher resolution than 1:1 million. The dataset with highest spatial accuracy is the NASA geocover (downoadable as MrSID images – see **page 141**), which is within 50-m and can hence be used for geocorrecting other data sets. Another problem is that the semantics used, for instance, for the land use/cover maps are not coherent with those used in different parts of Africa. This is also due to a poorly developed unanimous semantic cover of natural geography in Africa. Semantic inconsistencies lead to information loss and prevent sound conclusions being drawn.

For most spatial studies, it is necessary to merge data. Most satellite images must be georeferenced to a projection that fits the geographic location (and the framework data) before they can be used for analysis and further studies. This is not a trivial task. Field data sets continuously vary over time and space on different scales. A satellite image is already an aggregation of the land surface over the pixel size. Field data, including socio-economic data are often up- or downscaled, or aggregated. The quality difference between data sets of the same origin but presented at different scales is seldom reflected in the metadata. It is extremely important to know the timing of acquisition, grain size and scaling of field data when analysing, interpreting, and applying such data. For most of the global data sets available this is seldom a problem. However, most local users are ignorant of the problem and secondary data sets derived from such sources often lack meta-data.

Points to remember

- Increased data availability and the ease with which distributed data layers are created from point and line data, and remote sensing, have led to a widespread coupling of GIS and remote sensing to existing (non-topological) cause-effect models in, for example, hydrology and erosion studies, and to updating and downscaling land cover and population density maps
- Data availability for Africa has now reached a point where it is possible to do such studies, often with freely available data
- The major problem for the individual researcher is in accessing the data, and in acquiring the skills of GIS and RS needed to 'massage' the data into a coherent database
- The free GIS software programmes available today are powerful enough for you to learn GIS, and to create basic databases
- The bottleneck for using GIS for research in Africa is poor Internet access and poor GIS skills
- If you want to use GIS you should download the necessary data or order it via CD/DVD (usually possible for a small fee), and you should learn GIS by using one of the listed free software programmes
- As most of the software programmes have very similar interfaces, learning one means that learning a second becomes an order of magnitude easier. So going from DIVA-GIS to ArcView is very simple (also because they share data formats).

Expert systems

GIS and RS (or geoinformatics) have developed from being tools for data storage and presentation to also include analyses and modelling. Overlaying two or more thematic maps (see Figure 9) is a simple but often illustrative means of identifying relations in spatial

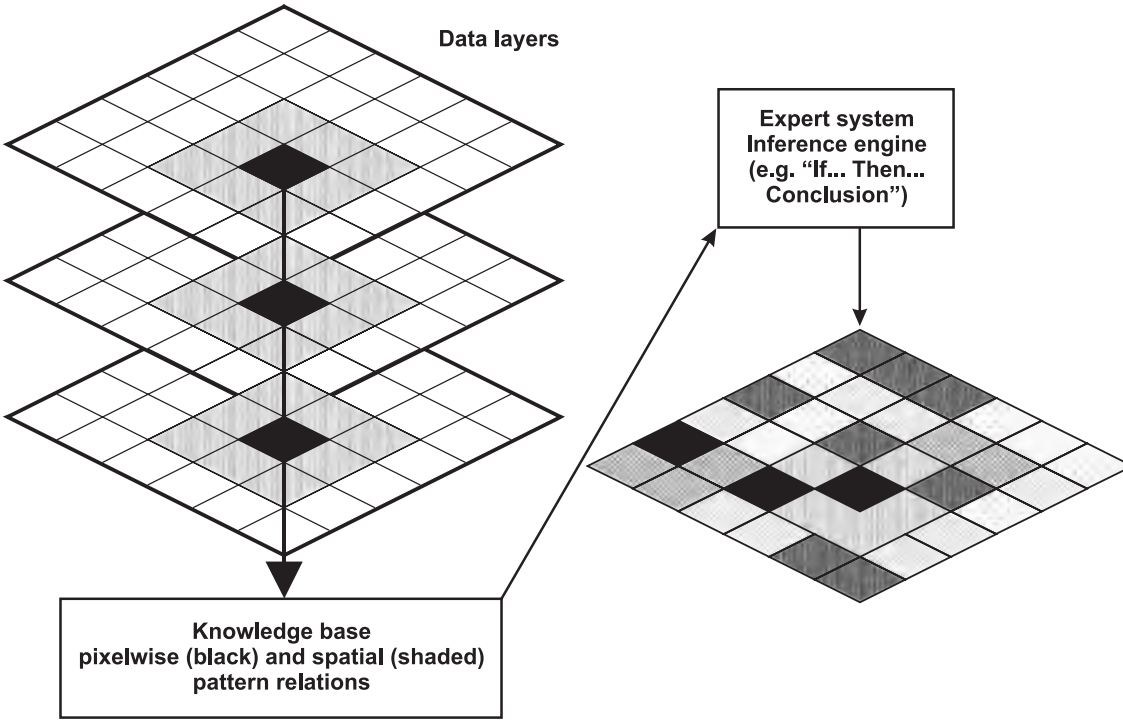


Figure 9. Schematic structure of an expert system approach for spatial data analysis

patterns. More advanced analyses include using map algebraic formulae combining several thematic layers. Such 'expert system' approaches are widely used to rank vulnerability of natural resources, food security, or water availability. One example is the DRASTIC method (Depth to groundwater, Recharge, Aquifer media, Soil, Topography, Impact of rootzone, Conductivity) for groundwater vulnerability analysis, where each of the seven factors has a physical related value. Development is towards more advanced expert systems including object-oriented methods, and considering ancillary and multi-temporal data, and spatial relations (Figure 9). Expert systems are like the game of chess – unambiguously defined with a set of strict rules. Expert systems are thus said to be data- or forward-driven. However, GIS is also becoming a decision-support system (DSS), e.g., for ill-structured (localisation) problems. Used as a DSS GIS becomes more of a tool for discussions and illustration of decision alternatives. Formal methods have been developed to involve various stakeholders in such discussions, including multi-criteria evaluation (MCE). In contrast to expert systems based on predefined rules and weights of physical parameters, DSS are related to different stakeholders perceptions, and as the aim is to reach a solution (for allocation of land use/development, water or nature protection), the method is said to be goal-driven.

Whether studying natural or social science, GIS can be very useful, and there is a plethora of methods, models and techniques that you can apply to analyse or present data that deals with spatial relations. But it is critical that you formulate a sound hypothesis and use adequate data of sufficient quality. To avoid mistakes a parsimonious approach, and rigourous meta-data description is essential. This will make it easy to update and eventually publish your data and results.

Resource material and references

Burrough, P. and McDonnell, R.A. 1998. *Principles of Geographical Information Systems*. Oxford University Press, Oxford, UK.

Fotheringham, S. and Wegener, M. 2000. *Spatial Models and* GIS. Taylor and Francis, London, UK.

Goodchild, M.F., Parks, B.O. and Steyart, L.T. 1993. *Environmental modeling with* GIS. Oxford University Press, New York, USA.

Goodchild, M.F., Steyart, L.T., Parks, B.O., Johnston, C., Maidment, D., Crane, M. and Glendinning, S. 1996. GIS *and environmental modeling: Progress and research issues*. GIS World books, Fort Collins, Colorado, USA.

Framework databases for Africa

Geography network
http://www.geographynetwork.com/

ESRI downloadable data
http://www.esri.com/data/download/index.html

Data Exchange Platform for the Horn of Africa: UN organisations Geo data depot.
www.depha.org

Digital Chart of the World (DCW): Basemaps for all the countries of the world.
http://www.maproom.psu.edu/dcw/

Food and Agriculture Organization of the United Nations (FAO)
Very good land cover maps over East and Central Africa
www.africover.org

Global GIS database – Digital Atlas of Africa
http://webgis.wr.usgs.gov/globalgis/

Field databases for Africa

GTOPO30 global topographic data
http://edcdaac.usgs.gov/gtopo30/gtopo30.html or
http://www.ngdc.noaa.gov/seg/topo/globe.shtml

Hydro1K HYDRO1k Elevation Derivative Database
http://edcdaac.usgs.gov/gtopo30/hydro/index.html

Landscan population data Oak Ridge National Laboratory
http://web.ornl.gov/sci/gist//landscan/index.html

Global landcover from NOAA-AVHRR (1992–1993 data)
http://edcdaac.usgs.gov/glcc/glcc.html

MODIS land cover from Boston University (2000–2001 data)
http://duckwater.bu.edu/lc/mod12q1.html

MODIS Global Vegetation Continuous Fields from 500m MODIS data 2000–2001
http://modis.umiacs.umd.edu/vcfdistribution.htm

CIESIN (Center for International Earth Science Information Network) Colombia University
Including climate data and global gridded population data from 1990 and 1995
www.ciesin.org

CARPE (Central African Regional Program for the Environment)
http://carpe.umd.edu/products/

Satellite imagery and related data

University of Maryland Global Land Cover Facility
A very good source of free Remote Sensing (Landsat (ETM) and TERRA MODIS) scenes
http://glcf.umiacs.umd.edu

USGS Land Processes Distributed Active Archive Center
http://edcdaac.usgs.gov/main.html

USGS Earth Explorer
http://earthexplorer.usgs.gov

USGS NOAA-AVHRR used to create global landcover – 93 original scenes (1992 to 1996)
http://edcdaac.usgs.gov/1KM/1kmhomepage.html

Africa Data Dissemination Service (United States Geological Survey – USGS)
http://edcw2ks21.cr.usgs.gov/adds/data.php

NASA Earth Observing System (EOS) Data and Information System
http://edc.usgs.gov/

NASA Earth Observing System Data Gateway.
http://edcimswww.cr.usgs.gov/pub/imswelcome/ or
http://redhook.gsfc.nasa.gov/~imswww/pub/imswelcome/

NASA global Hydrology and Climate Centre (Weather satellite data)
http://wwwghcc.msfc.nasa.gov/GOES

Goddard Institute for Space Studies
http://www.giss.nasa.gov/data/ and http://xtreme.gsfc.nasa.gov/

National Geophysical Data Center
http://www.ngdc.noaa.gov/

MrSID images (excellent geocorrected – can be used for georeferencing other spatial data)
https://zulu.ssc.nasa.gov/mrsid/

ICIPE (International Centre for Insect Physiology and Ecology) Africa Data Bank
(Including remote sensing data and weather station data Over Africa)
http://informatics.icipe.org/databank/

Japan Aerospace Exploration Agency
(Free JERS-1 radar images over most of Africa can be ordered on CD)
http://www.eorc.jaxa.jp/eorctop.htm

SRTM (Shuttle Radar Topography Mission)
http://www.jpl.nasa.gov/srtm/

Visible Earth – NASA site with preprocessed satellite images of Earth
http://visibleearth.nasa.gov/

Microsofts image database (terraserver)
http://terraserver.com/ or http://terraserver-usa.com/

Digital Globe (very high resolution data sets over selected cities)
http://archive.digitalglobe.com/

Geocommunity spatial news (incl Landsat viewer)
http://spatialnews.geocomm.com/

Free software sources

Dynamic Maps (A free ware GIS with many predefined functions for natural resource management)
http://www.skeinc.com/ or via www.africover.org

DIVA-GIS (A fully functional GIS developed by the International Potato Center)
http://diva-gis.org/

Arc-Explorer (Light-weight GIS by ESRI that also produced Arc-Info, Arc-View and Arc-GIS)
http://www.esri.com/software/arcexplorer/index.html

ERViewer (viewer for many image formats)
www.ermapper.com

WINDISP (A simple freeware for image processing from FAO)
http://www.fao.org/WAICENT/faoinfo/economic/giews/english/windisp/dl.htm

Mud Springs Geographers (A fully functional GIS bundled with free GIS data for Africa)
http://www.mudsprings.com/home.aspx

Mapmaker basic (Light-weight GIS freeware)
www.mapmaker.com

SILVICS (Satellite image processing for forests)
http://eurolandscape.jrc.it/forest/silvics/

GRASS (Advanced GIS and image analysis for UNIX or Linux)
http://grass.itc.it/

Microdem
http://www.usna.edu/Users/oceano/pguth/website/microdemdown.htm

4.3 Designing experiments

Richard Coe

- **Experiments are a central part of the scientific method because they allow you to test cause-effect hypotheses**
- **Many students learn about experiments in the context of studies of small field plots, but the key principles of experimental design are equally important in all studies**
- **All aspects of the design of an experiment depend on its objectives, so the objectives have to be carefully and thoroughly developed**
- **The details of the design of a good experiment will balance theoretical optimality with practicality**
- **Every experiment should have a written protocol that can be shared with others, so the design can be improved before the experiment starts**

Experimenting as part of research

Experimenting is a part of everyday life. In an informal way you experiment when you check whether your tea is too hot to drink, whether the bus is less crowded if you leave for work earlier or whether your supervisor approves of your style of writing. Within formal agricultural research, experimentation has long been the key tool. To many people 'agricultural research' is synonymous with field plot experiments. If you visit an agricultural research station one of the main things you can see are small plots for comparing different crop varieties or different management techniques. Much of the current theory and the methods for carrying out experiments were developed in the context of agricultural experiments, most notably by R.A. Fisher, a geneticist and statistician working at Rothamsted Experimental Station in UK.

Today field plot experiments on research stations are not the only avenue for agricultural research, but the ideas and methods of experimentation are still central to good research. Why?

Experimentation is concerned with the 'testing theory' step of research **(Chapter 4.2)**. Theories which help in problem solving often describe what will happen if a change is made:

- If we use this new variety of maize there will be less damage from stem borers
- If we substitute dairy meal with calliandra fodder milk production will not be reduced
- If we train farmers in pest management they will be able to grow cabbages more profitably
- If communities are better informed they will be more effective in managing common grazing.

Now in order to test your theory the obvious thing to do is to make the change and observe whether the predicted outcome occurs. This is the basis of experimentation, and the reason it is so important.

There are situations in which it is impractical or unethical to experiment, in these situations other ways of testing theories have to be found. It is not feasible to experiment if your prediction is:

- If the average annual temperature rises by 2°C then maize production in Kenya will drop by 15%

You could test the prediction by setting up simulation models **(Chapter 4.8)** that describe the relationship between production and temperature. But those models will themselves be based on theories tested by experiment. Here is another well known prediction made some years ago:

- Regular smoking will lead to an increased chance of lung cancer and other diseases.

It was not possible to test this by experimentation as that would have involved taking a group of people and requiring some of them to smoke. This theory was tested largely by surveys **(Chapter 4.4)** which are distinct from experiments. In a survey you observe what is happening without making deliberate changes. Thus the effect of smoking was investigated by comparing the health of people who smoke with those who don't, and a clear correlation emerged. The limitations of the study design are clear: if the smokers have a higher rate of lung cancer we can not be sure the smoking **causes** the lung cancer. Perhaps there is some unknown factor that tends to lead people to both smoke and get lung cancer. This is a problem of the survey approach to investigation, and means that theory testing is harder using surveys than using experiments. In the case of the health impacts of smoking, various possibilities for such factors were suggested (diet, genetics), and then eliminated by surveys which controlled for them, each providing evidence in support of the theory. However there will always be people who think of one more factor that could be the explanation. This would not be the case if the theory could be tested by a well designed experiment.

This chapter summarises the key decisions that have to be made if you are to conduct a well designed experiment. A prerequisite is understanding the principles and language of experimental design, described in the next section.

Basic ideas of experimental design

Think of the simple problem of determining whether the new maize variety 'Boreproof' is less susceptible to stem borer damage that the commonly used variety M512. That is the **objective**, and the objectives of the experiment will determine all other aspects of the design.

You could plant a field of Boreproof and measure the stem borer damage. But with what will that be compared? The objective requires checking it has less damage than M512. You need to make a **comparison** and so plant a second field with M512. The two varieties being compared are the two **treatments**. They are compared on fields, so the two fields are the **experimental units**.

How effective will this design be? What might you conclude when you get the data? If 50% of the plants in the M512 field are damaged, but the Boreproof field only has 20% damage has the theory been confirmed? Hardly! Any agriculturalist will tell you that stem borer damage can vary greatly between fields, as well as between different parts of the same field. So an alternative design is suggested. Have several fields each of Boreproof and M512. Suppose the results show the fields of Boreproof as having damage levels of 20%, 10%, 40% and 30%, while M512 has 50%, 60%, 40% and 35%. The **replication** of the fields allows you to check the consistency of the results. These results show a tendency for Boreproof fields to have less damage than M512 (the mean is 25% compared with 46%) but the results are not very convincing, with some M512 fields having less damage than some Boreproof fields.

A third alternative design is tried. Since you know that the pest pressure will vary between fields irrespective of the varieties being grown, maybe you can increase **precision** by growing both varieties in the same field. Make the experimental unit a **plot** (say 10m x 10m) of maize, with two plots in each field. Then put M512 on the left-hand plot and Boreproof on the right in each pair. You can now compare the two treatments within each field, and differences between fields become less important. Using the fields in this way is described as **blocking**, with each field being a **block**.

The results from this design are shown in Table 1.

Table 1. Results of a simple experiment

	Stem borer damage (%)	
Field	M512	Boreproof
1	50	20
2	20	10
3	30	20
4	60	30
5	60	40
6	20	5
7	0	0
8	40	10

Are these results more convincing? They certainly show consistency: Boreproof had less damage than M512 in every field except field 7, which has no stem borers anyway. But look carefully at the way the design was described. The M512 was always placed on the left-hand plot. Maybe the difference in stem borer damage is nothing to do with variety, but due to some other consistent difference between left- and right-hand plots. Maybe the wind blows from the left, bringing the pests or stressing the plants. You may know that is not the case, but could have trouble convincing others. And you can never be sure that there is not some other systematic difference between left- and right-hand plots. The solution is to **randomise** the allocation of treatments to plots. In field 1, toss a coin to decide whether Boreproof or M512 goes on the left-hand plot. Then randomise again in field 2, and so on. In field 1 you might end up with Boreproof on a plot with less stem borer damage for reasons unconnected with the variety, but over the whole experiment you can be sure that the only systematic difference between plots with Boreproof and plots with M512 is indeed the variety.

The basic ideas of experimentation described above should also help you understand studies which, though they involve comparison, are not experiments and can not demonstrate cause. For example, suppose a study showed that farmers in Central district have less stem borer damage in their fields than farmers in West district. They also have higher adoption rates for the Boreproof variety. This study involves comparing districts, but is not an experiment as the differences in adoption of Boreproof were not imposed by the researcher. It is also common to make comparisons over time, for example, by comparing stem borer damage levels before and after introduction of Boreproof in Central district. In such studies the change may be devised by the researcher, but it should only be considered an experiment if other features are present, e.g., some other districts in which Boreproof was not introduced, random allocation of the introductions, and some replication.

Diverse applications, common principles

The simple example in the previous section explains the basic ideas and terminology in the context of a 'classical' agricultural experiment – a variety trial. This section shows the correspondence with three other studies of different types. Each discipline tends to produce its own language and standard practices and it is important to recognise the commonality between them, and to make sure that you really understand the logic of the design. The same topic is discussed elsewhere in different contexts.

The examples in Table 2 are all different. One investigates field plots, one animals, and two people. Of the last two, one focuses on individuals and the other on communities. The practicalities of carrying out each of these studies will be quite different, but the fundamental logic of the design is the same in each case. The social sciences do not use the terms **treatment** and **unit** but the rationale for their approach is similar to research in the natural sciences. The roles of comparison, replication, randomisation, and controlling variation are the same for all of them. It is common for these aspects to be forgotten in studies involving people, particularly community-based studies like the last one in Table 2.

Table 2. Four examples of experiments

Objective	Treatments	Units	Measurement
Determine if Boreproof is more resistant to stem borer than M512	1. Boreproof maize 2. M512 maize	10m x 10m plots of land	Percentage plants damaged by stem borer
Find the effect on milk production of substituting dairy meal with calliandra fodder	1. Base diet + dairy meal 2. Base diet + calliandra 3. Base diet + 50% calliandra + 50% dairy meal	Dairy cows for 2 weeks of third month of lactation	Milk production in the second week
Check whether training in pest management allows farmers to produce cabbages more profitably	1. No training 2. Attendance at farmer field school on pest management	Farmers for whom cabbage production is a main enterprise	1. Farmers' knowledge 2. Profitability of cabbage enterprise
Evaluate the effect of community information and organisation on common grazing management	1. No intervention 2. Information provided 3. Information provided and village 'grazing committees' facilitated	Villages in areas where common grazing is degrading	1. Community views on grazing problems 2. Range quality

Design decisions

Now you understand the basics of designing experiments you can start thinking about the design of your experiments. You may need one substantial experiment, several smaller related experiments or possibly no experiments. If you are going to experiment then there are many decisions you will have to make about details of the design. How can you make those decisions? There are several sources of help:

- The fundamental principles of experimental design – the outlines above and more details in other texts
- The more practical ideas in the following sections, and in other texts
- Papers and reports describing similar experiments that others have done
- Other researchers who have worked on a similar topic (maybe in a different region) or used a similar method
- Your observations of other experiments
- Your imagination
- Pilot studies in which you try out techniques and arrangements before committing yourself to an expensive or long-term experiment.

There is no single correct way to design your trial, but there will be plenty of ways that are wrong – designs which will not lead to valid conclusions meeting your objectives. Even if you design a trial that will give valid results it may be inefficient – not give you as much information as possible for the time and effort spent. Avoid these scenarios by:

1. Thinking.
2. Using all the sources of help listed above.
3. Showing your design to others and getting their comments.
4. Envisaging the data your design might produce and the way in which you would then interpret it. Some researchers sketch out the tables and graphs they would use in the

analysis of the data, then making sure the design will generate the required numbers to complete them.

5. Thinking of the practical as well as the theoretical requirements. You have to manage your trial (set it up, look after it), cope with the travel requirements, have enough time and equipment to measure all the plots, and so on. And you have to be able to afford it!
6. Iterate. Start with a possible design, think through the consequences then go back and revise it until you have something sound.
7. Thinking.

In the following sections the main ideas you need to make decisions on each of the key points are described together with some of the common mistakes that you must try to avoid.

Objectives

All aspects of the design depend on the objectives. Therefore you must get the objectives right! Objectives must be:

- **Clear.** If the objectives are vague it will not be possible to decide on the rest of the design
- **Complete.** Often the statement of objectives is incomplete so that the experiment can not be designed.
- **Relevant.** In applied research, experiments are made to help solve real problems and fill knowledge gaps in the process. The objectives of the experiment must be relevant to solving the problem. It must be clear how you will be a step nearer solving the problem once you have the results from the experiment
- **Reasonable.** The objectives must be reasonable given current understanding of relevant phenomena and other observations. Avoid objectives that contain elements of alchemy or wishful thinking
- **Capable of being met by an experiment.** Some research questions do not need an experiment. Two problems which often arise here are:
  - objectives that require a survey rather than an experiment
  - objectives that require two or more experiments rather than a single one.

Make sure that the objectives fit in well with the overall strategy of the project. You have to be able to explain what the next step will be after the experiment is completed.

Common mistakes to avoid

1. Objectives which are too vague. The objectives in Table 2 all fall into this trap! Real experiments would need to have objectives that made it clear, for example, what sort of base diet is to be fed to the dairy cows, how much training in pest management should be given, or where the rangelands are located.
2. Objectives which just say 'the objective is to compare the treatments'. Treatments should be a consequence of the objectives, not the other way around.
3. Loading too many objectives into a single experiment, so no design can be found that meets all of them. For example, in trials in farmers' fields, understanding details of biophysical processes usually requires a high degree of uniformity, and hence the researcher taking control. Eliciting farmers' assessments of the technologies requires them to have a free hand. Thus the two objectives will probably not be met in a single trial.

Treatments

There are four ideas you need when choosing treatments:

1. **Comparison and contrasts.** Experiments involve making comparisons. The exact comparisons that meet the objectives can be defined as contrasts, i.e., the numerical expression of the comparison. Make sure your experiment has all the treatments needed to make all the comparisons implied by the objectives.
2. **Controls.** 'Controls' or 'control treatments' are the baseline treatments against which others are evaluated. In the stem borer experiment M512 might be considered the control.
3. **Factorial treatment structure.** Many experimental objectives require looking at several 'treatment factors'. For example, in the stem borer experiment you may also want to look at the effect of sowing date (early, mid, or late). Then the experiment might have 6 treatments (Boreproof sown early, mid, or late and M512 sown early, mid, or late). Factorial treatment structures are important for two main reasons:
   - they tell you about **interaction** – such as whether the difference between Boreproof and M512 depends on when they are sown
   - if there is no interaction they give information about both factors with the same precision as would be obtained if the same amount of experimental effort went into investigating just one of them. This is the 'hidden replication' described in textbooks.
4. **Quantitative levels.** Some experiments require varying a quantity that could have many different levels, such as sowing date or amount of fertilizer applied. Choosing the levels to use as treatments in the experiment depends on the exact objectives and what you already know about the response to varying it. Generally fewer rather than more levels are needed, and there is rarely a reason for using more than 4 different levels.

Some common mistakes to avoid

1. Including extra treatments 'because they might be interesting' rather than because the meet a clear objective.
2. Missing suitable controls, so, for example, the new varieties are grown but there is nothing against which to assess them.
3. Thinking 'control treatment' means 'do nothing' or 'zero input', even though those might be appropriate in some cases. Control treatments are just treatments needed make the required comparisons, so you may have two or more controls corresponding to different objectives.
4. Using too many levels of a quantitative factor. Using 10 levels, say 0, 10, 20, 30, 40, 50, 60, 70, 80, and 90 kg N/ha will give more information about response to fertilizer than just using 0, 20, 50 and 90 kg N/ha. But if you can make a total of 20 observations (e.g., if you can only afford 20 plots) then 2 replicates of those 10 different treatments will almost certainly give less information about response to N than 5 replicates of the latter set of 4 levels.

Another key idea is **confounding**. Suppose that in the stem borer experiment M512 was sown on 20 March but Boreproof was not sown until 2 April, because there was a delay in procuring the seed. When the stem borer damage is observed to be less in Boreproof we can not conclude that the variety is resistant. The difference in damage may be due to the different sowing dates or different varieties. Sowing date and variety are said to be 'confounded'. **Treatments must be defined in a way that does not confound different effects.**

Units

With crop experiments, decisions have to be made on size, shape, orientation and arrangement of plants within the plot. There are a few guidelines based on theory:

- Many small plots often give more precise results than a few large plots taking up the same area
- Long thin plots often give more precise results than squarer plots.

These guidelines have to be modified by practical considerations:

- Plots have to be large enough to manage (sow, weed, spray, harvest) in a way that represents what a farmer could do
- Plots have to be large enough to take measurements, allowing for the possible disturbing effects of destructive measurements during the experiment **(Chapter 4.5)**.
- Borders may have to be left around each plot to make sure that anything happening on one plot does not influence what goes on in the next plot.

These considerations, particularly the last point, often overrule the theory.

If the units are not plots of land but animals, people or communities then there are often more decisions to make and few general guidelines. **Base the design of the experimental unit on the experience of others who have done similar experiments.** What did they use as the unit? What problems did they have? How will your experiment differ from previous ones? Does that imply any changes in unit?

Think of the experiment with factorial treatment structure, with two varieties (Boreproof and M512) each sown early, mid, and late. A common design for this type of experiment is the **split-plot**. Large plots are defined and the early, mid, or late sowing date allocated randomly to each one. Then each large plot is divided into two, with M512 and Boreproof randomly allocated to the two halves. A split-plot design can have practical advantages, for example, you are less likely to disturb the early sown plots when sowing the later ones. However it does have disadvantages. There are two sorts of plot (large plots and split plots). This complicates the analysis because variation between both types of plot has to be considered. The precision of a split-plot trial is generally less than for that of alternative of random allocation of all treatment combinations to the smaller plot. **Don't use a split plot design unless practical considerations require it.**

Some common mistakes to avoid

1. Plots are often too small, so it is not possible to manage or measure them realistically. An extreme example occurs when measuring labour. It is not possible to estimate the labour required for a task such as weeding if the plot is very small, because weeders will not work at the same rate per unit area as they would in a larger plot.
2. In situations other than annual crop experiments, interference between plots can be hard to see, but can seriously bias results. Water and insects can move from one plot to the next. Tree roots can grow into neighbouring plots. If the unit is a farmer, he or she may talk to another farmer and influence results in an unplanned way.
3. Some researchers seem to believe that experiments with factorial sets of treatments have to be done in split-plots. This is untrue.
4. Split-plot designs are useful but overused.

Replication

Replication (having several units of each treatment) is important for four reasons:

1. **Estimating precision.** The uncertainty in an average is estimated by the variation between the observations being averaged.
2. **Increasing precision.** Calculating an average over more values from a replicated experiment will increase precision since the calculated value will be closer to the true value.

3. **Insurance.** More replications in an experiment will provide some insurance against things going wrong with one or two replicates. Without such insurance an experiment may be rendered useless by, for example, goats getting into the field, or some participating farmers dropping out of the study.
4. **Replication.** This can increase the range of validity of results if a comparison is repeated under a range of conditions.

There should be enough replicates to satisfy all the above reasons for replicating:

1. **Estimating precision.** Look at the error d.f. (degrees of freedom) from the analysis of variance, 10 d.f. can be considered a reasonable minimum. Much more than 20 has no particular advantage.
2. **Increasing precision.** If you have an idea of the precision you need and the variation in your experimental material then it is possible to estimate the number of replicates needed. Details are in books, and software is available to help.
3. **The number required for insurance must depend on the risks.** A long-term trial in a risky environment (e.g., one that might be burned in the dry season) may be worth insuring, by adding replicates. A short-term trial that can easily be repeated if something goes wrong is not worth insuring.
4. **Increasing the range of validity.** Suppose the stem borer trial had some replicates on sandy soil, some on loam and some on clay soil. Then you could be more confident that the results were generally valid than if the experiment had only been done on sandy soil. The importance of this will depend on the objectives.

Common mistakes when planning the number of replicates

1. Using the 'usual' number of replicates (in field experiments this is often 4, for some unknown reason) rather than rationally selecting the number of replicates.
2. Forgetting the 'hidden replication' in experiments with factorial treatment structure.
3. Insisting that all treatments have the same number of replicates. In the absence of any other information this is sensible, but it is not necessary. Suppose you decide you need 10 replicates of Boreproof and M512, but only have enough Boreproof seed for 8 replicates. That does not mean you have to also use 8 replicates of M512. It may be sensible to continue with the 10 replicates.
4. Forgetting that in split-plot and similar types of experiments there is more than one type of unit and each has to be replicated sufficiently.
5. Estimating the number of replicates needed, finding it is too large to manage or afford, and proceeding with far too small a number. The experiment will not meet its objectives! If you can not afford to meet the original objectives then modify them rather than carrying on with an experiment that will almost certainly not be useful.
6. Assuming that sub-samples from one plot are really replicates.

The last point is particularly important and is a common problem in student projects. Suppose you measure stem borer damage by selecting 10 plants at random from the 10m x 10m plot and measuring the damage on each one. 10 plants from one plot do not tell you the same thing as 10 plants from different plots. If different treatments are applied to different plots, it is variation between plots which is important, not variation between the plants within a plot. The parallel mistake in the community experiment would be confusing the information from responses of several people in the same village with responses from several different villages.

Site

The site(s) for the experiment will be determined by the objectives. It has to be representative of the problem area, both on a large scale (for example, in the same agro-ecozone) and on a small scale (for example, having the appropriate soil type and previous management).

The site also has to be practical. It should be:

- Accessible
- Secure
- Large enough

An experiment will have to be made at more than one site if any of the following apply:

1. The problem area is too variable in key characteristics for a single representative site to be found.
2. You are unsure of the key environmental (biophysical, social or economic) characteristics that may determine the outcome of the experiment, so cannot be sure they are represented by a single selected site. Getting consistent results from several sites will give you confidence that these results really do apply to a wider area.
3. The objectives of the trial require conditions to be compared that cannot be controlled as treatments, such as soil type, rainfall or soil depth.

Cases 1 and 3 require sites to be selected in the same way that single sites are selected. There is an argument in case 2 for sites to be chosen by random selection, but that is rarely practical.

The same considerations apply when experiments are carried out with farmers and communities. Do not simply choose the villages or farmers in which last researcher worked, but look carefully at the objectives and decide on which characteristics it is important to have represented.

Some mistakes to avoid

1. Choosing a site, such as a university research farm, for its convenience rather than its suitability in meeting the objectives.
2. When working with farmers and doing experiments in farmers' fields, biasing experimentation to wealthier farmers and more fertile fields. There are techniques to avoid this.

If practicality requires you to do either of these things, then you need to be upfront and clear about how you expect them to influence your results.

Blocks and allocation of treatments

Once you know where the experiment will take place and what the units or plots are, you can define a set of units for the experiment. If the units are field plots then you could mark out the field into plots, avoiding places that are clearly unsuitable (a patch with surface rocks, an old termite mound, the strip adjacent to large trees). If the units are not plots you can do the equivalent:

- **If the units are farmers.** Produce the list of farmers who are willing to take part and meet the criteria determined by the objectives. (Make a special note of the potential bias of using only farmers who are willing, or able, to spare the time to take part)
- **If the units are tracts of rangeland.** Map out their location and negotiate their use with the communities who look after them
- **If the units are villages.** Contact those villages that meet the criteria determined by objectives.

Next, determine which treatment will be applied to each unit. Random allocation should be used. Random allocation does not simply mean 'mixed up'. Avoid any possible bias by using an explicit random process. For example, use pieces of paper with treatment names put into a 'hat'. The number of pieces of paper for each treatment will be the number of replicates. Then decide the treatment for the first unit by drawing a paper from the hat without looking, again for the next and so on. There are computer programs to help with this.

The precision of almost every experiment can be improved by blocking. Whatever units you have, you know they will vary. Some variation is predictable. Try to arrange the units into homogeneous groups, each of which will become a block. Table 3 gives some suggestions on characteristics that might be used to block different types of unit, but what is suitable for your trial will depend on the objectives of your trial. For the stem borer experiment, the level of stem borer damage in the previous season may be a good characteristic to use to group units into blocks. However it would be irrelevant if the trial was about N leaching or weeding regimes.

Table 3. Possible factors to use in definition of blocks

Units	Characteristics used in blocking
Field plots	Soil type Previous crop yield Slope Weeds present
Animals for dairy experiment	Weight Previous milk yield breed
Farmers in pest management experiment	Education level Length of time growing cabbages Size of farm
Villages in community resource management experiment	Ethnic group Presence of community organisation

If:

1. Every treatment will have the same number of replicates and
2. Every block has the same number of units and
3. The number of units in a block is equal to the number of treatments.

Then the best design is to put exactly one replicate of each treatment in each block. **The allocation of treatments within a block should be random.** This is the **randomised-block design.**

If the blocks are not all the same size, or the number of units in each block is not equal to the number of treatments, then you will have an **incomplete block design**. Take care when deciding which treatments go into each block. Software is available to help you with this.

Some common mistakes to avoid

1. Assuming blocking is only useful in field experiments. The idea and terminology was developed in the context of field experiments, but it is just as important in all other experiments, though researchers often fail to block experiments with people or experiments carried out in laboratores and nurseries. **Blocking gives extra precision at little or no cost and is almost always worth doing.**

2. Assuming blocks have to be the same size and equal to the number of treatments. Incomplete block designs can be very useful. If you have to get a bit of help designing and analysing them it will be worth it.

Management

In a field experiment, 'management' means preparing the land, sowing, weeding and all the other agronomic practices needed to raise the crop. In other types of experiments there are equivalent management activities. The management of an experiment is often not considered part of the design, yet it can have a large impact on the success of the trial.

1. Decide whether the objectives demand that you manage the experimental material to a very high level (e.g., zero weeds) or a realistic level (e.g., farmers' weeding practice). The first may be appropriate if you are studying processes such as water or N uptake, and don't want weeds to obscure results. The second will be appropriate if you are evaluating technologies and want them to represent farmers' systems.
2. Avoid confounding treatments with management differences.
3. Aim for uniform management. Often the difference between a successful and a failed trial is in how well the crops (or animals or people) were managed, and whether this was done uniformly. You can improve uniformity by, for example, training fieldworkers and monitoring the way they execute operations.

Measurement, data management, analysis and reporting

The measurement and analysis of the trial will also have implications for design, and have to be thought through at the design stage. Details are in **Chapters 4.5** and **4.7.** You will also have to plan a scheme for looking after the data, details are in **Chapter 4.6**.

Writing it up – the protocol

The protocol is the written description of everything that will be done in the experiment, starting with the objectives and going right through to the analysis, interpretation, and use of the data.

- A protocol should be written for every experiment.
- The protocol must be shared **before** the experiment starts to get input from others. There are always other people with expertise that will help increase the quality of your protocol. You should share the protocol with at least:
  - scientists at your location (who understand the local context and constraints)
  - scientist in the region (who understand what else of relevance is happening in the region)
  - a subject-matter specialist (who should be aware of relevant developments around the world)
  - other students
  - your supervisor
  - a biometrician
- The protocol must be sufficiently detailed for someone else to take over the experiment part way through, or to make sense of the data at the end of the experiment, even if you are no longer around
- The protocol should be kept up-to-date. It is not just a plan, but a record of exactly what you actually do.

The protocol must be securely archived so that information about the activity can be found in the future.

Finally: involve a biometrician

Biometricians and statisticians are trained in the art and science of experimental design. They may not understand all the practical constraints and opportunities in your particular study. But they will be able to help with such technical details as choosing the number of replicates and allocating treatments to blocks. They will also be able to spot flaws in the logic of the design, and help you make sure it will really meet the objectives.

Many researchers only consult an biometrician when they get stuck with statistical analysis of the data. That is too late. Get a biometrician's advice early, thereby guaranteeing a good design for your study.

Resource material and references

Appendix II. ICRAF. 2003. *Genstat Discovery Edition and Other Resources.* World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Coe, R., Franzel, S., Beniest, J. and Barahona, C. 2003. *Designing On-Farm Participatory Experiments: Resources for Training.* World Agroforestry Centre (ICRAF), Nairobi, Kenya. 135 pp.
Available from www.worldagroforestycentre.org

Mead, R., Curnow, R.N. and Hasted, A.M. 2003. *Statistical methods in agriculture and experimental biology.* Third edition. Chapman and Hall, London. 472 pp.

Robinson, G.K. 2000. *Practical Strategies for Experimenting.* Wiley Series on Probability and Statistics. John Wiley and Sons, Chichester, UK. 282 pp.

Schroth, G. and Sinclair, F.L. 2003. *Trees, Crops and Soil Fertility – Concepts and Research Methods.* CABI, Wallingford, UK. 437 pp.

Stern, R.D., Coe, R., Allan, E.F. and Dale, I.C. (Eds.) 2004. *Statistical Good Practice for Natural Resources Research.* CABI Publishing, Wallingford, UK. 387 pp. (In press).

4.4 Designing surveys

Erica Keogh

- **A survey is a well organised, reliable observation of what is going on in the world, that can be used to show the current status, compare different situations and identify relationships between variables**
- **The same principles of good survey design apply whatever the subject, whether the observations are of people, land, plants, animals or institutions**
- **Design of a survey requires choosing the unit of study, defining the population of these units, selecting a sample of units to measure and designing a measurement tool**
- **A successful survey requires good management of the planning, fieldwork and resulting data, not just application of sound statistics**

What is a survey?

A survey is an observation of what is going on in the world at a particular point in time, but we use the term 'survey' in those situations where:

- Data collection is well defined and organised
- Data collected can be shown to be 'representative' and reliable
- Data are competently interpreted
- Resulting information is utilised while the data is of current value.

In research new empirical information about the world can be collected in two ways – by surveys and experiments. Surveys can be distinguished from experiments by the fact that surveys observe what is there. They do not deliberately make changes to observe the effect. Experiments impose planned changes (the treatments) in order to measure the effects they cause, whereas a survey will investigate one or more characteristics of a population. A survey is distinguished from a census in that a census attempts to cover the entire population while a survey attempts to cover a pre-determined portion (or sample) of the population.

Some books, courses, and researchers imply that surveys are only used to study people. People, households, and villages are commonly the 'objects' studied in a survey. But surveys can be used to study just about anything! In agricultural research you might have to carry out a survey of crops, soils, weed populations, or farm animals or the trees, or sediment in the rivers. Many of the principles of survey design and execution are the same for all types of survey.

Survey information can be collected in an extremely structured manner, or may be more informal, or a mixture of the two approaches, or something in between. Whatever the 'tools' used to collect the information, one thing must be made clear – it is essential to maintain consistency throughout the exercise and to avoid errors arising from inadequately prepared tools.

Why conduct surveys?

Many situations present problems into which you can gain insight by the collection and analysis of survey data, thereby allowing you to:

- Determine existing conditions
- Monitor change over time
- Evaluate new projects

- Forecast future needs.

A survey may seek to

- Describe existing conditions
- Establish relationships between different quantities
- Make comparisons
- Test hypotheses,

or a mixture of all of these.

A survey may target a large sample or a small one – the size will be determined by the sampling methods used and will have an impact on the future use of the results.

Example 1

a. Problem 1. Increased elephant damage has been reported in some villages. Are the elephants moving along normal migration routes or are they roaming more widely than before?

Here you would be interested in **describing existing conditions** and, possibly, trying to **make comparisons** with conditions in previous years. The research could be extended over a period of time, thus **monitoring** the situation over a number of years.

b. Problem 2. Is infestation of maize fields by *Striga* worse when fields are suffering from soil erosion?

In this case you could do a preliminary investigation into whether there is a measurable **relationship** between *Striga* infestation and the level of soil erosion. Alternatively, if there is prior information about this relationship, you can **test the hypothesis** that this relationship exists and is quantifiable.

Types of surveys

There are various ways to distinguish one type of survey from another, but perhaps in the present setting it is best to provide examples which will illustrate the wide variety of studies that are possible.

- Street interviews to assess public opinion about price increases of seed maize
- Household interviews to measure food production and consumption for monitoring food security
- Field observations to estimate earworm infestation in the current maize crop
- Field observations and community discussions to quantify the effects of elephant damage to crops
- Household interviews to gauge the effects of HIV/AIDS on labour availability for household agricultural activities
- A case control study to compare old and new tillage practices in different communities
- An enumeration of tree species in quadrats within a specified area for assessing biodiversity.
- A study to estimate soil fertility prior to land preparation
- A study of a sample of records from the meteorology department to track rainfall patterns over the last 50 years
- An investigation of sections of river banks to determine silting levels arising from gold panning.

From the examples you should realise that a survey may entail interviewing people, or collecting specimens, or measuring items, or studying records, or a combination of one or more of these activities. Thus, the type of survey you are planning dictates what measure-

ment instruments you will be using (e.g., a questionnaire for interviews or a tape measure to check the area planted), and also the sampling scheme (the rules for choosing exactly which things will be measured) you will be using. This matching of 'tools' to the type of study is one of the classic features of surveys, with each survey having a unique set of instruments and methodology for efficient data collection.

Example 2

Referring back to the problems introduced in Example 1, some of the terminology you are going to meet when designing and implementing surveys can be illustrated.

	Problem 1	Problem 2
Population	Farmland in area reporting increased elephant damage	Maize-growing areas in western Kenya
Unit	Village	2m x 2m quadrat
Sampling scheme	30 villages selected at random	10 villages selected at random 10 fields selected at random in each village 2 quadrats per field, placed 1/3 and 2/3 of the way across the field from the entrance.
Measurement tools	Questionnaire for village meeting Visual assessment of damage	Counts of *Striga* plants visible in a quadrat Visual assessment of soil erosion in the field

Setting up a survey

An effective survey encompasses many activities, which must all come together to provide a useful and timely report. The actual planning for a survey is as important as its implementation, and the amount of work involved in the planning should not be underestimated. The efficient and successful management of a survey depends to a great extent on a thorough understanding of the population, of the survey topic, and on having well structured administrative backup available throughout. Available resources will often dictate planning decisions, but it is essential to aim to maintain the quality of all procedures by adopting a 'global' viewpoint, i.e., **by considering the impact of each decision made at a particular stage, on the whole project, thereby achieving balance and consistency throughout.** Some examples of surveys have been given. Next we look at the details of survey design and implementation.

Planning the survey

First and foremost you should specify the objectives of the survey.

- What should be accomplished by the survey?
- What should be measured in order to reach your goals?
- What is the analysis plan for the measured variables?

Ask yourself such questions as **'Why do we need to collect this data?'** Many surveys are multi-purpose, information on more than one topic will be collected, and you need to have some ideas of the precision required in the various areas to be studied. **Define the objectives as simply as possible and ensure they are not self-contradictory,** e.g., will analysis of one variable confuse or assist in the understanding of another? The target and study populations (see later) need to be carefully defined in conjunction with those population characteristics to be studied.

Familiarise yourself with all possible sources of existing knowledge from previous studies. Such information can be used not only to identify gaps and thus emphasise the need for the present study, but also to provide checks on possible sources of bias, to help avoid duplicating work already competently carried out, or to improve estimates previously obtained. It is also important to identify all possible secondary users, i.e., those who may have use for your data in the future. Such users can be of great help with planning, avoiding conflict, and suggesting alternative approaches. You may also be in the situation where your research project is but a small part of some on-going larger research project – in this case it is essential to:

- Maintain contact with those implementing the larger project
- Receive information about results being obtained from other sections of the project
- Ensure your project fits in with the overall larger objectives
- Provide timely feedback on your progress to all other players
- Work with others as part of the larger team.

The flow chart shown in Figure 1 illustrates the phases of a survey, each of which needs careful planning right from the beginning.

Right from the beginning, it is essential to:

- Be aware of all resource limitations
- Be able to identify, for each task:
  - Who is going to be responsible
  - How much it will cost
  - How much time it will take.

Surveys involve large amounts of documentation, all of which have to be prepared in advance and tested for ease of usage. Sometimes you will need to recruit persons who can assist you at one stage or another.

Timetables and budgets

Organising the timetable and fixing the budget are major components of survey preparation. The time and funding that are available are the major factors determining the scope and extent of your study. **It is extremely useful to use a Gantt chart (e.g., Table 1) for timetabling, since it enables you to maintain an overall view of all that the study will entail.**

There will always be time restrictions to be adhered to in any survey project. It is better to over-estimate, allowing lee-way for unforeseen happenings. It is often at the beginning of the data entry and processing stages that delays occur, but these can be minimised by adequate pre-testing checks in advance. Allow for realistic staff turnover, which may result in delays. Add contingency amounts of time allowing for weather effects, breakdown of equipment, or errors. Identify critical activities which, if delayed, will hold up other activities, and try to foresee possible alternatives. Previous research in the same area can prove useful in the time-planning context since from this you can identify what may have gone wrong before. **Asking the experienced workers is most useful since they have the first-hand experience you wish to know about. It is essential that the data be analysed while it is still relevant; so the report can be published within a realistic time after data collection.** Decide whether an initial overview preceding full analysis will be of benefit and plan accordingly.

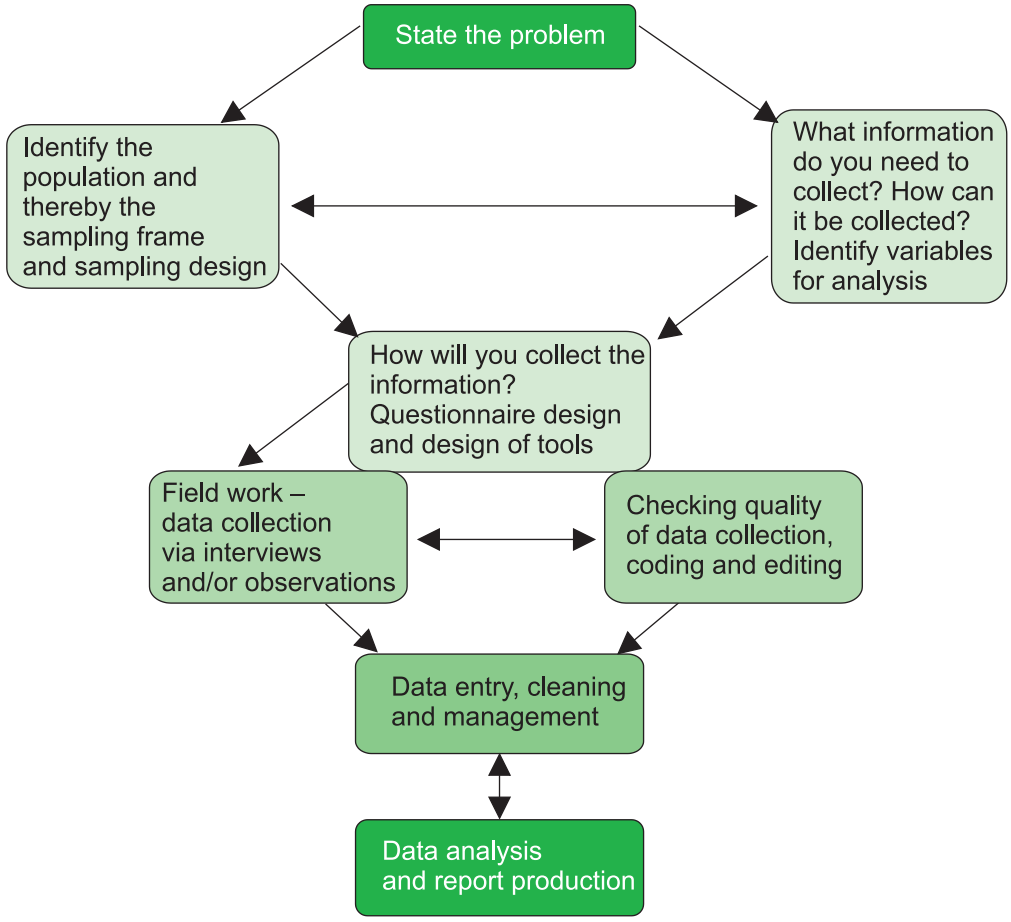


Figure 1. Steps in carrying out a survey

Table 1. Example of a draft timetable for a crop management survey

Task	Week number																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Consultations with communities /publicity	•	•				•	•						•	•	•	•	•
Questionnaire design and testing	•	•	•														
Sampling design and sample selection		•	•	•	•	•											
Design of data entry			•	•													
Data analysis planning			•	•	•												
Field staff recruitment		•	•	•													
Training of enumerators and pilot					•	•											
Printing of tools (questionnaire)					•	•											
Fieldwork and checking						•	•	•	•								
Data entry and validation						•		•	•	•	•	•					
Data cleaning and analysis						•					•	•	•	•	•	•	
Production of graphs and tables														•	•	•	
Report preparation					•			•	•	•		•		•	•	•	•
Archiving				•	•											•	•

Source: United Nations (2004)

Budgeting

Hand-in-hand with timetabling for the survey, is the survey budgeting. This is probably the **most difficult task of all since the survey design is totally dependent on the budget**, and *vice versa* – so which comes first?

The main components of a survey budget

- **Publicity and information** – including meetings, agreements, workshops
- **Wages and salaries** – including contingency planning for ill-health, adverse weather, inflation, resignations, after-hours working, field allowances
- **Transport and communications** – including phone, fax, postage, and e-mail usage, fuel, hire charges, bus fares
- **Meals and accommodation**
- **Equipment and consumables** – including hardware and software, printing equipment, clip boards and note books, maps, files
- **Printing and duplicating** – a major component of the budget
- **Hidden costs** – equipment usage.

Errors

A survey requires and combines the techniques of sampling, design of tools, data collection and data analysis, and **the accuracy of the methods employed will determine the quality of the information finally produced**. In any survey there are many potential sources of error which may be broadly classified as sampling errors and non-sampling errors.

Sampling errors. These are errors arising because, by chance, the sample is not fully representative of the population. Such errors can be estimated and are a random result of the sampling procedures. Broadly speaking, the larger the sample size, the smaller the sampling errors.

Non-sampling errors. This category includes all of those errors which can arise from other sources:

- Variation between data-collection personnel
- Inadequate tools
- Inadequate sampling frame
- Data-entry errors
- Coding errors
- Non-response
- Errors in response
- Effects caused by the way questions are worded.

Each of these can give rise to bias which is often not measurable. **Bias** means that the results based on the survey are not, even on average, the same as those that would have been derived from a total census of the population, but consistently over- or under-estimate quantities.

Increasing the sample size, so as to reduce sampling error, can very well increase non-sampling errors due to resulting poorer-quality enumeration and lower levels of supervision. **Sampling and non-sampling errors and their relative magnitudes must be considered simultaneously when determining sample size.** Often only sampling errors are mentioned since non-sampling errors are usually not measurable, and sometimes unknown. **You must remember that you will be making many measurements on your sample and that the precision of estimates is likely to vary from factor to factor.**

Sampling

The theory of sampling is covered adequately in many texts and only some brief notes are made here. There are two inter-related decisions to make: the type of sampling and the sample size. **Decisions** depend on blending the theoretically optimal with what is really practical, in the light of the survey objectives.

Decisions about sample size must be taken in the global context of the project and must include consideration of the following factors:

- Available resources
- Objectives of the study
- Sub-groups, within a population, that you wish to study
- Practical constraints
- The precision needed
- Homogeneity of the population.

Following are definitions and examples illustrating them.

The population

The **target population** is defined as all those units in which you are interested. The **study population** is defined as all those units that you can reliably identify. Ideally these two populations should coincide but, unfortunately, this is often not the case, particularly when the population consists of people.

Units

When you implement your survey, you are going to be dealing with a **unit**, i.e., you are going to interview a **person**, or generate discussions with a **group of people**, or count the number of *Striga* plants in a **quadrat within a field.** These are the units of study. In many studies there is a hierarchical arrangement of units. We measure things on people, but also record something about the household they are in, the village in which the household is found, and the district where the village is located. This hierarchy may be used in sampling, even if measurements are taken at only one level.

Sampling frame

The **sampling frame** is a 'list' of all the items from which you are going to select your sample, noting that you need a separate frame for each level in the hierarchy of units. Careful construction of the frame is needed since, as mentioned above, unexpected errors can easily arise if the frame is out of date, or if it has inaccurate or duplicate records, and so on. If the frame is inadequate we say it exhibits **over-coverage** or **under-coverage** – these terms simply reflecting the non-match of study and target populations (Figure 2). If you are sampling households, then the ideal frame is a list of all the households. If sampling fields or rivers, for example, the sampling frame may be a map or aerial photo, i.e., an implicit list.

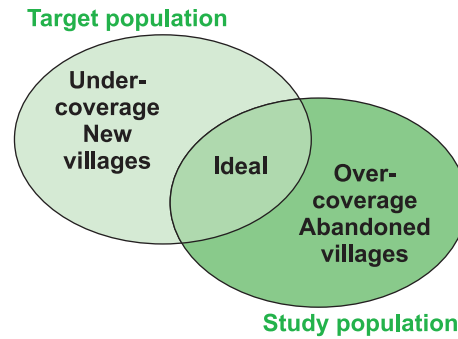


Figure 2. Target and study populations

Example 3

Refer back to Problem 1 in the previous examples. Not only do you want to observe and measure the actual damage in the fields, but you also will wish to interview the villagers and discuss with them their methods for protecting their crops. Another aspect of interest will be gender differences in managing crop damage. Suppose the district authorities provide you with a map on which the locations of villages in the study area are marked. Your first stage of sampling will be to select villages and thus your **target population** is all villages in the area, whilst your **study population** is all villages marked on the map provided to you. If the map is out of date it may mark a village which, no longer exists because its inhabitants moved out to another area 2 years ago precisely because of high rates of crop damage. We call this **over-coverage** since that village would potentially be selected into the sample (according to the map) and yet it does not really exist. Conversely, if a new village has been formed, with some inhabitants of one village moving away and making it their own new settlement area, then this village may not be marked on the map at all and so will not be available for selection into the sample. We call this **under-coverage** since that village is not (and yet should be) available for sampling. Both of these situations will give rise to **non-sampling errors** that cannot be measured and you may never know they exist.

Different approaches to sampling

There are various approaches to sampling and each survey will entail its own unique sampling design. Sampling texts will provide you with the theoretical details and what is aimed for here is to provide a 'feel' for knowing which approach to select. Firstly, two ways of selecting a sample are described.

The simplest type of sampling is known as **simple random sampling**. This involves assigning a unique identification (ID) (e.g., a number) to each item in the sampling frame and then randomly selecting the number of units you require, e.g., numbered pieces of paper placed in a hat and randomly selected one after another. Another very useful method is **systematic sampling**. Here the sampling units are listed in some order that bears no relationship to the topic under study, e.g., listing names in alphabetical order. A starting point is then randomly chosen, and thereafter the sample is determined using what is called the **sampling interval**. This method is often applied to a situation when you have a map or a grid of an area, which can be sectioned into **cells** each of which is then numbered and a systematic sample is easily selected. This latter approach is often referred to as **sampling in space**.

Next, there are a number of ways of classifying the population in different ways before carrying out the sampling, whereby aiming to make use of existing knowledge of the population to ensure the sample is an adequate representation of that population.

Stratification of the population is an approach used when the population is heterogeneous and can be subdivided into homogeneous sub-populations, each of which will be of interest in themselves. Random samples are then selected from each sub-population or strata. In **cluster sampling** the population is divided into clusters that are groups of sampling units which are not similar and one cluster can exhibit the whole range of variability of the population. Using simple cluster sampling, divide the population into clusters, then select a random sample of clusters and investigate each study unit in each selected cluster. In **multi-stage cluster sampling** you can select a sample of clusters but then, within each cluster, further select a random sample of study units.

Clearly it is often useful to classify the population in more than one way, and thus you can use techniques of stratification and clustering together. The final selection of the units you are going to study is usually done using either simple random sampling or systematic sampling. The following examples will clarify these notions.

Example 4

Referring back to Example 3 – recall we have a map (hopefully up-to-date) showing the location of villages in the area of interest. As noted before, the villages marked on the map will be the items in the sampling frame for the first stage of sampling. The actual people resident in each selected village, i.e., the households, will represent items in another sampling frame, for a second stage of sampling. Focus for now on this selection of households and recall that you are interested in the gender dimensions (of head of household) of crop protection from animal damage.

Discussions with the district officials and some initial contact with communities in the area will provide you with information about the overall picture of wild animal marauding in the district. You discover that in one area the main proponent of damage are elephants, with lesser damage caused by baboons and jackals, whilst in another part of the district with a different vegetation type, there was apparently an influx of Quelea birds which caused terrific damage during the past month. The remainder of the district suffers little from large animal damage, with only baboons and jackals causing any measurable loss. Obviously, it will be of interest to study the whole district, even though the elephant damage is only restricted to one area – by looking at the whole district you would hope to be able to compare and contrast areas with different levels of elephant damage.

On the basis of the above observations you decide that there should be three strata within the district. Within each stratum, villages can be randomly selected – probably using a systematic sample from the map. This will constitute the first stage of sampling.

The second stage of sampling, that of households within villages, can be approached in a number of ways. Firstly, for each selected village, the village head could be asked to prepare two lists of names of heads of households, a male list and a female list, and a random sample of households can be drawn from each list. Alternatively, you could, with the assistance of the village head, draw up a map showing all households in the village, marking each one as male- or female-headed. Within each group of male and female household heads, each household will be given a number and then, for each gender group, a systematic sample of households can be selected.

Heads of households can be interviewed using a prepared questionnaire to extract demographic details and obtain estimates of crop damage that has occurred in the past two seasons. In addition, focus group discussions can be held with key informants in each village in order to obtain in-depth information and opinions on the issues of crop damage and ways to reduce it.

The process described in Example 4 is called **multistage sampling** – in other words you sample at various stages of the population hierarchy, ensuring that at each stage you select an adequate sample from each sampling frame. **Issues of sample size at each level of sampling will need to be discussed and finalised with someone who understands the theoretical aspects of random sampling.**

Example 5

Now let us take Example 4 further, and address the issue of data collection from the fields

and storage places of the villagers. This will constitute a third stage of the multi-stage sampling process. One approach is to use the selected sample of households as a starting point for selection of actual sites for measurements of crop damage. Each selected household will have a number of fields under cultivation. Depending on the size of area under cultivation, it may be sensible to sample areas within fields for exact measurements, or another approach could be to sample whole fields from those under cultivation. Sampling areas within fields can be done by mapping the field, dividing it into plots of the size to be examined, and then randomly selecting a sample of plots or quadrats – this can be easily carried out once the map is drawn up. Additional considerations that must be taken into account when planning this third sampling design include the type of crop planted, direction(s) from which animals invade, location of water sources, and any other factors that may have a bearing on crop damage.

An alternative approach would be to ignore the sample of selected households and begin afresh, requesting the community to draw up a map of all planted fields, including crop and animal access information, location of water sources, etc., as above. Planted areas may then need to be clustered before sampling quadrats within each cluster.

Selection of storage places for recording types of storage and amount of damage can again be approached in several ways.

Sample size

Decisions on sample size depend on a number of factors, including:

- What is required in terms of precision of variables measured?
- Just how much variability is there expected to be in each item to be measured?
- The practicalities – how big a sample can you actually deal with, in terms of both time and resources?
- What sub-groups of the population are really of interest? You need to decide on the sample size for the smallest sub-group of interest to ensure that the sample for this sub-group is adequate for realistic estimation
- Which variable should be used to calculate sample size?

A good way to think about sample size is in terms of obtaining a **confidence interval,** i.e., what width of confidence interval will be acceptable for decision-making, based on the survey results? The width you need will be used to determine the sample size, for each sub-group of the population. Expressing the results in terms of confidence intervals helps in interpreting the results more realistically. **If the confidence interval is too wide then no meaningful conclusions can be made.** As mentioned earlier, **the larger the sample size the narrower the confidence interval** – but increasing the sample size is likely to increase the cost and non-sampling errors. **Managing a large sample survey requires extensive resources and personnel if quality is to be maintained**, and it is only the large agencies who can afford this type of survey. But if the sample size is too small, then once again the quality of the estimates is at stake, and results will not be meaningful.

It is not true that the fraction (*f*) of the population sampled greatly influences the accuracy of the sample. The information in a sample of 50 from a population of 10,000 (f = 50/10000 = 0.5%) is much the same as that in a sample of 50 from a population of 100,000 (f = 0.05%), other things being equal. The sampling fraction is not something to consider when fixing the sample size, and aiming for a 10% sample or a 5% sample is not logical. The only exception to this is when f starts to get large – say over 20%.

You should be constantly aware that **each survey study planned and implemented is a unique case and thus 'standard' sample sizes do not exist. You** should familiarise yourself with previous research, but only use it to provide guidelines for your own study. The sample size used last time may or may not be suitable for the current study and you should make your own considerations and do your own calculations, rather than assuming that those used in a previous study were suitable.

Designing measurement tools

The design of the measurement tools needs to be done in conjunction with:

- Formulating and stating the objectives
- Planning which variables to collect
- Deciding how to analyse the information collected
- Consideration of time and resources available
- Bearing in mind the eventual report to be produced.

It is all too easy to imagine that you will be able to collect vast amounts of information from each unit studied – the reality is that it is usually not possible or desirable. Hand-in-hand with development of measurement tools, must be the preparation of the data analysis plan and drafting the outline of the final report.

For each item of information collected, there should be one or more corresponding sections in the analysis plan.

If your survey involves communications with groups of people, then you have to be aware of the time you are going to demand from them to assist you in your data collection. Even if your data collection only means laying out and measuring quadrats in someone's fields, they are going to need to accompany you to do this and you have to be able to rely on them.

People are busy and, in addition, many other researchers may be demanding their time. Thus deep thought should be given to the design of the data-collection tools and, for each variable selected for study, you have to ask yourself 'What useful information am I going to get from this?'

Questionnaires

When communicating with people, either via a structured interview or via focus group discussions, or by any other means, it is wise to lay out a questionnaire ahead of time and to know in advance the type of answers you can expect from each question. Questionnaire design is extremely important – when you are interviewing people, you are assuming that:

- Everyone has the same understanding of each question
- Each question does have an answer
- Each question can be relatively easily answered
- Each question should be relevant to your study
- The question is not 'leading' the respondent towards a particular answer.

Remember that sensitive questions can upset people, which will lead to inaccurate information being provided. Good questionnaire design can only come with experience and it is wise to always ask for assistance.

Questions can be classified as **open** or **closed.** An **open question** is one for which any answer is accepted and recorded in full. A **closed question** is one in which you supply pre-

determined **response categories** into which each and every response should fit. Thus, **response categories** should be:

- Non-overlapping, i.e., mutually exclusive
- Exhaustive
- Permit an overview of the situation
- Neither too many nor too few
- Placed in a logical order.

Open questions provide more information than do closed questions but they are correspondingly harder to analyse and **wherever possible it is best to use closed questions. Focus group discussions (using open questions) are extremely useful for finding out general information and situations on the ground.**

Finally, remember that you should place your questions in a sensible and logical order so that the interview/discussion will flow.

Other data-collection tools

If your survey involves measuring one or more items you will need to prepare for this in advance of data collection. Usually, you should keep in mind the data entry format you will eventually use, since if the formats for both are similar it makes the data entry easier and less prone to errors. Thus, you should design a **spreadsheet** which can be imitated in **data-entry format**, containing clearly defined rows and columns in which values can be recorded, and including a column for comments that will remind you of the circumstances of the collection at data entry time. As with questionnaires, **it is important to collect only the information you really need and which you can really use.** Site details – date, place, time, methods, personnel – can be coded, but must be part of each record. **It is also essential to draw up a protocol for data collection** – this will be a series of detailed instructions and a description of how to actually go about collecting the information required, e.g., how to lay out the quadrats and how to count *Striga* within each quadrat. **The purpose of the protocol is to ensure consistency in data collection methods and implementation,** particularly if more than one person is to be involved in the exercise.

Testing the tools

Once you have designed your basic data-collection tools, the next step is to test them. We call this first testing exercise a **pre-test**, and it serves not only to see whether the tools are suitable, but also to gauge the responses to be expected and thereby to refine, adjust, and further develop the tools. This testing should be carried out using a small sample of units that will not be involved in the main survey. Those people testing the tools should be experienced researchers so that they can react properly to needs which will be highlighted.

After pre-testing it is 'back to the drawing board' again to prepare the final draft of the tools, and the final draft of the data-analysis plan. Mobilising resources and personnel will also be undertaken during this time, until finally you are ready to conduct the **pilot study.** The purpose of the pilot study is to test not only the measurement tools, but also to act as part of the training for personnel, and to test the data entry and data analysis plans. Usually the pilot study is carried out in the same area as where the intended study will take place, using a sample (from each sub-group of the population) that will not be involved in the final study – thus **the sampling design must be completed and known prior to the pilot study.** All personnel to be involved in the final study will also be involved in the pilot study – in

some ways it is like a 'mock' study of all procedures. When the pilot is complete you can finalise the measurement tools and reproduce them in bulk as required.

The time between the pilot study and the main study should be as short as possible so that all personnel remain in the correct frame of mind for the main study.

Fieldwork and data collection

The pilot study is part of the preparation for fieldwork. Once the pilot study is complete and the tools finalised and reproduced, you should start the main study as quickly as possible. The training of personnel should be thought of as an on-going exercise – before, during and after the pilot study personnel will be becoming more and more familiar with the measurement tools and all the needs of the survey. During the initial start-up period of the survey it is wise to meet with all personnel on a daily basis so as to maintain standardisation of data collection. The team leaders should be moving from one person to another to:

- Check that each is collecting the information in the required manner
- Check through collected data
- Pass on the completed forms for further checking
- Code and data entry
- Liaise with other team leaders.

Once team leaders are satisfied that their members are acting as expected, the teams can disperse, but the team leaders should continue close monitoring and liaison with each other.

Data management

A survey generates a huge amount of data and thus it is essential to be absolutely organised for every aspect. Data collection forms should bear unique ID numbers which, by means of codes, will enable the data manager to know exactly where that data was collected, and by whom. The team leaders should check each completed form in the field and, if there are problems, the person who collected the information will have to return to the site and repeat the process. Once the team leader is satisfied with a form, he/she will pass it on to the data manager. If any coding is to be done it is now that it should occur – for instance, categorisation and consequent coding of the content of open questions can take place at this time. Thereafter the form is ready for data entry. Often data will be entered twice – **double data entry** – an approach that is recommended since it nullifies errors of entry – many statistical packages offer this facility. Those doing data entry should have been involved in the planning so that they are aware of the survey objectives, familiar with the measurement tools, and thus in a position to spot inconsistencies and/or errors on the data forms – in this way **cleaning** of the data begins even at the data-entry stage.

Full-scale cleaning of the data usually takes place once all data has been entered and the data files merged into one. **Cleaning** involves examining each variable in turn, looking for **outlier values** and **inconsistencies**, particularly in respect of other variables that provide complementary information. Once the data is pronounced clean the data analysis plan can be put into action and results obtained for input into the final report. Additional **recoding** may take place during the data analysis, e.g., merging categories of responses for more realistic analysis.

Data storage

The importance of **backing up** your data files cannot be emphasised too often. At least three copies of each of the following files should be kept, preferably on CD's

- Original data entry files, obtained before cleaning
- Cleaned data files
- On-going analysis files
- Records of comments on data collection
- Records of progress on data collection
- Records of coding and recoding
- Tables and other results of the analysis plan
- Reports.

All of this information will, eventually, feed into the data archive that should be set up on completion of the survey.

Reporting

The survey report should follow the phases of the survey. **Each phase should be reported upon fully, including both good and bad aspects.** Full details of measurement instruments, training instructions, field reports, coding procedures, cleaning procedures, and the data analysis plan, should be available as appendices to the main report. **Don't forget to report on the non-sampling errors!**

Resource material and references

Appendix II. ICRAF. 2003. *Genstat Discovery Edition and Other Resources*. World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Alreck, P.L. and Settle, R.B. 2004. *The Survey Research Handbook*. Third edition. Irwin-McGraw Hill, Burr Ridge, Illinois, USA.

Ballar, B.A. and Lanphier, C.M. 1978. *Development of Survey Methods to Assess Survey Practices*. American Statistical Association, Washington DC, USA.

Belson, W.A. 1981. *The Design and Understanding of Survey Questions*. Gower, Aldershot, UK.

Casley, D.J. and Kumar, K. 1988. *The Collection and Use of Monitoring and Evaluation Data*. IBIRD/World Bank, Washington DC, USA.

Casley, D.J. and Lury, D.A. 1993. *Data Collection in Developing Countries*. Oxford University Press, Oxford, UK.

Errington, A.J. 1985. Sampling frames for farm surveys. *Journal of Agricultural Economics* 36(2): 251–259.

Ferber, R. 1980. *What is a Survey*. American Statistical Association, Washington DC, USA.

Hayek, L.C. and Buzas, M.A. 1997. *Surveying Natural Populations*. Colombia University Press, New York, USA.

Kish, L. 1987. *Statistical Design for Research*. John Wiley & Sons, New York, USA.

Lohr, S.I. 1999. *Sampling Design and Analysis*. Duxbury Press, Pacific Grove, California, USA.

Moser, C.A. and Kalton, G. 1985. *Survey Methods in Social Investigation*. Gower, Aldershot, UK.

Oppenheim, A.N. 1978. *Questionnaire Design and Attitude Measurements*. Heineman Educational Books, London, UK.

Scheaffer, R.L., Mendenhall III, W. and Ott, R.L. 1996. *Elementary Survey Sampling*. Duxbury Press, Kent, UK.

Thompson, S.K. 2002. *Sampling*. Second edition. John Wiley & Sons, New York, USA.

United Nations. 1964. *Recommendations for the Preparation of Sample Survey Reports* (Provisional issue), Statistical Papers, Series C, ST/STAT/SER,C/1/Rev.2, New York, USA.

United Nations. 1984. *Handbook of Household Surveys* (Revised edition), Series F, No. 31. New York, USA.

United Nations. 2004. *Technical Report: Household Surveys in Developing and Transition Countries: Design, Implementation and Analysis* (Section D, Chapter 14) (in preparation) ST/ESA/STAT/AC.85/www. New York, USA.

Internet resources

www.reading.ac.uk/ssc/develop

http://unstats.un.org/unsd/hhsurveys/index.htm

4.5 Measurements

Jane Poole

- **Measurements generate the primary data in your study, whether it is a survey or experiment**
- **You will have to measure not only the primary quantities that meet your objectives, but those data that help explain and qualify them**
- **There are always alterative ways of measuring anything. Choose the method that best meets your objectives while being practically feasible**
- **Pay attention to quality control: careless measurement can jeopardise the whole study**

Introduction

Measurement is a general term that encompasses many types of data collection. Measurements may be numbers that a scientist collects, such as yields of a crop in a field trial. But they can also be notes made of a farmer group discussion, climatic data provided by a local meteorological station, or responses to survey and interviewing questions.

Every aspect of your research study needs careful design. This includes choosing what measurements to take, when to take them and why. You must also consider how to measure and how much to measure. For large trials and surveys it may be necessary to delegate data-collection to other scientists, local extension officers or farmer representatives – you will need to decide who takes the measurements. This chapter provides general guidance on how to make these choices and highlights important issues to be considered.

What are measurements?

Measurements generate the data you need for your research. You require these data and their analyses to make your research conclusions. There are many different types of measurements and your choice of which to use will depend on the objectives of the study, and on other details of the design. The measurements needed will also determine some aspects of the design.

The following are examples of measurements that may be taken for different types of agricultural research. These examples are just a small selection of the hundreds of possible measurements you could take.

- **Laboratory trials** – chemical properties of soil and water samples, pathogen growth on petri dishes, insect mating and offspring production, eating routines of insect pests
- **On-station and on-farm field trials** – plant heights, insect pest and disease levels, crop and biomass yields, root damage of plants, farmer-participatory evaluation of varieties, labour requirements
- **Participatory research** – farmer group characteristics, farmer perception of new technologies, farmer evaluation of on-station demonstration trials
- **Biophysical surveys** – site location and characteristics, plant varieties, crop management, scientist-evaluated disease infection levels, farmers' perception of disease infection levels

- **Socio-economic surveys** – site location and characteristics, household and farmer characteristics, farmers' perception of crop management practices, farm labour information
- **General/environmental measurements** – weather data (rainfall, temperature), soil type and properties.

Types of measurement

Qualitative and quantitative

Both qualitative (farmer opinions of new technologies) and quantitative (crop yields) data require measurements. Quantitative measurements are necessary for many analyses and interpretations. Qualitative data can often add insights and explanations that are hard to capture in numbers. The distinction between the two is not always clear. Qualitative data (farmer reasons for crop failure) can be quantified after coding, (e.g., by noting whether or not 'drought' is given as reason for crop failure and then reporting the proportion of farmers who give different coded answers such as the proportion believing 'drought' to be the reason).

Example 1

An on-station researcher-managed trial was conducted to investigate sorghum varietal resistance to stem borers. Quantitative measurements were taken of the number of stem borers in the stems, stand count and crop yields. Local farmers were then invited to the station to view the different treatments and group discussions were held to elicit farmers' opinions on the performances of the varieties. These additional qualitative data provided the researchers with information about: characteristics farmers found important, the opportunities for transferring the experiments on-farm, and the likelihood for farmer uptake of the most resistant varieties.

Repeated measures

Measurements taken on the same unit (plant, plot, household) repeatedly during a study are called **'repeated measures'**. These type of data are frequently used in laboratory and field trials, e.g., plant disease levels estimated every week, the growth of a fungal pathogen on a petri dish measured every 3 days.

When you are collecting repeated measures how often should you collect the data? In some cases the answer to this question is simple, as when data are required after chemical spraying or rain, then the occasions are defined. In other instances it is up to you to decide how frequently to measure.

General guidelines when choosing the number of repeated measures to take:

- If you want to fit a (growth) curve to your data then 4–5 time points are usually sufficient
- When you don't know which time points will give you information (plant disease levels in a field trial may stay constant for some time) then you may need to take measurements regularly (once a week). Note that for the plant disease example there is no point taking measurements at the start of the trial if there is no disease present. In this case you should be checking the site regularly and then start taking measurements when the disease starts to appear, otherwise you will spend a lot of time collecting a lot of zeros!
- It is not essential that the observations be taken at equal time intervals. However, it is important to record details of each time point so that the patterns observed can be accurately plotted (time plotted on the x-axis on the correct scale).

Destructive and non-destructive

An important option to consider when taking laboratory and field trial measurements is whether they are to be made **destructively**, as when a plant is cut down to measure root sizes, or **non-destructively**, as when you simply measure plant height. If it is your trial, make sure destructive measurements will not disturb further observations. If you intend taking destructive measurements in farmers' fields make sure they understand and agree. The 'destructive' option is not usually applicable to socio-economic surveys and participatory methods of research!

Example 2

A researcher wishes to measure above-ground biomass in an agroforestry trial, over a period of 3 years. The plot size is set at 10m x 15m. He/she has several measurement options for evaluating the amount of biomass, some 'destructive' and others 'non-destructive'. What measurement(s) could he/she take? Some options are in Table 1.

Table 1. Options for measuring biomass in an experiment

Measurement	Advantages	Disadvantages
Destructively sample a few plants per plot at regular intervals	• Collect large amounts of data on biomass production	• Lower precision of yield estimates (increase plot size to overcome this) • If plant size within a plot is highly variable then a large sample is needed for a precise estimate of biomass • Time requirements are high
Destructively sample a few plants from the guard rows at regular intervals	• Collect large amounts of data on biomass production	• Guard rows may not be representative of the plot • Over time the guard rows will lose their ability to 'protect' the crop • Time requirements are high
Destructively harvest the whole plot at the end of the experiment only	• Time requirement is low • Does not require extra plot area	• Have no idea of the biomass production over the 3-year time period
Record the plant heights at regular intervals and harvest the whole plot at the end of the experiment	• Does not require extra plot area • Large sample (whole plot) can be measured • Plants can be followed over time	• The height measurements may not be representative of the biomass yields
Record the plant heights at regular intervals. A sample of plants grown close to the trial are harvested regularly	• The harvest measurements from neighbouring plants can be used to calibrate the non-destructive measurements	• Requires a lot of experience to correctly calibrate the measurements

Bulked samples

Some variables, like the soil samples and chemical properties can be measured by **bulking** together samples collected in the plot (or laboratory, site, etc.). You take N samples from a plot/location and mix them together to form a single composite sample. M sub-samples are then extracted from the composite mixture and measurements taken for each. Things to note about this type of measurement:

- The variation you observe between the M sub-samples is due to measurement error and/or poor mixing. It has nothing to do with the variation in the plot
- The closeness of the measured values to the plot value will depend on how close the value in the bulked sample is to the plot value. This is determined by the N field samples. The more samples you bulk together (i.e., N is large) the more representative of the site your composite mixture will be
- If the N field samples are highly variable, or collected in a way that introduces bias (e.g., all samples taken from one corner of the plot), then increasing the number of sub-samples you take (M) will not help.

Think carefully about the information you really need. Do you want to know how soil P, for example, varies between different samples from the same plot or how it varies between different plots? If you only need the latter then maybe M can be 1, but N may still have to be large to make sure the bulked sample really represents the whole plot.

What measurements to take and why?

Your choice of measurements (the what, when, why, how and how much, who to measure?) depends primarily on your **research objectives.** You must ask 'What data do I need to collect and analyse in order to achieve my objectives?' Careful consideration of your detailed objectives, together with the practicalities of measurement – this means the resource availability, will assist you to collect the relevant data. **Researchers often take measurements that will not help them to answer their objectives and/or take measurements which duplicate the information. This is usually because they have not given sufficient consideration to the data they need.** Collecting data you don't really need is a waste of time, and in some cases is even unethical (for example, in a household survey in which you take up the householder's time). Failing to collect data you do need will mean you can not achieve your objectives.

Measurements may be **primary responses** that are central to answering your research objectives or **variables** that help to explain them. Examples of primary responses may be crop yields (Objective: compare yields under different management methods) or disease infection levels (Objective: map the geographical distribution of the disease in a region). Primary responses are usually highly variable, with variation at every level of the design hierarchy. You therefore need to investigate the reasons for this variation and may also want to make comparisons with similar research. In the second example above the researcher collects additional data on potential sources of variation such as soil type, climatic conditions and crop management. The following are examples of measurements that may be relevant to specific research objectives.

On-station field trial

Objective – Evaluate the effects of 5 treatments on cabbage crop aphid numbers. (Detail – treatments include chemical, biological and un-treated control, 4 blocks). (Table 2)

Table 2. Suggested measurements in cabbage experiment

Primary response measurement(s)	Additional variable(s)
Plant aphid counts (every 7 days)	Yield at harvesting (to investigate the aphids' effect on yields)
	Rainfall and temperature (changes in climatic conditions may affect aphid numbers, how do these relate to the treatment effects?)
	Soil fertility measurements

Socio-economic survey

Objective – Investigate farmer perceptions of the impact of *Striga* on their maize yields. (Detail – 200 farmers interviewed in one district of Kenya).

NB. This survey could be combined with a 'researcher observed' level of *Striga* to compare farmer perception to the actual levels of infection.

- Look carefully at your research objectives
- What are the primary response measurements you should take so that you can answer your objectives?
- What are the additional variables you could measure (Table 3) that will help you to explain the patterns you observe and enable you to compare your research to similar work?

Table 3. Suggested measurements in *Striga* survey

Primary response measurement(s)	Additional variable(s)
Farmer perception of *Striga* levels	Maize management methods (that may affect levels of infection)
Farmer perception of yield loss due to *Striga* infection	Importance of maize to farmer's livelihood (looks at the impact of the perceived *Striga* levels)

So, your measurement options are determined by your research objectives. But often you will find that several different measurements could be used to answer the objectives so how do you decide which ones to use, without duplicating the information? The answer to this question depends on your research design, available resources, and practical considerations.

Research design

Almost all experimental designs have more than one level of hierarchy (villages/ farms/fields/ plots, or plot/row/plant/leaf) and you have to decide what measurements to take at each level. Different quantities should be measured at different levels of the hierarchy, for example, the wealth of a farmer is usually measured at the household level, the crop yield may be assessed for each plot, and tree height has to be measured on individual trees. Other variables may be measured at higher levels, for example, discussions with a farmer group will generate village-level variables.

The type of research you are doing also determines which measurements are appropriate. For a researcher-designed and managed trial it makes sense to take measurements on every plot and location. In a farmer-designed and managed trial measurements may only be taken on some plots. In a farmer-designed and managed varietal trial the objectives require crop yields to be measured. However, on some farms the level of crop management was very low and weeds greatly reduced yields. Yield measurements were taken on the sub-set of well managed farms and conclusions applied to this environment. The reasons for varying management input were recorded on all farms to explain the differences between the well managed and poorly managed sites. In this example measuring the yields on poorly managed plots would not have provided the information necessary to explain varietal differences.

Research resources

The type and number of measurements you can take will depend on the resources available in terms of time, money, and human resources.

It is often not possible to take as many measurements as you would like due to a lack of these resources. So, should you take small samples of many different types of measure-

ments or fewer types with more samples? The answer depends on how precisely (i.e., the size of measurement error) you want to evaluate each type of variation. It is often possible to simplify your measurements, by using indicators and proxies, so that a larger sample can be measured. Review the following two situations and decide which measurement option you would take.

Situation 1

Conduct a biophysical survey to evaluate the levels of coffee berry disease in five coffee-growing districts. You have enough resources to sample 1000 trees. The majority of farms have around 200 trees and there are approximately 500 farms in each district (Table 4).

As an alternative to this option you could increase the number of farms to 20 and decrease the trees sampled per farm to 10 – thereby increasing the precision at the district level but decreasing precision at the farm level.

Table 4. Measurment options in coffee survey

Measurement options	Gain/loss considerations
Sample (all) 200 trees on each farm Visit 1 farm in each district	A precise estimate of disease level on each farm but you only have one observation per district and therefore no idea of variation within the districts
Sample 20 trees on each farm Visit 10 farms in each district	Estimation of variation within each farm and also within each district. Comparison of the two is also possible
Sample 1 tree on each farm Visit 200 farms in each district	A good estimate of disease levels within each district but no idea of variation within each farm

Situation 2

Carry out a farmer-managed experiment to evaluate the yield potential of 4 sorghum varieties. You have 50 farmers who are willing to participate in the trial, but you are the only scientist on the project. The crop matures on all farms in the same 2 weeks (Table 5).

Table 5. Measurement options in sorghum variety trial

Measurement options	Gain/loss considerations
Visit every farm and carry out the harvesting yourself, avoiding the edges of plots, taking into account damaged plants and gaps in the plot, etc.	1. Time – you don't have enough of it!! 2. Should the researcher control the harvesting of a farmer-managed trial? 3. The assistance provided by you may give a bias to the farmers' perception of the varieties 4. Does harvesting the whole plot at the same time (which you would have to do) accurately reflect the actions of a farmer?
Take proxy measurements such as stand count and height prior to harvest	1. Requires less of your time and can be carried out before harvest 2. May not always be a good proxy for the crop yield
Ask farmers to harvest their own plots and provide you with sorghum yields for each plot, in kg/plot or as a score such as, 'poor' to 'excellent'	1. There is less time needed for you to interview the farmers about their yields and perceptions 2. Farmers maintain 'ownership' of the trial and the trial remains 'farmer managed' 3. You do not obtain a precise researcher-controlled crop yield, although you can use the farmer evaluations to answer the objectives of the trial

- What measurements do you need to take at each level of your design hierarchy?
- What resources do you have for your research – in terms of time, money and labour?
- What use of resources will give you the highest precision for your most important measurements?

When to take measurements?

One way to categorise your measurements is using the 'Before – During – After' approach. In any research project, experiment or survey, you can take measurements at each of these three stages.

Before

Measurements taken at the start of your research can:

- Provide you with a baseline for your experimentation, e.g., soil fertility measurements of a field-trial site
- Be used to characterise the plot/farm, e.g., wealth categorisation of farmers prior to their participation in an on-farm 'uptake of technology' trial
- Assist with your design, e.g., characteristics of regional farming population used to select a representative sample for participatory work.

During

You want to collect data on 'interim' responses whilst conducting your research:

- Common measurements on crop trials include plant stand and height. Other measurements may include labour use for different operations, insect pest and disease levels etc.
- You might opt to include the use of participatory research tools, e.g., participatory rural appraisal (PRA) during on-farm experimentation
- Whilst evaluating the use of agricultural information centres or extension offices you may choose to record the daily attendance numbers.

After

Towards the end of your research there may be follow-up measurements that can help you to complete your understanding of the results:

- You could measure soil fertility levels at the end of an on-station field trial, or farmer perceptions and technology uptake at the end of an on-farm trial, for instance, do they choose to continue using one of the tested technologies?
- Data could be collected to demonstrate the impact of your research, by comparing to your baseline data.

Data collection and quality control

Taking all measurements yourself will help to maintain high data quality but may not be possible if you are collecting data over many plots, farms and locations. If enumerators are used then it is harder to ensure they are using common methods, and you must monitor their performance and follow-up on difficulties and questionable data. Problems are especially likely to arise when carrying out socio-economic surveys or using PRA methods as these require considerable interviewing skills such as probing, performing arithmetic calculations to confirm responses are reasonable, and assessing the attitudes of the farmer. In this type of research it may be better to keep the sample size smaller and conduct the interviews yourself.

Field trials

- Ensure data collectors are trained in how to take each of the measurements. Give your enumerators a demonstration of the data-collection methods
- Monitor enumerators' performance by observing at least some of the data collection
- Check through the data as soon as it is given to you and follow-up on any problems with the enumerator immediately – before their memory fades
- Remember to take photographs of significant results or events, and label them carefully for later reference.

Surveys, interviews and participatory research

- Interviews and surveys should be kept as short as possible. If your questionnaire has, for example, 50 questions then it means you are collecting a lot of irrelevant information, repeating measurements and are not focused on your research objectives!
- When conducting farmer interviews for on-farm trials, socio-economic or biophysical surveys and other participatory research it is preferable to conduct the interviews in the field. Farmers will find it easier to evaluate and quantify their observations if they can see the crop/trial in front of them
- If you are using enumerators to assist in data collection then they must be appropriately trained. Meetings to discuss the survey measurements are useful so that you can agree on common methods of data collection. To standardise the perceptions/abilities of the enumerators you should carry out a few 'mock' interviews with farmers
- As with field trials, check through the data as soon as it is given to you and follow-up on any problems with the enumerator immediately.

Even if you have other scientists, technicians and enumerators working on your study you should be spending time in the field collecting data. Only by being personally involved can you monitor the quality and understand difficulties and things that are not going as planned. You will also gain insights into the problem that you would not get simply by analysing data that someone else collected.

Record ancillary observations, comments and notes along with all your planned observations. These can be anything. Include comments on data such as, 'plot 17 did not look so well weeded as the others' or, 'Mrs Njoroge was recovering from malaria when we conducted the interview'. Include notes on things to follow up, such as, 'I could see trees on the hill top but no one mentioned these in the interview' or, 'there were many more bees on the local varieties than on the introductions'. Include ideas that occur to you as you spend time in the field – 'Maybe we have to sort out insect damage before soil fertility will make any difference'. All these things will add to your interpretation and real understanding of the research you are doing. When researchers used pencil and paper exclusively their field books were usually full of such notes. Now computers are used to capture data, often all you will find are files full of numbers, with no marginal notes and comments. This is not very useful. Such information can easily be recoded in your computer, or you may still use a notebook. But if you do not write down the notes and comments when they occur to you, you will not remember them.

Photographs can also be an important record of the ancillary information. You can use them to illustrate the points you are making in your report and presentations. They will also help you recall the field situation, allowing more relevant and effective analysis and interpretation of numerical data.

How much data should you collect?

Often too much data are collected, just in case it might be useful. Researchers believe some measurements require little cost and therefore are worth taking. However, even if you have the necessary amount of staff/labour time some 'costs' may remain hidden. For example, staff asked to collect frequent growth data, which they know are rarely used in the final analysis, may not pay sufficient attention to data quality. Additionally, computer entry of the data may take up too much of the researcher's time. **When large volumes of data are collected it is often because staff have not yet given sufficient thought to the analysis.** Careful thought about what measurements are really needed to answer the objectives of the research should help to avoid this pitfall.

Measurement tools and equipment

The tools available for taking measurements are varied and you should use the ones most appropriate to the objectives and practicalities of your research. It is unlikely that your research will require only one type of tool and it is usually a good idea to combine such rigourous studies as quantitative field trials with PRA-type methods like farmer evaluations.

- For quantitative field data standard tools such as balances and tapes are used. **Ensure that you and/or the enumerators are fully trained in the use of each tool so that data collection is of a consistently high quality**
- When collecting data from farmers other tools will be needed. **Formal individual farmer questionnaires are often used but they are not always the most efficient method of data capture.** They are an intensive method of data collection and resources may limit the number of interviews you can do. Using a range of methods from PRA and other types of social enquiry may provide you with just as much information, for less work. For example, group discussions can be used to record a single consensus measure that is often useful when collecting baseline study data. **Use of PRA methods requires intensive training and quality control of enumerators.** Many PRA methods are designed for open-ended, exploratory enquiry. They can be very useful in more-structured investigations, but take care to use them in a consistent way
- Choose tools most appropriate to your research objectives and the practicalities of your work, i.e., available equipment and resources
- This may mean combining tools used in rigourous studies, such as the tapes for measuring tree heights with PRA-type methods like farmer evaluation of the tree
- Use quality control methods to ensure measurement tools are being used appropriately and accurately.

Resource material and references

There are very few books and papers written specifically to tackle the issue of measurements. 'Survey sampling' and 'Experimental design and analysis' books sometimes contain sections or chapters on data collection methods. There are also subject-specific books with details of measurement methods, some of which are listed below. You may find that scientific or discussion papers covering research topics similar to your work provide the best 'measurement' ideas.

Appendix 8. Presentations and Style – Tips on Photography and Writing. Eric McGaw. On CD.

Appendix 11. ICRAF. 2003. *Genstat Discovery Edition and Other Resources*. World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Ashby, J.A. 1990. *Evaluating Technology with Farmers: A Handbook*. CIAT Publication no. 187. Centro Internacional de Agricultura Tropical (CIAT), Apartado Aereo 6713, Cali, Colombia. 95 pp.

Ackroyd, S. and Hughes, J.A. 1981. *Data Collection in Context*. Longmans, London, UK. 155 pp.

CIMMYT Economics Program. 1993. *The Adoption of Agricultural Technology: A Guide for Survey Design*. Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT), Mexico DF, Mexico. 88 pp. cimmyt@cgiar.org

Coe, R., Franzel, F., Beniest, J. and Barahona, C. 2003. *Designing on-farm participatory experiments*. Resources for trainers. World Agroforestry Centre (ICRAF), Nairobi, Kenya. www.worldagroforestycentre.org

Feldstein, H. and Jiggins, J. 1998. *Tools for the Field: Methodologies Handbook for Gender Analysis in Agriculture*. Kumarian Press, Hartford, Connecticut, USA. 270 pp.

Franzel, S. and Scherr, S.J. 2002. *Trees on the Farm: Assessing the Adoption Potential of Agroforestry Practices in Africa*. CABI, Wallingford, UK. 208 pp.

Lal, R. 1994. *Soil Erosion Research Methods*. St Lucie Press, Florida, USA. 352 pp.

Philip, M.S. 1994. *Measuring Trees and Forests*. CABI, Wallingford, UK. 336 pp.

Spencer, D. 1993. Collecting meaningful data on labour use in on-farm trials. *Experimental Agriculture* 29: 39–46.

Schroth, G. and Sinclair, F.L. 2003. *Trees, Crops and Soil Fertility – Concepts and Research Methods*. CABI, Wallingford, UK. 416 pp.

Internet resources

- Reading Statistical Services Centre (SSC) website (http://www.rdg.ac.uk/ssc/) contains several downloadable booklets and papers. They provide 'easy to read' discussions and advice on various aspects of experimental and survey design and analysis. 'Measurements' are often discussed within these topics.
  Examples of useful information on the SSC website:
  N. Marsland, I.M. Wilson, S. Abeyasekera and U. Kleih (2000). A *Methodological Framework for Combining Qualitative and Quantitative Survey Methods*.

DFID Good Practice Guidelines:

- Guidelines for planning effective surveys
- On-farm Trials – Some Biometric Guidelines
- Centro Internacional de Agricultura Tropical (CIAT) website (www.ciat.cgiar.org)

Online publications:

Horne, P.M, and Sturr, W.W. (2003). Developing Agricultural Solutions with Smallholder Farmers: How to get started with participatory approaches.
TSBF Institute of CIAT (2001). Legume Cover Crop and Biomass Transfer Extension Leaflets

- Food and Nutrition Technical Assistance (FANTA) website (www.fantaproject.org)
  (Downloadable: Agricultural Productivity Indicators Measurement Guide)
- International Livestock Research Institute (ILRI) website (www.ilri.cgiar.org) – check the 'Capacity Strengthening – Training Materials' page for some on-line materials.
- International Institute of Tropical Agriculture (IITA) website (www.iita.cgiar.org)
  On-line publications (a few of these need to be ordered from IITA):

IITA Research Guides:

- IRG2 – Soil sampling and sample preparation
- IRG7 – Use of maps for planning research farms
- IRG9 – Morphology and growth of maize
- IRG18 – Farm records and work planning on agricultural research farms
- IRG31 – Tips for planning formal farm surveys in developing countries
- IRG50 – Socio-economic characterisation of environments and technologies
- TE/124 – A field guide for on-farm experimentation

4.6 Data management

Gerald W. Chege and Peter K. Muraya

- **'Data management' refers to all the steps in looking after and processing your data, from observation in the field until the end of the study, and after**
- **Attention to data management is important to ensure your observations are valid, they can be processed efficiently and will remain available for follow-up analysis at the end of your study**
- **Your project must have a data management strategy that describes procedures and responsibilities**
- **Computing will be an important part of a data management strategy. If your data are simple then spread-sheets may be suitable tools for data management. There are good and bad ways of using spreadsheets**
- **If your data are complex then spreadsheets will not be sufficient and you will need to learn something about database design and use**
- **Misunderstandings over data ownership can damage projects. Make sure all ownership issues are resolved before data are collected**

Introduction

Research work, irrespective of whether experimental or survey type, generates data. Data are the resources used by scientists to make conclusions and discoveries. As in other human activities, if you plan to use resources you need to take care of them, because lack of care may have disastrous effects. For example, a computer file containing medical data collected over a number of years could become corrupted. If there was no other copy elsewhere the total value of the resource would be wiped out.

Data management can be defined as the process of designing data collection instruments, looking after data sheets, entering data into computer files, checking for accuracy, maintaining records of the processing steps, and archiving it for future access. It also includes data ownership and responsibility issues.

Data management is important for the following reasons:

- **To assure data quality.** Since conclusions are based on data, accuracy is paramount and errors resulting from wrong data entry, incorrect methods of conversion and combining numbers must be avoided
- **Documentation and archiving.** Documenting or describing data and archiving it are important so that anybody can make sense out of the volumes of rows and columns of numbers for ongoing research and future use
- **Efficient data processing.** Scientists spend a great deal of time preparing data for analysis. This includes converting data to suitable formats, merging data sitting in different files, and summarising data from field measurements. The time spent in this pre-processing step can be greatly reduced if data are properly managed.

To see why data management is important, it may be worthwhile considering how organisations manage financial and accounting data. Whole departments spend huge resources on tracking transactions to ensure quality, on keeping records to document and describe those transactions, and on ensuring the records are available for future reference, to generate invoices, or to make payments and summary accounts. Specialist accountants are trained and hired to do this. Unlike accountants, scientists are expected to perform similar tasks with research data without the benefits of training.

The key steps followed in research data management are summarised in Figure 1.

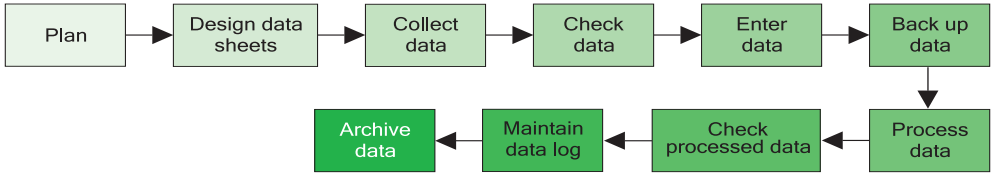


Figure 1. Data management processes

Planning for data management takes into account research objectives, resources and skills available. Appropriate field data recording sheets are designed. **Data collection** includes appropriate quality control. **Raw data** should be checked for errors. It should be entered into well organised computer files. **Captured data** must be backed up to safeguard against catastrophes. Data are processed for analysis, the results of which are checked again for any errors. Any **data processing** is logged to track data changes. Finally, data are archived for future reference possibly by other scientists.

After reading this chapter, we hope you will be better able to manage your research data. To appreciate the difficulties involved, some of the problems will be discussed. Such problems are both technical and people-oriented.

Technical problems include such issues as: lack of skills, lack of data documentation so future access is not straightforward, joint access for team projects, lack of proper design so as to meet data requests, incompatible data sets in cases where similar data are gathered at different locations or times, or files backed up on software that is no longer supported.

'Soft' or people issues include: time wastage in searching for data, re-processing old data sets, collecting data that had already been collected, and reformatting data.

Data capture

As shown in Figure 1, data capture is the activity that combines collection, checking, entry and saving data in some permanent electronic medium. You can get lots of help from specialists in carrying out the data management steps before and after data capture, but this is the one step you cannot shortcut and have to do yourself without much help. **It is the step that takes most of the resources (time and money) meant for the research; and that's why it is so critical.** The quality of the data processing that comes after this step will be determined by many factors, including which data you capture, how you lay it out, and which tools you use to do the job.

The tools

Some data capture needs can be sufficiently met by using word-processing software to publish the final results in a simple table. That's important, but it's not the main reason you enter data. It is to help you turn raw data into more meaningful results, an operation that is more difficult to archive with **word processors** than other software tools. **Databases** are the other type of tools available for data manipulation. However they are not in common use because their use is not intuitive for users who do not have much programming experience. Between the word processor and database extremes lie the **spreadsheets** that some people prefer to use for data capture because they are easy to use for data entry, **limited** manipulation and to display simple graphics. Here, the word limited is emphasised because the extent of the spreadsheet limitation is something that is under your control. Used without any discipline, a spreadsheet can be as severely limiting as a word processor; with discipline you

can use it to process your data with a flexibility coming very close to what you can achieve with a well designed database application.

Disciplined use of identifiers

Most of the data types that you will need to capture will be numeric, but these will very often need a few **non-numeric identifiers** for labelling rows or columns of numbers. Some people use long descriptive names as identifiers to make it easy for humans to understand and process the data. Others will use short often cryptic codes as identifiers, so that when the data are exported to other processing environments, the codes are very useful for formulating data manipulation commands. The majority of users use a mix of the two types of identifiers in an unplanned way, thus making it very difficult to understand them and limiting the extent to which the data can be processed either in or out of spreadsheets. Table 1 shows a spreadsheet that attempts to capture both types of identifiers and is laid out in a form that is easy to export for processing.

In this layout you will notice that:

Table 1. Using long descriptive and short coded identifiers in a spreadsheet

Plot identifier	Name of the village	Size of plot in square meters	Maize yield in kg	Is the plot infected or not?	Number of insects counted
plot	village	size	yield	infected	insects
1	Kesen	2.5	40.7	no	0
3	Sabey	2	53.6	yes	144
4	Sabey	5	50.7	no	0
4	Kesen	8	27.6	yes	107
6	Kesen	4.5	48.7	no	0
8	Sabey	4.5	37.7	no	0
8	Kesen	7	25.8	no	0
9	Kesen	1.5	40.4	no	0
10	Sabey	4	30.6	no	0
11	Sabey	3.5	24.3	no	0
12	Sabey	4.5	59.3	yes	35
13	Sabey	5	44.6	yes	340
14	Sabey	5	56.8	no	0
19	Sabey	5	62.1	no	0
20	Sabey	1.5	33.6	yes	489

- The long names are captured in single cells and formatted in word wrapping style – instead of the more common way of using multiple cells to break the label into small displayable chunks
- The short names are all alphabetic. Avoiding other characters or alphanumerics is a good discipline since most other applications will strip them out, or replace them with codes that may change the column names to something unexpected
- The short labels are entered on the last row, just before the numbers – allowing a **data export range** that excludes the long titles to be formulated and named.

Other descriptors

To further describe data sets, users will often go to great lengths to formulate folders and filenames that document the data. So a folder/filename like, /Western Kenya/Eva/*Striga*

research/2000.xls is not uncommon. There are two problems with describing data sets like this. The first is that you lose this description if the file is copied to another folder. The second is that the folder/filename structure gets very convoluted if you attempt to cram in all available documentation. One way to get round these problems is to enter these other documentations directly into the spreadsheet, rather than coding them into folders and filenames. Entering them at the header is less likely to interfere with other spreadsheet operations than anywhere else. A good example is shown in Table 3.

The body of a data set

Data identifiers will normally be few, placed at the top of a worksheet, and are needed to provide meanings to the values that form the main body of a worksheet. The quality of a dataset and its processing efficiency is determined by how much discipline has gone into the construction of the spreadsheet's main body. Here we look at a few tips that are easy to follow and that have profound effects on your data-processing efficiency.

One value per cell: when to create a new worksheet

One general rule for capturing and storing data is the concept that data in a single cell is **atomic**, i.e., only one data item occupies one cell. It is difficult to analyse multiple data entries per cell. The solution is to create another sheet with repeating data and to link it to the first sheet. For example, suppose variable Q1 in the variable set Q1, Q2, Q3, ..., Qn, has repeating values as in Figure 2a.

a.

Qno	Q1	Q2	Q3	Q4 ...
1	2, 6, 7	5	10	20
2				
3				
4				
.....				

b.

Qno	Q1
1	2
1	6
1	7
.....	

Figure 2. One entry per cell principle: a. Error in column Q1, b. Solved by another sheet for Q1

The circled entry is in error. The solution is to create another worksheet for Q1 as shown in Figure 2b. The two worksheets are then linked, using special formulae in Excel, e.g., vlookup(...) that are more difficult to use than exporting the data to a database package like Access.

Data body: row consistency

In the body of a spreadsheet, all the rows should represent the same entities. A further addition to this rule is that each row should be so completely filled in that sorting the body in any way should not result in the loss of meaning for that particular row. Table 2 shows a data set that clearly violates this rule. It is clear that:

- Line 9 represents a new sub-plot heading, and not crop row measurements as in all the prior rows. If you sorted the rows using some order, say ascending total grain fresh weights, you would no longer be able to tell which cropping rows came from which sub-plot. The solution is simple: create a new sub-plot column and fill it accordingly. The other

Table 2. A spreadsheet body with inconsistent row entries

1		Sub-plot 1				
2	Block	Maize plot	Row	Cropping system	Total fresh grain weight (g)	Total fresh cob weight (g)
3	1	1	1	control	2291.0	528.0
4			2	control	1156.0	228.0
5			3	control	871.0	199.0
6			4	control	missing	missing
7			5	control	505.0	88.0
8						
9		Sub-plot 2				
10			1	control	571.0	147.0
11			2	control	564.0	132.0
12			3	control	430.0	113.0
13			4	control	188.0	108.0
14			5	control	649.0	236.0
15						
16		Sub-plot 3				
17			6	control	861.0	201.0
18			1	combi	lost	lost
19			2	combi	381.0	121.0
20			3	combi	536.0	143.0
21			4	combi	438.0	140.0
22			5	combi	617.0	169.0

solution of moving the second part to a different sheet is not recommended because you lose the integrating effect that allows you to analyse the data set as a single unit.

- Lines 8 and 15 are blanks, which represent entities that are different from the prior rows. The user may have inserted them for some sort of clarity, and not to indicate the end of a data set range, which is how Excel would interpret them. So, if you used your sort function, only the top part, up to the blank row, of your data set would be sorted, which is probably not what you intended.
- The case for data lines 4–7, 10–14 and 17–22 also needs attention. Without rearranging these data, it is clear to us what the implied values are in the blank entries. But this would no longer be the case if the data were sorted. This is a very common problem when users try to make a spreadsheet look exactly like the paper forms. The solution, of course, is to fill in the implied values.

Data body: column consistency

The column entries in the body of a data set should all be the same type of data. This is important to prevent errors during data conversions using some formulae, or during data exports. Some entries in Table 2 violate this tip. Note that 'missing' and 'lost' values for Total fresh grain weights in rows 6 and 18 are text data types which data processors would understand differently from the numbers. What you should put in these cells depends on the software that will be used to further process these data. For instance, for Genstat you would use the star (*); for SAS you would use the dot (.) In some cases Excel would treat these labels as 0, resulting in an incorrect result when the columns are used in calculations. Our

recommendation is to leave the columns blank. Should you need to explain further why no value existed, use the comment feature. Unfortunately, comments are ignored when data are exported to environments outside of Excel, thus limiting their usefulness.

Putting it all together

Table 3 shows a spreadsheet whose preparation has considered most of the tips given in this section. It does not matter that this one has been designed for experimental data; the same tips are applicable to survey types of data.

Table 3. An example of a spreadsheet design that uses some of the tips discussed in the previous sections

Program	Domestication of Agroforestry Trees					
Project	Genetic Resources of Agroforestry Trees					
Experiment	Leucaena family trial					
Location	Kenya, Muguga					
Investigators	James Were, Tony Simons					
Start Date	5/1/1996					
Statistical Design	Incomplete block design					
Assessment	Tree growth					
Date	Feb-97					
Replicate number	Blocking id	Plot number	Leucaenea family id	Tree identifier	Tree height (cm)	Number of stems
rep	**block**	**plot**	**family**	**tree**	**height**	**stems**
1	1	1	20	1	214	4
1	1	1	20	2	252	6
1	1	1	20	3	153	2
1	1	1	20	4	183	4
1	1	2	18	1	98	1
1	1	2	18	2		
1	1	2	18	3	201	3
1	1	2	18	4	192	1
1	1	3	9	1	232	8
1	1	3	9	2	201	7
1	1	3	9	3	198	4
1	1	3	9	4	152	4
1	1	4	10	1	175	2

Data entry and validation schemes

A well designed experimental data sheet is ideal for data collection. If the design is done before field data collection, it should be printed and used by all those who are collecting data in the field. Data entry should be done as soon as data collection is complete so that any clarifications are sought while people's minds are fresh. Some initial checks should be done for obvious errors. Missing values should be carefully treated. Ensure non-available data appears as blanks in the worksheet (not as zero since zeros are included in statistical calculation). You can use a number of techniques to aid data entry and avoid transfer errors.

During data entry, and especially for long or wide lists, it is a good idea to be able to see column and row headings all the time even as you scroll through the worksheet. This is

achieved by freezing or splitting the window panes. In Excel this is achieved by selecting **Window ➠Freeze Panes** (or **Window ➠ Split**) when the row below the column heading is selected. To remove this effect, select **Window ➠ Unfreeze Panes.**

Validation during data entry

Data should be entered quickly and in raw form to minimise the chances of making errors in transcription. You should enter all the data. Partial data entry that can be quickly analysed is not recommended. If you enter it all, you can cross check during entry to minimise errors.

To identify data records, each field should be unique, for example, plot number. Derived data should not be entered. Since computers are good at calculations, you are well advised to simply enter primary or fundamental data, thus avoiding possible errors caused by hand calculation. For field experiments, it is a good idea to enter the data as they appear in the plots (for logical mapping to the physical plot) or to use two columns to identify the location of the plot – (x, y) coordinates using some reference frame.

Drop-down lists can be used to avoid typing a sequence of data more than once and to avoid typing errors. A drop-down list is a set of data (such as crop names) from which you can chose one for entry into a cell. This is created by highlighting the data set, and selecting **Data ➠ Validation➠Allow:➠ List**. For **Source** of list give the range of the data set.

Data validation is also done on a range of cells. When new data is entered in cells with range checks, any data values outside these ranges generate error messages. For example, if values for a variable *wheat % moisture* is in the range 13 to 29, you can set the range check by highlighting the data area for that variable, then select **Data➠ Validation ➠ Allow: Decimal** and set **Minimum** as **13** and **Maximum** as **29**.

Validation after data entry

Scatter plots can be used to spot data outliers once data have been entered. These are data values considered outside the allowed range and easily seen from a plot (see Figure 3). Line plots can also be used to spot outliers.

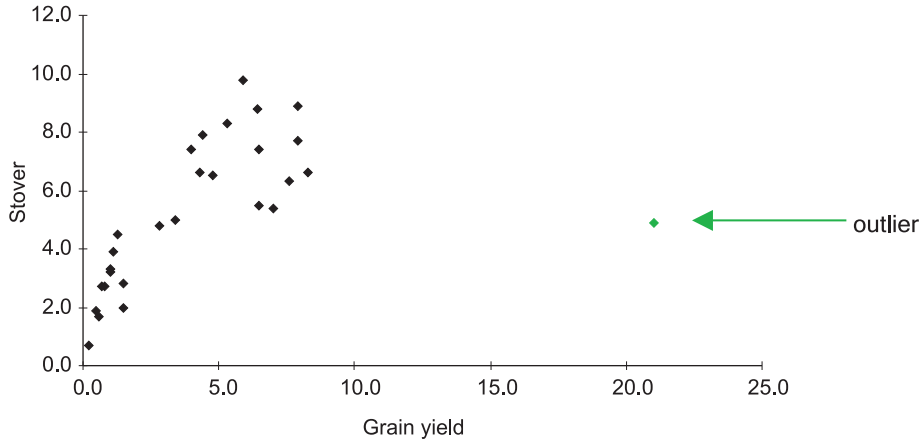


Figure 3. Scatter plot example showing data outlier (for details see Appendix 11)

Adding comments to cells

Unusual instances occur in data capture and subsequent entry. For example, when no value is recorded for a variable, it is a good idea to indicate the causes of that anomaly. This can be done by inserting comments in the worksheet using **Insert ➠ Comment** on the subject cell.

Data auditing

For already existing data, the auditing tool allows you to check for some errors. Auditing is created by: **Select Tools ➠ Formula Auditing** ... and then click on **Show Formula Auditing Toolbar.** On the toolbar, click on the icon **Circle Invalid Data** (second from right). This draws a ring around invalid data.

An illustration of auditing, with validation rules for both ***species*** and ***rcd*** is shown in Figure 4. In the figure errors are circled for the variables ***rcd*** and ***species.*** All cells in a column should have the same data type. In the case of column **D** (Figure 4), the data type is numeric. The string **'DEAD'** in cell **D10** is therefore inappropriate. The entry **12.7, 13.3** in cell **D2** is also invalid (it is ringed) because of two decimal number values and so is **198** (in D21) because it is out of range. In cell C19, **A. *polyantha*** is wrongly spelled and hence ringed. You can see how the auditing tool helps you to spot data entry errors before processing.

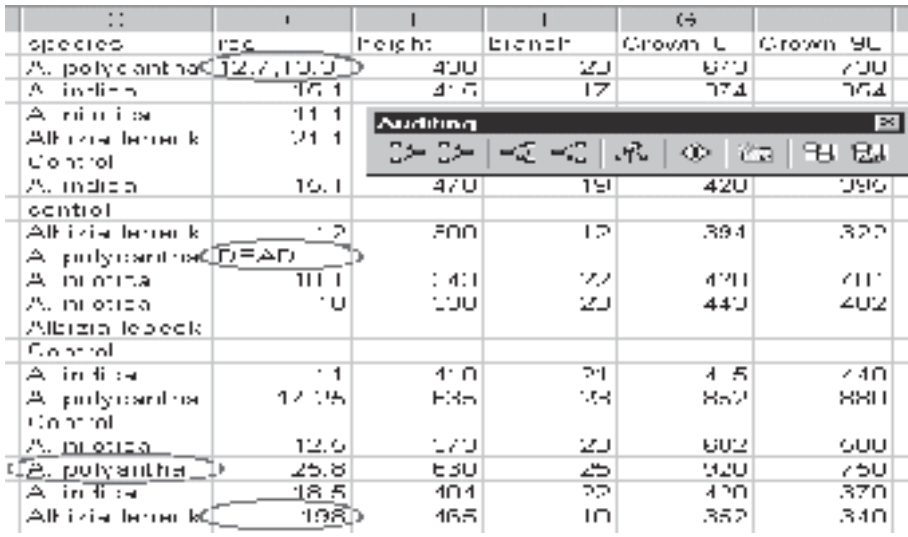


Figure 4. Auditing of existing data (see SCS–University of Reading: Disciplined Use of Spreadsheet Packages for Data Entry)

Saving and protecting files

Once data have been entered, it is important to use a good **file naming** and **saving strategy** so you can easily refer to the same data in the future. Files and where they are stored (directory structure) should be named with sensible names that suggest the research type. Choose a directory structure that best describes the structure of the files. Data files should not be mixed with program files. A common scheme is to use the **directory \USER** for storing all data files. Each research location should have its sub-directory and each experiment within that location should have its sub-directory. For example, 'Eucalyptus tests' at

'Kakamega' might have data stored in the directory \USER\KAKAMEGA\EUCALYPTUS. All files related to this experiment would be stored in that directory. Such files might include field data, reports, charts, documentation and statistical results for the Kakamega eucalyptus tests.

It is also a good idea to document the workbook using Excel's *summary information* which shows the *title, subject* and *keywords* for the workbook. This is particularly important if many workbooks are likely to be used. Summary information is achieved by clicking on File ➠ Properties.

It is important to **save** and **backup** your data regularly. Hard disks get corrupted for a variety of reasons. The computer could get stolen, burned or simply fail to work. Backing up can be done on diskettes, tapes or CD-ROMs using a CD writer. For diskettes, keep at least two sets (a copy may take several diskettes) in different places besides the one on hard disk. To avoid having several versions of the same data, it is advisable to create a **master copy** (with the same file name) which is updated and backed up every time there is a change of data, and keep it away from the computer.

It is a good idea to **protect data files** from unintended changes and when they occur, to keep track by highlighting them. This is achieved in Excel by selecting: Tools ➠ Track Changes or Tools ➠ Protection (this allows you to set up rules for file protection).

When to use an advanced tool

MS-Excel is a powerful tool when used properly with single data files. When multiple data files are used, it becomes difficult to maintain data in those files. Often you end up with several files of the same information, and it is hard to keep them consistent with each other. Querying these files is not easy and becomes cumbersome and inefficient if you have several of them.

Relational databases were designed specifically to handle many related data files and are optimised to allow efficient data querying, ensure data integrity and sharing of data between different individuals when necessary. A database allows you to have data of the same subject in a table, and then, assuming there a number of different data sets, have each on a different table in a database. Because research is usually on related objects, the next step is to relate the different objects. This ensures **referential integrity** so that changes to a data item are made through a controlled environment. During your analysis phase, you should use a statistical tool like SPSS, **Genstat (Appendix II)**, or SAS. The database data can be exported to the specific analysis tool of your choice.

Designing a database

If you are handling several related worksheet files, the logical step to take is to use a database for these data. The key issue in a database approach in capturing research data is the initial design of the base tables. Each data object (called **entity** – like a plant), has characteristics that define that object. For example, a plant has leaves, height, maturity stage, shoot size at a given time, description, etc. These are the **properties** of the plant object. In Access-speak, the plant entity's properties are the **fields** or **attributes** and each has a **data type**. (sample data types are integers – identified as **integer** and **long integer**, decimal point numbers – identified as **long** and **double**, and **text**, amongst others).

Database design means defining each of these attributes with their data types, selecting one or a combination of attributes as a **unique identifier** for the plant object (called the

primary key) and then repeating this process for other entities. After this, you can create **relationships** between the different entities. For more complicated databases some real design work is necessary (which includes **normalisation**). An example is the relationship between an employee and her dependants shown in Figure 5 (both employee and dependant are entities). This type of relationship is called one-to-many – one employee can have several dependants.

Figure 5. Relationship between an employee and her dependants

Once the fields for each entry are choosen you can define a table to hold the data. The table design screen in Figure 6 shows the design of a person-level table. Names for the fields and their data types are defined. Once the table is created you can enter data via the datasheet or the spreadsheet view. This is shown in Figure 7. The datasheet resembles the Excel worksheet.

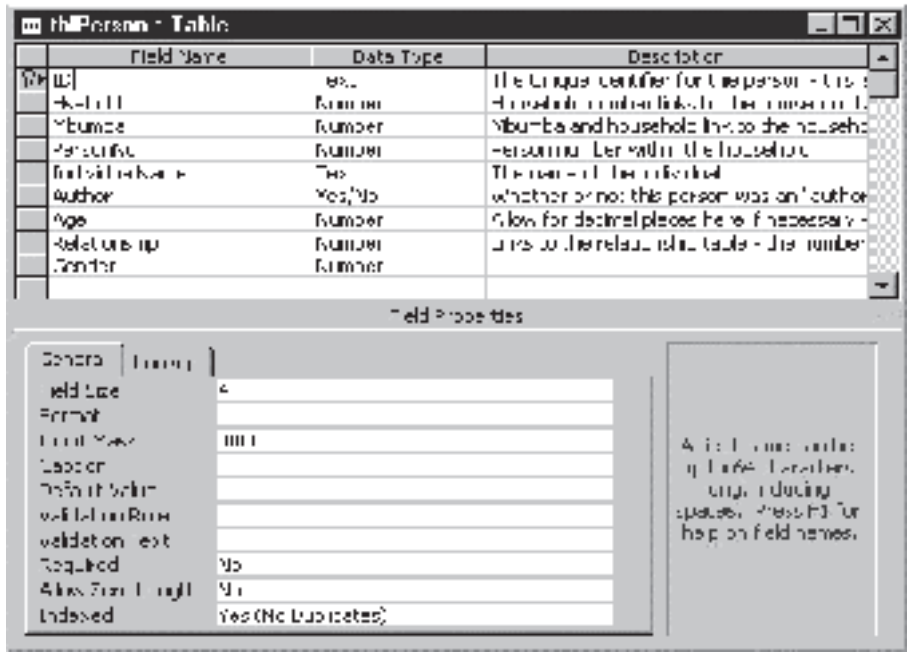
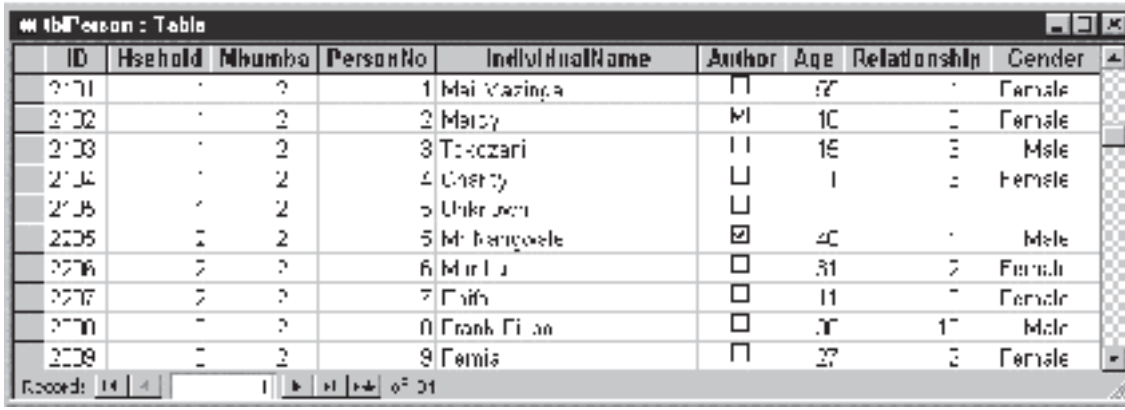


Figure 6. Table design in MS-Access for a person-level table

As with the use of a spreadsheet, it is important to use a database package '**with discipline**'. With minimal discipline, defining the number of fields and their data type is enforced, but you should normally do more than the minimum.



tblPerson : Table

ID	Hsehold	Nhumbo	PersonNo	IndivlNualName	Author	Age	Relationship	Gender
2101	1	2	1	Mai Kazinga	☐	52	1	Female
2102	1	2	2	Mercy	☑	10	2	Female
2103	1	2	3	Tsvezari	☐	15	3	Male
2104	1	2	4	Cherry	☐	1	2	Female
2105	1	2	5	Unknown	☐			
2205	2	2	5	Mr Nengoshe	☑	45	1	Male
2206	2	2	6	Martha	☐	31	2	Female
2207	2	2	7	Faith	☐	11	2	Female
2208	2	2	8	Frank Tinson	☐	30	1	Male
2209	2	2	9	Temis	☐	27	2	Female

Record: 1 of 31

Figure 7. 'Datasheet' view of person-level data

Selecting data for analysis

There are two alternative ways of getting Excel data into a database. The first is by **importing** it into the database. This leads to two copies of the same data set and can be a major source of data inconsistency when changes are made in the database but corresponding changes are not made on the worksheet. A good practice in data management is to designate either the worksheet or the database the **master copy** so that data changes are only done on the master and all data sub-sets are extracted from the master for analysis.

An alternative to importing Excel data into a database is **linking**. Here, the database and the spreadsheet use the **same datasheet copy.** Changes made in the database or in Excel are reflected in this copy. **This option is the best in terms of data management.** You will have no worries about managing multiple data sets if you use this method.

Using queries for calculations

Databases are designed to store **fundamental data. Computed data** is not fundamental since it can be derived from other data. To get computed data in databases, you can use a **query** and create a new field where a formula for the derived data is entered. For example, to compute the average height of a tree in a certain experiment, assuming individual heights are stored in a field called height, you would enter the following in a blank field of the QBE grid – **avheight:sum([height])/count([height]).** Once computed, this field cannot be updated manually with data, unlike in Excel where a formula can be overwritten with data.

Using queries to select records

In Excel, you can select rows of data by using **Data ➠ Filter** and specifying a criterion. This is quite simple but the resulting rows of data must be copied to another sheet before use. Both the filtering and copying of results are manual processes and prone to error, and the results are a duplicate of the original data. As observed, creating copies of the same data is a bad data management practice. In databases, you can get the data sets you want by creating a query and setting criteria (e.g., all trees with a height between 20m and 40m). It is possible to have complex queries using **and** with **or** logic combinations. The important point is that the **query** is stored in the database (not the results) and you simply execute it whenever the dataset is required. Similarly, queries can be created to select fields, link

multiple worksheets so that data sets can be extracted from all of them, and also to check data validity.

Data archiving

Archiving is the process of storing data for future use. The user of archived data is not necessarily the person who did the experiments, or carried out a survey. Indeed a well archived data set can be used by others to derive new relationships in the data or to compare primary data with secondary data. Funding agencies may even be attracted by the possibility of archiving data from the findings of a proposed project.

The process of archiving data requires three basic principles:

1. The data about the project rather than the results of the study itself (sometimes called meta-data – description of the data itself) should be archived.
2. The description of why data was collected should be archived.
3. You must archive a description of the data files – their types and structure.

The latter makes future retrieval easier. The first point makes it possible to easily understand the rationale of the data-collection exercise, while the second gives additional information on the procedures and processes of data collection. This means a future researcher would be able to replicate the experiment or survey for scientific validation of the findings. Data files need to well structured, in the majority of cases they are computer files.

Backups of computer files should be made regularly with a strategy of keeping a master copy far away from the archival site. This ensures continuity in case of natural hazards like fires, floods or earthquakes.

To summarise, a good data archive should be/have:

- **Accessible:** hence easy to access by many users who have commonly available software
- **Easy to use:** so that the field data collection forms, and what will be entered into the computer are similar
- **Clearly defined variables:** the units of measure and codes used (labels for names of variables) should be as clear as possible
- **Consistency:** of names, codes, units of measurement, and abbreviations
- **Reliable:** archive should be as free from errors as possible
- **Internal documentation:** documentation should be complete with regard to: procedures for data collection, sampling methodology and sampling units used; the structure of the archive (how different files are related); a list of all computer files in the archive; a full list of variables and notes on how to treat missing values; summary statistics for cross-checking the information in the archive; and any warnings and comments that need to be observed for data usage
- **Confidentiality:** ensure that the data remain confidential if this is required by the sources
- **Complete:** if possible you should store copies of: the data capture field forms; the data management log-book; a description of derived/calculated variables.

Storage and access to the archives is also an issue for you to consider. A good archive includes information on how to get into the archive with rules of use and replication to other third parties. The **storage medium** for computer archives is in most cases, **hard disks**. With the new pervasiveness of the Internet, access to the archives is mainly by downloading archive files. This is true for different types of data including text, graphics, maps, photos, and audio and video material. Using **diskettes** and/or **CDs** as **distribution media** is of course an option. The other medium is good old **printouts** sent by mail for long-distance access.

To show the seriousness of data archiving and its place in research, it is now possible to publish peer-reviewed **data papers** (http://esa.sdc.edu/Archive/E081-003/main.html).

Data ownership issues

Any data set must belong to somebody. **Ownership** in the wider context means who can access data for reading only, updating, deletion or creation. In the scientific community, a data set does not necessarily belong to the individual who generated it. It generally belongs to an institution that can ensure continuity through hosting the data and providing access to it to individuals, or other institutions.

If data are generated by more than one institution, for example, through collaborative research, then they belong to the participating institutions. If data are generated by publicly funded projects, then they are public property, held in trust by an institution for the public. There is need to recognise **intellectual property** for scientists who may generate data. If scientists generate data using public funds, they must use the data for the purpose for which it was intended.

There are data ownership issues that need to be agreed on by several different parties. Institutions have an obligation to give **public access to data**. They do this by adopting some policies and procedures that must be communicated to all the interested parties.

It is important to balance the rights of individuals who collect data with the need to ensure future public access to data. Two approaches are used:

1. **A time limit** is set beyond which a scientist can not claim ownership to data. For example, 1 year for field research is typical, because this gives the scientist some time to carry out the analysis and publish before the data become public property.
2. Data owned by an institution may be released to a researcher if a good case is made for that access. Acceptable reasons may include: to check analysis, to improve on analysis, to correct an analysis, to analyse new questions using the same data, for integration with other data, and for meta analysis.

The **subjects of a data set**, for example, the people interviewed or the farmers who took part in an experiment, also have some rights. These may be set by common values or may be determined by law. The minimum all subjects can expect, and that all researchers should ensure, are:

1. That all personal information will be kept confidential.
2. That the data will only be used for the intended purposes, and the people concerned agree to this before participating.
3. That results of the study are made available to all people participating.

Data management strategy

A **data management strategy** is a set of policy guidelines developed for an institution to help it cope with different facets of data management. A data management strategy requires the following:

- **Commitment.** This is important for all stakeholders of a project including project managers and researchers. It is a pre-requisite for a successful data management strategy since it will enable commitment of resources to develop and maintain the strategy
- **Skills.** These are required by all players to do what is necessary for the data management strategy. They include data entry skills for junior members of the team, form design skills, data entry and validation skills, and data archiving skills amongst others

- **Time.** This is necessary for a good output. Funding agencies have recognised that besides the final research output (analysis results), data is also an important output achieved by archiving. Clearly enough time ought to be devoted to data management in the overall research timeframe
- **Financial resources.** It is important to include data management within the project proposal, otherwise the necessary tasks it involves will not be done.

Key components of a strategy

The four key components of a strategy are:

1. **Transformations and their products** – these are the steps in research data management.
2. **Managing meta-data** – the process of defining and managing descriptions about the data.
3. **Data management plan** – this is the overall plan of the strategy and how steps in the strategy can be measured for performance.
4. **Data management policy** – these are principles that guide structure and contents of meta-data and the strategy plan.

1. Transformation

This describes the entire data management cycle starting from problem definition, formulation of research objectives/hypothesis, development of data capture tools, data entry using some validation rules, selection of data for analysis, the actual data analysis, management of results and finally publication of findings. This is a cycle because it is possible to go back to any point in the process in case there are errors.

2. Managing meta-data

Meta-data is a description of the data to be handled in a research project. It can be used to describe data sets, enable effective management of data resources, and to enable other researchers to understand the data sets of a project.

The key areas of a meta-data are:

i. **Title.** The name of the data set or the project
ii. **Authors.** Names of researchers (principal researcher and others) with addresses, phone, e-mail, and web contacts
iii. **Data set overview.** Introduction to the data set, location of data, time of experiments/survey, and any references
iv. **Instrument description.** Brief description of data capture instrument with references
v. **Data collection and processing.** Description of how data were collected, computed values, and quality control procedures
vi. **Data format.** Structure of data files and naming conventions, codes (if used), data format and layout, version number and date.

Meta-data description must be done for every project.

3. Data management plan

A plan shows how data will be recorded, processed and managed for the duration of the project. It includes roles for staff, back-up procedures, quality control checks and how to handle errors, procedures for managing the data management strategy e.g., discussions in meetings, procedures, software upgrades, methods of creating archives, and how the archive will be maintained.

4. Data management policy

These are policy statements that guide data management to ensure consistency. They are high-level objectives of data management. A typical policy consists of the following objectives:

- Establish and distribute high quality data sets
- Standardise quality control procedures
- Ensure data and other project materials are archived and reviewed regularly
- Reduce time between data collection and analysis
- To maintain data securely
- Facilitate data access and usability through improved meta-data.

The **policy** includes the **roles and responsibilities** of individuals in the project, specifies the data owner, the data custodian (a manager in charge of the data management process), data user (individual with access rights to the data), security administrator and an information systems group.

Conclusions

Data management in research is very important. It is the entire process encompassing project initiation, through all the phases up to the time a paper is published as a result of that research. For quality results, all the phases as described in the strategy above must be managed according to the stated principles. The scientific community now accepts **data papers** for publication, in addition to the traditional research output documents. Data archives are a rich source of information so your data should be archived following the guidelines discussed above. These should give you enough reasons to embrace research data management and practice it. After reading this chapter, you should have sufficient information to better manage your research data.

Besides the referenced material, additional sources of relevant information are provided below.

Resource material and references

Appendix 9. Muraya, P., Garlick, G. and Coe, R. 2003. *Research Data Management.* World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Appendix 11. ICRAF. 2003. *Genstat Discovery Edition and Other Resources.* World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Ecological archives. A *Publication of the Ecological Society of America*: (http://esa.sdc.edu/Archive/E081-003/main.html)

Coe, Richard. 2001. *Audit Trail in Research Data Management.* (Draft Report), World Agroforestry Centre (ICRAF), Nairobi, Kenya.

Coe, Richard. 2001. *Issues in Data Ownership and Access.* (Draft Report), World Agroforestry Centre (ICRAF), Nairobi, Kenya.

Statistical Services Centre (SSC), University of Reading, UK. http://www.reading.ac.uk/ssc.

Case Study No. 6 – Good practice in data management (2000)

The role of a database package for research projects (November 2000).

Project Data Archiving – Lessons from a Case Study (March 1998).

Good Practice Guidelines (2000).

4.7 Analysing the data

Susan J. Richardson-Kageler

- **Analysing the data means turning the raw observations into summaries that can be interpreted**
- **Appropriate methods for analysis depend on the objectives, the study design and the nature of the observations**
- **Exploratory and descriptive analysis, that displays the main patterns in the data with summary tables and graphs, will allow you to tentatively meet many of your analysis objectives**
- **Formal or confirmatory analysis will add information about uncertainty and allow you to disentangle complex patterns**

Introduction

This chapter summarises the key points to look out for when you analyse your data, and the dos and dont's. It is assumed that you have taken a statistics course at some stage. If you need more details of the statistical techniques that are used in this chapter, then refer to the relevant text books (see Resource material and references at the end of the chapter for some of these texts).

Students often get stuck when they have to start analysing their data. This can happen to you for a number of reasons. Perhaps you are scared of the analysis stage. You have approached the research confidently, enthusiastically collected the data and now find that you do not know how to proceed with the analysis. Sometimes the problem is made worse because you have left the analysis till the last moment. The analysis should not be considered the final stage in the research process, but should be done as soon as data become available. A good researcher will think about the analysis at the research proposal stage. In the research method section of your study proposal you should include an analysis plan with descriptions of the possible tables, figures and methods you will use. This will help you to think about analyses early and will ensure that the analysis will be possible and will meet your objectives.

Descriptive methods are more important in real analysis than their emphasis in many statistics courses would suggest. Often the most elegant analyses and main results are obtained from well thought out summary tables and skilfully designed graphs. They require little more than common sense and a clear idea of what you are trying to find out. So, even if you are unsure about formal statistical methods, you should be able to start your analysis.

The following are the usual steps in the analysis of data that are considered in this chapter:

- Define the analysis objectives
- Prepare the data
- Descriptive analysis
  - Tables
  - Graphs
  - Summary statistics
  - Identify oddities
  - Describe data pattern
- Confirmatory analysis
  - Adding precision
  - Improving estimates

- Interpretation
  - Understand the results
  - Combine new and old information
  - Develop models
  - Develop new hypotheses

As you progress through your research and analyse the data as you go along, you will find that the data analysis is **iterative**, this means it is not a simple matter of following straight through the process outlined above. You will need to stop and revisit previous steps as new information is discovered. Even though you analysed the data as you progressed through your work, you may need to re-organise your data and reanalyse, so you must leave plenty of time to complete the analysis and write up the results after data collection. Remember, it always takes longer than you expect to get the tables, figures and analyses compiled and written up effectively.

In this chapter, two examples are used. The first is a survey which investigates farmer's perceptions to, and use of, planted fallows. A questionnaire was administered to 121 farmers who had experience with planted (improved) fallows grown with or without rock phosphate fertilizer in Western Kenya.

The second example is an experiment to evaluate whether the pumpkin (*Cucurbita maxima* L.) variety Flat White Boer can be used as a smother crop when planted at the same time and intercropped with the long-season maize variety PAN86 at the University Farm, Mazowe, Zimbabwe. This evaluation is done by comparing sole maize, sole pumpkin and a pumpkin–maize intercrop.

Analysis objectives

Two key points for this section:

- Analysis objectives are determined by, but more specific than, the overall research objectives.
- Analysis objectives will evolve during the research as you gain insights and experience.

When the research proposal is put together at the beginning of the research, it is usual to state an **aim** and a set of **objectives** in the introduction. **Analysis objectives** often need to be stated separately from objectives of the research work. This is because the analysis objectives are dealing with the **specifics** of the analysis. The original objectives may appear vague in comparison as they often do not specify precisely which variables are to be analysed and how they are to be processed. The analysis objectives will determine such specific things as:

- What the relevant variables are and to which level they are summarised (see later in this chapter in the section on Preparing for analysis)
- The specific comparisons that will be made
- The relationships between variables that will be investigated.

For the pumpkin–maize intercrop example, the analysis objectives include:

- Comparing the maize grain yield between sole maize and pumpkin–maize intercrops
- Comparing the weed density of the sole maize, sole pumpkin and pumpkin–maize intercrop
- Determining how the maize yield depends on pumpkin and weed cover.

The analysis objectives should be refined as you proceed through your data collection. Your experience in the field will give you ideas and insights you did not have when you planned the research. It is a good idea to keep a notebook and record your observations and

ideas as you go along. Things often happen for which there is no place on your data entry field sheets. You often find that when you come to complete the writing up that you cannot remember these important things that have occurred during the research process. Even go so far as to keep a notebook beside your bed at night. This will help you to sleep better as you can write down the things that you think of, and so you won't have to keep yourself awake to make sure that you remember your ideas in the morning! Examples for your notebook include the fact that one of your test animals broke out and ate your neighbour's vegetables, or ideas for more informative graphs and data summaries.

Preparing for analysis

Some key messages for this section include:

- Data must be well organised – see **Chapter 4.6**
- Data must have been checked – see **Chapter 4.6** – but remember that more mistakes will become apparent as you do the analysis
- Some data preparation is necessary once the analysis objectives are clear
  – Summarise to the right level
  – Use suitable format for the software.
- The preparation of data for analysis starts with the project proposal and ultimately fulfils the objectives of the research. From the start, think through how the data are going to be entered onto the computer and which software packages are going to be used.

The stages involved in the preparation of the data for analysis are:

- Raw data entry and checking
- Organisation of the data to the form needed for the analysis to meet the objectives
- Archiving of the data so that it remains available.

Once the data are entered, check the data entry using simple data analysis. This includes transforming and plotting the data, summaries which show extreme values (minimum and maximum values, trimmed means), boxplots, scatterplots, tables of the data in treatment order, frequency tables of coded data, ANOVA and plots of the residuals.

After checking the data entry, or sometimes before it is entered, the data must be summarised to the appropriate level for data analysis. To do this, you will need to recognise the correct data structure. You need to decide at which level you will do the analysis to satisfy the objectives of the study. If data have been collected on several individuals per house, do you want to analyse the data about individuals or about households? If data have been collected at sub-plot level, for example, 10 plants have been randomly sampled from each plot, you may need to calculate a summary statistic for these 10 plants to be used in an analysis at the plot level. The data may be complicated and involve many sites over several districts. Some of the objectives may be fulfilled by summarising data at the site level, clustering similar sites into groups and then comparing the sites across these groups. A survey on disease in coffee involved many districts, farmers and trees. The huge data matrix looked like a nightmare, yet the first objective was satisfied by simply calculating the proportion of farms in each district that had the disease. The next objective required calculating the same thing for two different coffee varieties.

You may also need to calculate new variables from those you have measured. For example, you might have measured fresh weight and moisture content, but later discover that you need an analysis of dry weight. The whole process of understanding your data structure and deciding on the summaries or variables or units to be used needs to be continuously

related back to the objectives of the study so that the analysis does not become misdirected. You may need to revisit this step again after carrying out part of the analysis as you may realise then that the summaries you have calculated may be inappropriate, or some analyses have indicated patterns that will be best examined at a different level or with different variables.

The results of this stage are data sets that are in the correct form to answer the research objectives. If this stage was not done in a statistics package, then the data should now be ready for transfer to a statistics package for analysis. Software for handling data entry, modification and analyses should all be **compatible** and includes:

- Database management software
- Spreadsheets
- Statistics packages
- Word processors.

Compatibility means that you can move the information from one package to another. For example, you may want to add your data to your final report as an appendix. You should be able to copy the data into a data-entry package such as Excel, and paste them into the word-processing package in which you are writing your project report, e.g., Word. Further, the graphs generated in a statistics package such as Genstat or Minitab, can be copied from the statistics package and pasted into your project report in the word-processing package.

Exploring and describing the data

Important points to remember are:

- Descriptive analysis uses tables and charts of summary statistics to show the main patterns in the data
- It also reveals unusual or surprising observations
- Preliminary reports and conclusions can be based on the descriptive statistics.

The aim of exploring and describing the data is to find out what the data has to tell you. The data can be split into two parts:

Data = pattern + residual

Pattern is the underlying structure or shape of the data, in which your primary interest lies. Knowing the pattern should mean that you satisfy the objectives. An experimental pattern is often the result of the treatments that you have applied. The pattern is summarised by descriptive statistics, e.g., the mean of the treatment. **Residual** is the remaining, unexplained variation. There should be no pattern in the residual part of the data. If there is pattern in the residual part of the data this indicates that some effect has been forgotten, perhaps due to the layout, treatments or measurements. **The ultimate aim of the data analysis is to describe the pattern.**

Data analysis starts by exploring and describing the data. This is the point at which you begin to understand what is really happening. When this step is carried out effectively, you can make subjective conclusions about the research and write a preliminary report. Students sometimes forego this step and go straight to the confirmatory analysis (see the next section). The **confirmatory analysis** in agricultural research often includes an **analysis of variance (ANOVA).** ANOVA can be used in a variety of circumstances: both as a descriptive tool and in inference where it is used to identify which parts of a model are important.

Example

One student who went immediately to an ANOVA missed out the most important finding of his two and a half years of research. He was investigating the effect of different diets on the fat in ostrich meat. He collected the data and then carried out an analysis of variance. He did not plot any graphs, calculate any summary statistics, or check the residuals resulting from the analysis of variance. Later, when the residuals were examined, it was found that there was at least one outlier generated by each analysis of variance. On examination of these outliers, it was found that the birds concerned all came from the same farmer. He was raising them so they had consistently lower fat in their meat than any of the other ostriches. Further investigation of this farmer could reveal he used a diet that will answer the aim of the research – which was to develop a diet for ostriches which results in low body fat. The farmer was doing exactly what was required as an outcome, he was already producing birds with low body fat and high muscle yield – but the student had missed the point entirely. The lesson from this is that **data analysis is not complete without a proper investigation of the pattern before carrying out a confirmatory analysis.**

The preliminary analysis of the data (exploration and description) should reveal the following:

- Structure/shape of the data and pattern as related to the objectives
- Outliers or unusual observations
- The need to modify the data
- Patterns suggesting new questions and the data analyses.

Methods used to explore and describe data

- Descriptive statistics
- Tables
- Graphs

All of these must correspond to the objectives of your research. These methods will carry you a long way through the analysis when added to the formal statistical techniques that you will use.

Table 1. Numbers of people interviewed by village

Count of village	
Village	Total
Eb	4
Ed	7
Ei	18
Ek	1
El	12
Em	13
Es	4
Et	1
Eu	7
Ey	4
Lu	7
Mu	4
Ny	17
Sa	18
So	1
Sr	3
Grand total	121

The process of describing data sets is probably best illustrated with examples. The data for the first example, the survey of farmer's perceptions, were first entered into an Excel spreadsheet. When starting the data analysis of surveys it is common practice to start with a series of tables summarising the data. The **Pivot Table** and **Pivot Chart** facility in Excel is possibly one of the most powerful and useful tools that Excel provides (Table 1). Table 1 is a one-way table that summarises the numbers of people interviewed by village. It is usual to start the analysis of survey data by describing the demographics of the population interviewed.

The summary would be more informative if more information was included. For example, it could include gender (Table 2). This table is now a two-way table. However, examination of Table 1 reveals that there are a number of villages with few respondents. At some stage in the analysis it may be worth combining the smaller villages into like groups. Similar groups can be established using common

Table 2. Numbers of people interviewed by village summarised by gender

Count of village	Gender		
Village	F	M	Grand total
Eb	2	2	4
Ed	3	4	7
Ei	11	7	18
Ek	1		1
El	8	4	12
Em	1	12	13
Es	2	2	4
Et		1	1
Eu	4	3	7
Ey	1	3	4
Lu	3	4	7
Mu	4		4
Ny	12	5	17
Sa	12	6	18
So	1		1
Sr	1	2	3
Grand total	66	55	121

Table 3. Summary of the use of natural fallows by gender

Count of natural fallow	Gender		
	F	M	Grand total
No	31	22	53
Past	5	6	11
Still	29	26	55
Unknown	1	1	2
Grand total	66	55	121

sense, for example, villages close together, or by using a data-driven method such as **cluster analysis**. The clustering could be based on variables decided upon after examining the objectives of the research.

But also think about the analysis objectives: Do you actually need to know about differences between different villages? Maybe 'village' is only recorded as part of the logistics of data collection, and need not appear in your summary tables. **The point is: the tables should relate to what you need to know.**

One objective in this survey is to assess the use of fallows in the past. This is given in Table 3.

For the intercropping experiment, the data were first entered into an Excel spreadsheet. It was noted, while entering the data, that a mistake was made in carrying out the experiment. One of the plots that should have been Treatment 3 was accidentally assigned Treatment 6. This is not the end of the world, and the data can still be analysed with slight adaptations to some of the methods. As these data are in Excel it is possible to carry out some analyses using Excel. The output shown here is not the default from Excel. The Excel output showed many decimal places. This has been modified to give one decimal place for each average. Displaying too many decimal places is a common mistake made by students – often because they just copy and paste the output from one package into their document.

A quick examination of Figure 1 reveals that the treatment with the highest weed biomass was the pumpkin-only crop with no weeding. The lowest weed biomass in the sole maize and sole pumpkin crops was for those weeded at 3+5+8 weeks. All the pumpkin–maize intercrops that were weeded showed similar weed biomass levels. It is not clear if these are different, however it appears that the best return for effort is to weed the intercrop at 3 weeks. Bar

Table 4. Two-way pivot table of average weed biomass for the three crops and four weeding treatments

	Weeding				
Crop	None	3 weeks	3+5 weeks	3+5+8 weeks	Average
Sole maize	83.5	79.3	28.3	6.8	51.4
Pumpkin–maize intercrop	87.6	6.2	3.5	5.5	24.2
Sole pumpkin	162.2	30.9	40.8	3.1	59.3
Average	111.1	35.6	23.7	5.1	44.2

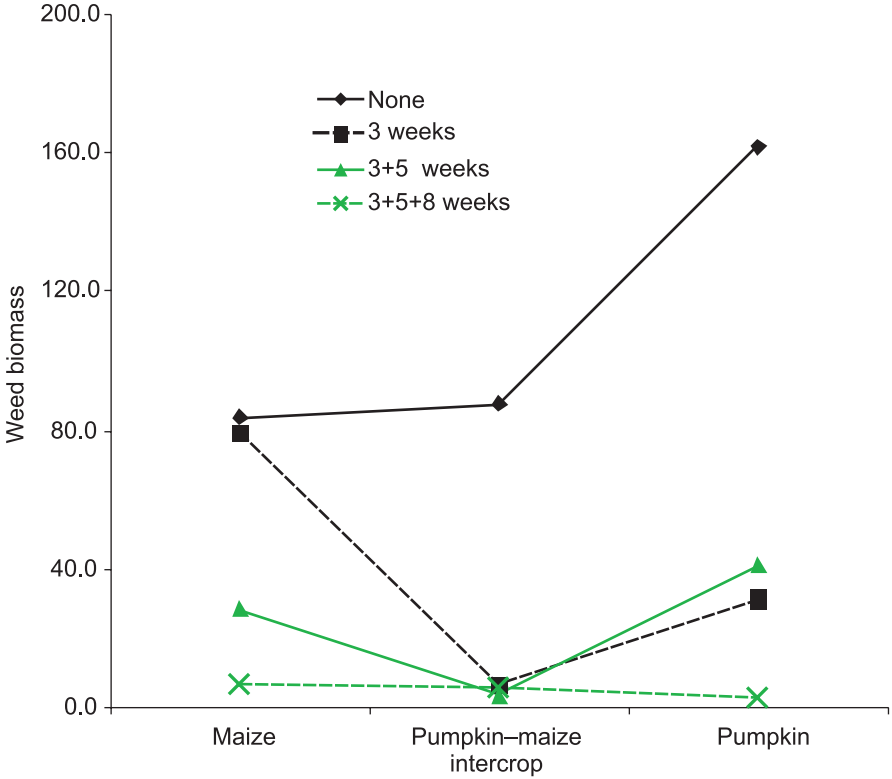


Figure 1. Graph of data in Table 4

charts can be useful displays when presenting results because they give a quick visual display of what is going on that most people intuitively understand. However, the x-axis would be the crop treatments, the plots would be each weed biomass and separate lines could be used to join each weed treatment for the three crop treatments. (Figure 1). Note that the lines connecting the points highlight which ones are from the same weeding treatment. They do not suggest that there is a weed biomass for some intermediate treatments.

So far, the analysis has summarised the data using **means**. Other summaries such as the **minimum and maximum values, trimmed means, standard deviations** and **standard errors** can be calculated. These summaries are useful when dealing with larger data sets, as they may give indications of outliers, but they are not useful when dealing with small data sets like the pumpkin–maize intercropping data. Take care in your use of a spreadsheet, it may contain statistics calculated in a way that you are not sure about. For example, the way the spreadsheet deals with missing values, or whether the spreadsheet is calculating a population or a sample standard deviation. If in doubt, take the data into a statistics package.

Table 5 shows a printout for the summary statistics calculated on the pumpkin–maize intercrop. This is not suitable to be presented in your report as an unmodified printout for a number of reasons. First, notice that the variable name 'Treatment' has been reduced to eight characters by the statistics package. Next, unhelpful treatment numbers rather than informative names are given. Most importantly there are statistics given by this printout that may not be appropriate, or that you may not even understand. The statistics may not be

Table 5. Descriptive statistics of the weed biomass of the pumpkin–maize intercrop data (see text for explanation as to why this table should not be presented in this form)

Descriptive Statistics

Variable	Treatmen	N	Mean	Median	TrMean	StDev
Biomass	1	3	83.5	84.5	83.5	27.8
	2	3	79.3	98.4	79.3	58.5
	3	2	28.30	28.30	28.30	4.92
	4	3	6.81	5.44	6.81	5.84
	5	3	87.6	73.3	87.6	36.3
	6	4	6.24	7.03	6.24	2.27
	7	3	3.527	2.960	3.527	1.136
	8	3	5.49	7.32	5.49	3.24
	9	3	162.2	160.0	162.2	18.9
	10	3	30.92	34.89	30.92	9.14
	11	3	40.83	38.03	40.83	6.01
	12	3	3.14	1.73	3.14	3.10

Variable	Treatmen	SE Mean	Minimum	Maximum	Q1	Q3
Biomass	1	16.1	55.2	110.8	55.2	110.8
	2	33.8	13.7	125.9	13.7	125.9
	3	3.48	24.82	31.78	*	*
	4	3.37	1.79	13.21	1.79	13.21
	5	20.9	60.7	128.8	60.7	128.8
	6	1.14	2.91	7.98	3.87	7.81
	7	0.656	2.785	4.835	2.785	4.835
	8	1.87	1.75	7.39	1.75	7.39
	9	10.9	144.4	182.1	144.4	182.1
	10	5.27	20.48	37.41	20.48	37.41
	11	3.47	36.73	47.72	36.73	47.72
	12	1.79	1.00	6.70	1.00	6.70

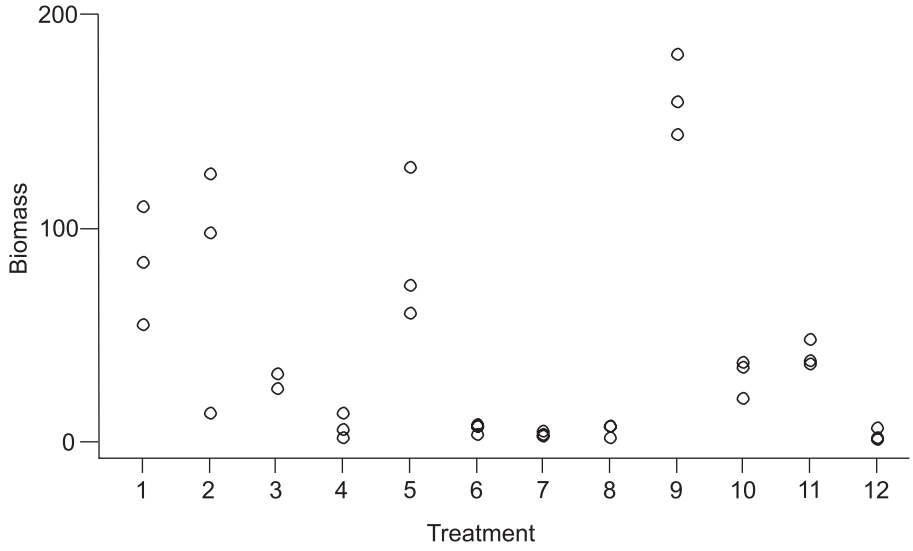


Figure 2. Dotplot of the weed biomass for the pumpkin–maize intercrop experiment showing each of the treatments

appropriate for the data structure and the statistics may not answer the research objectives. The student is also allowing the statistics package being used to have undue influence on the analysis and the presentation of the results and to distract him/her from the objectives of the research. This is also an extremely untidy table which is difficult to read! Tables are better if they don't go over a row for each treatment, and 'treatment number' would make more sense if it was replaced by a text label.

Figure 2 shows a type of exploratory graph (dotplot). **Dotplots** show the spread of the data. Notice how the points for Treatments 1, 2, 5 and 9 are more spread than those of the other treatments. The graph is still labelled with unhelpful treatment numbers rather than names but maybe that does not matter. This is an example of a graph which is important to you in analysing the data – it shows that some treatments are much more variable in weed biomass than others. That is something you may need to take into account in your analysis. But, unless it relates to a key analysis objectives, you will not need to include this graph in your report. Hence, its inelegant layout is not a problem.

Other plots like **boxplots** and **stem-and-leaf plots** are available and should be tried. If the data have two related variables, for example, yield and amount of fertilizer applied, **scatterplots** should be plotted to check whether a **regression line** can be fitted.

The analysis shown so far should be repeated for each response variable in a data set. You can see that by this stage you could already write a fair amount on the data patterns and subjectively suggest results. For example, in the intercropping example you could suggest which is the best crop and weeding regime to use. But there are some severe limitations to this analysis. Two important ones are:

1. Only simple patterns can be investigated. You can look at how y varies as x varies by plotting y against x. But what if there are several x's, all to be considered simultaneously?
2. There has been no consideration of the uncertainty in any of the summaries that are used to interpret the data. Yet we know there is variation in the observations, so there is uncertainty in the results.

The formal analysis addresses these problems. But, you should note that although Excel is useful for the descriptive analyses described so far, Excel is not good for more formal analyses and modelling. Use a reputable statistics package such as Minitab, Genstat or SAS.

Formal analysis and statistical modelling

Key points for this section include:

- Complex patterns involving several variables at the same time can be investigated by fitting statistical models to the data
- Much of the formal statistics taught in introductory courses aims at providing information about the precision of estimates used to interpret the data
- Formal statistical analysis should never be an end in itself, but part of data interpretation.

The next stage in the analysis after describing the patterns and identifying any outliers is to confirm them. This is done by fitting **models** and carrying out a **confirmatory analysis** of the subjective results. Some of the confirmatory analysis may also be used to look at pattern and residuals, which means the exploratory data analysis stage (and the data checking) is not yet completed.

When doing 'research' you will usually wish to generalise from your data to some wider population. This is what statistical inference is all about. So you are likely to find that

descriptive stuff is insufficient. You are after all doing a research degree. If there is no generalisation, then there may be no research – and you might not get your degree!

It is at this stage that you will need to get some idea of the precision and accuracy of your results. **Precision** is the closeness of the data points to each other. It is often measured using **variance**, whereas **accuracy** is the closeness of the data points to the true population value.

Regression

Statistical models are mathematical representations of the pattern in data. The simple regression model is often taught in basic statistics courses as it illustrates many important concepts. A regression is the fitting of a straight line to data to describe and predict the relationship between two variables. These variables consist of a response variable or **dependent variable** and the variable to which it responds, the **independent variable.** In the following regression example (Figure 3) the relationship between inorganic soil nitrogen and crop yield was investigated.

Checking the nature of the relationship

Before any model is fitted you must investigate the **nature of the relationships** between the variables. With a regression model an attempt is being made to fit a straight line to the relationship, consequently you must check first if there is a straight line relationship. Any model can be fitted to any data, but this doesn't mean a relationship actually exists. One of the biggest errors you can make is to fit a model without checking that the model is sensible for the data.

Fitting the line

In Figure 3 there appears to be a straight-line relationship between the two variables, although there is some variation about this line. After carrying out the regression analysis it was found that in the model, the regression line is:

Yield = 1.23 + 0.142 inorganic soil nitrogen

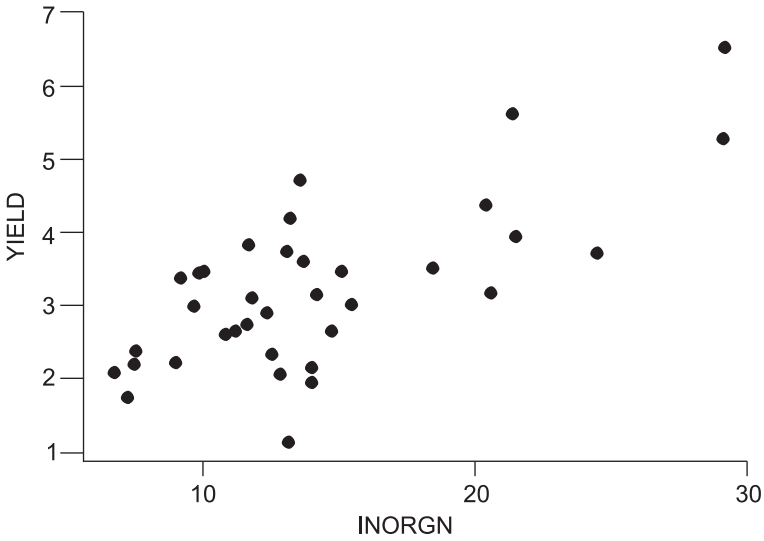


Figure 3. Relationship between soil inorganic nitrogen and crop yield

Another common error is to include all the regression output generated by computer software in the text of the report. If you really need to include the output, it is best put into an appendix with the key points summarised in the text. This is also a common mistake when carrying out the ANOVA and can be dealt with in the same way.

Your work is not complete when you find the model or regression line. You must still check to see if you have fitted an appropriate model and if each of the parameters should be in the model. This is done using the residuals and carrying out a number of significance tests. You must indicate in the methods section of your report that such checks have been carried out, or the reader will assume they have not been done and question the validity of the models that you have produced (see any good statistical text or ask a statistician for details on how to check residuals).

Part of exploratory data analysis is checking patterns of the residuals – there should be no patterns if you have picked up the pattern and structure of the data correctly. However, you often cannot check the residuals until you have actually fitted a model because the residuals are the result of fitting the model.

Confirmatory analysis

Having checked that a model is now possible you need to look at the confirmatory statistics supplied in the output. This is the **statistical inference** part of the data analysis. There are several concepts you need to understand before you can do the next part of the analysis. These are **estimates (point and interval)** and **tests of significance.**

It is the ideas of **inference** and **modelling** that students are usually taught in university statistics courses, but find difficult. So you may have to review the key ideas of estimation, confidence intervals and significance testing. Often this has been covered in your statistics course, but in ways that are difficult to relate to your needs in analysing your research data. This is perhaps because your statistics course was too theoretical. It may not have included realistic examples. Some courses still do not integrate the use of the computer with the discussions of the concepts of statistical modelling. Also, you may not have been very interested at the time, perhaps because you had convinced yourself that the ideas were difficult.

There are now many resources that you can use to review the ideas you need without using too much mathematics. The references at the end of this chapter are for students who need to review such ideas. But don't leave this too late in your thesis writing. The later you leave it, the more pressure you will be under to finish the thesis, so you will not be able to concentrate on reviewing ideas that are not central to the needs of a current chapter.

If learning statistics was a problem for you, then remember it might also have been a problem for your supervisors. They may be hoping that you will be more comfortable with statistical concepts than they were! Even if they now like statistics, be aware that some supervisors may cling to one or two favourite methods of analysis. These may not be the only methods that can now be used to process your data.

You should understand these ideas, because **you should not do analyses that you do not understand. The rule remains to analyse the data in ways that are dictated by the objectives of your study.** For example, suppose you use a method called principal components in your analysis. You do not necessarily need to understand all the formulae that underlie this method. But you must be able to explain (perhaps in an oral examination) why you have used this method and how the results have contributed to your understanding of the data in relation to the objectives of the analysis. It is not sufficient to say:

- 'It is the common method that everyone seems to use
- My supervisor said I should use this method
- An article I found in an international journal used this method.'

These are all sensible reasons, but it is your research and your data. You must be able to explain why each method is appropriate for your own work. This is rarely a big problem, but it can loom large, because you may not feel confident about the topics, and hence feel that you are unable to decide how to proceed. In such cases it is good if you encourage communications, perhaps the statistician and your main supervisor could meet to discuss their differences. If they meet, you may find they are discussing general issues of principle, and straying from your well defined problems of how to analyse your data. If so, then the key is (as always) that the analysis must help in the objectives of your study. In the end you may have to make some compromises (see Table 6, and the section on ANOVA).

Analysis of variance (ANOVA)

Another type of modelling that you will commonly come across in agricultural research is the **analysis of variance**. As the name suggests, it allows you to determine how much of the variation in the response can be attributed to different treatment factors or other effects. For further details on ANOVA, you can refer to any good experimental design text book (Mead *et al.*, 2003, or the Statistical Services Centre, University of Reading website).

Investigating pattern in ANOVA

The ANOVA can also be used for further investigation of pattern, by using it to generate plots of the two factors and the responses and tables of means and examination of the residuals resulting from the fitted model. This means the data are examined using more than one source of variation at a time (combining the two factors instead of examining just one). In Figure 1, for example, you can see the effect of weeding, of the crop and of any interaction (non-parallel lines).

Confirmatory analysis in ANOVA

Figure 1 shows us that weeding treatments 2, 3, and 4 generally give a much a lower weed biomass than weeding treatment 1, where no weeding was done. The plot also shows an interaction – the lines are not parallel. Such results are often presented with **significance levels** (**P-values**). Make sure you know what these mean and how to interpret them. You should not be using such statistics if the implications and assumptions on which they are based are not clear.

The same is true when it comes to presentation of results. For example, many supervisors (and journal editors) insist on placing 'letter values' adjacent to the numbers in tables of means, as in Table 6a. Here a supervisor insisted that the results of a multiple comparison test are included. He always does this, and it was the key in his thesis, which was in the same area as this work. A statistician may well have other ideas on how the results are best presented, perhaps as in Table 6b.The statistician does not see how these tests help in the analysis, and proposes that just the standard error of the differences between the means is presented instead, while also matching the layout of the table to the structure of the treatments. You need to understand enough of the statistical ideas to choose between these (and other) presentations, and to defend your choice in front of supervisors, examiners and editors.

Table 6. Tables of mean weed biomass for the pumpkin–maize intercrop experiment

a. Treatment	Mean weed biomass (g/m²)[1]
1. Sole maize with no weeding	83.50b
2. Sole maize with weeding at 3 weeks	79.30b
3. Sole maize with weeding at 3+5 weeks	28.30a
4. Sole maize with weeding at 3+5+8 weeks	6.81a
5. Pumpkin–maize intercrop with no weeding	87.60b
6. Pumpkin–maize intercrop with weeding at three weeks	6.24a
7. Pumpkin–maize intercrop with weeding at 3+5 weeks	3.53a
8. Pumpkin–maize intercrop with weeding at 3+5+8 weeks	5.49a
9. Sole pumpkin with no weeding	162.20c
10. Sole pumpkin with weeding at 3 weeks	30.92a
11. Sole pumpkin with weeding at 3+5 weeks	40.83a
12. Sole pumpkin with weeding at 3+5+8 weeks	3.14a

1. Means with the same letter are not significantly different (5% LSD)

b.	Weed biomass (g/m²)		
Weeding	Sole maize	Intercrop	Sole pumpkin
None	83.50	87.60	162.20
3 weeks	79.30	6.24	30.92
3+5 weeks	28.30	3.53	40.83
3+5+8 weeks	6.81	5.49	3.14

Average SED = 18.5

There is a need to examine the appropriateness of the model, and the model fitting itself can lead to further data exploration and understanding of the results. In the pumpkin–maize intercrop, the ANOVA table showed some significant effects, but the analysis of the residuals showed that the model was not appropriate (the residuals showed inconsistent variances across the treatments and the normal probability plot of the residuals was not a straight line). If you only fit the model for the significance levels to test your null hypotheses you will miss the real information in your data. You might produce significance levels with no understanding of what really happened. The question is, you found a significiant result, but so what? You have missed the patterns and information in the data and you have yet to prove that the 'significant' model is actually appropriate.

Mixed modelling

ANOVA can be difficult to apply to situations where there are multiple sources of random variation – such as those between villages, between farms within villages and between plots within farms. An approach to modelling, called **mixed modelling** is now available to deal with these situations. This is an important statistical development for analysis of many field studies, both survey and experiment. See Allan and Rowlands (2001) for further information.

The methods touched on briefly in this chapter are the main methods that have been used in agricultural research in Africa.

Note that the value of any statistical method depends on what you want to find out, and the nature of the data and the research design that generated it.

It does not depend in any way on whether your data came from a research station or from farms, whether the study was participatory, or whether it relates to the biophysical, social or economic aspects of a problem.

Making sure you satisfy the research objectives

Important concluding points to remember:

- Revisit objectives and make sure they have been met
- Check that all the analyses you include in your thesis really contribute to those objectives
- Don't work in isolation.

Once the analysis appears to be complete, you need to revisit your objectives and make sure that they have been fulfilled. Remember, the whole process of data entry and analysis should have been iterative and should aim to meet the objectives of the research. The analysis that you ended up doing may not be the same as that planned in the original project proposal. At this stage you need to revisit your objectives and relate them to the results you have obtained.

A way to start this process is by laying out the original proposal objectives, tables, graphs, outputs, descriptions, conclusions and interpretations in front of you. Lay out the results in the same order that you lay out the methods. The sequence should be logical and should not jump around from topic to topic. Now try the 'so-what' test. Check that every item of statistical analysis you are going to report actually contributes to the conclusions that you have reached. Check that the conclusions match the original objectives. Are your conclusions really conclusions? Make sure you haven't read into the data something that you would like to see there. This is really easy because you have been so close to the whole project it is difficult to divorce yourself from it and see it objectively. Now relate and interpret your results to the literature. At this, as in all stages of the research process, it is important that you don't become tempted to work in isolation. Talk to your colleagues, get help and give seminars periodically so that you can get feedback from those around you.

Remember, don't waste your data by only carrying out significance tests and that you don't have to be a hot shot statistician to get really good information out of your data.

Resource material and references

Appendix 10. Coe, R., Stern, R., Allan, E., Beniest, J. and Awimbo, J. 2002. *Data Analysis of Agroforestry Experiments*. World Agroforestry Centre (ICRAF), Nairobi, Kenya.

Appendix 11. ICRAF. 2003. *Genstat Discovery Edition and Other Resources*. World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Allan, E. and Rowlands, J. 2001 *Mixed Models and Multilevel Data Structures in Agriculture*. Statistical Services Centre, The University of Reading, UK. http://www.reading.ac.uk/ssc/

Jones, A., Reed, R. and Weyers, J. 1998. *Practical Skills in Biology*. Second edition. Longman, UK. 292 pp.

Mead, R., Curnow, R.N. and Hasted, A.M. 2003. *Statistical Methods in Agriculture and Experimental Biology*. Third edition. Chapman and Hall, London, UK. 472 pp.

Muzamhindo, N. 1999. *Analysis of Experimental Data for Maize Crop at University of Zimbabwe Farm*. BSc. Honours Project, Department of Statistics, University of Zimbabwe, Harare, Zimbabwe.

Stern, R.D., Coe, R., Allan, E.F. and Dale, I.C. (Eds.). 2004. *Statistical Good Practice for Natural Resources Research*. CABI Publishing, Wallingford, UK. 387 pp.

The Research Support Unit of the World Agroforestry Centre (ICRAF) have some training materials and other guides on analyses. www.worldagroforestrycentre.org/research support.

Velleman, P.F. and Hoaglin, D.C. 1981. *Applications, Basic and Concepts of Exploratory Data Analysis*. Duxbury Press, Boston, Massachusetts, USA. 354 pp.

4.8 Models

Catherine Wangari Muthuri

- **A model is a simplified representation of part of the real world. In this chapter we discuss models that can be described mathematically**
- **Models are based on theory. In research models help to test theory by making predictions that can be compared with observations**
- **Models also allow the implications of research results to be explored by making predictions for new situations**
- **Each model is built for a specific purpose. A model that is useful for one job may be inappropriate for another task on a similar topic**
- **Models vary in scope from the simple, which you can put together and use very quickly, to the complex that may take much of your project time to develop and use**
- **Computing tools designed for the job can make modelling feasible for students who are not specialists**

Introduction

Modelling can mean many things in research, and models of one sort or another play a crucial role in much research. Experience shows that the role and use of models is rarely explained in research methods courses. The result is that many students have only a vague idea about what models can and should be doing for them. Modelling is often regarded as the domain of specialists who sit hunched over computers, not of agricultural researchers who want to solve real problems in the field. The result is that much research is less effective than it might be. The aim of this chapter is to start to fill that gap.

The chapter is divided into three major parts. The first shows you how models are a natural part of the research process. This is to help you develop your ideas from the general 'models are everywhere' to the main focus of the chapter, which is concerned with mathematical or simulation models. The second part discusses your options if you plan to do some mathematical modelling. Finally, details of the steps you need to follow to construct, use and test simple models are described, using examples where modelling tools have been applied in research studies in Kenya. Research findings can be enriched by the use of simulation models and this is an attempt to encourage you not to shy away from using modelling tools just because you don't like maths!

Model types

Models are everywhere

You may not be aware of them, but you are using models all the time. They come as physical models in all shapes and sizes from dolls, miniaturised cars and aeroplanes and globes, or as visual representations in maps or pictures. They may be presented as verbal or mental models, or in more abstract arithmetic or algebraic form, in nearly all we learn. A model is just a simplified representation of part of the real world.

Physical models have been used for centuries in research. Engineers use models of boats to study their stability and resistance to movement through the water. In biological research one species is often said to 'model' another; in the early stages of medical research monkeys and mice are used to model man, because they represent *some* aspects of human physiology well. The images we carry in our minds, i.e., mental models, are simplified representa-

tions of complex systems. We use them constantly to interpret the world around us and we usually do not realise that we are doing so.

None of these models involve the complete similarity of real world and model, but similarity in key features. A model is useful if it behaves in a realistic way for your problem. The scale model of a ship may be useful for investigating its stability in the water, but it will be useless for determining the profitability of operating the ship. Different models of the same phenomenon are useful for different things. Take a 1-ha farm as an example. A map of the farm (a visual scale model) might be useful when the farmer is planning the location of different crops. Physical models of the landscape, built up from clay and painted, can be used to examine the interaction of the farm with neighbouring farms and other land areas. Numerical input–output models help in making investment decisions. Detailed numerical topological models can be used to understand water flow and erosion on the farm. Each of these is a 'model of the farm' and each is useful for its own purpose, but inadequate for other purposes.

Models in the research process

Research involves developing a theory of the real world and testing it with observation, then perhaps using it to explain and predict further phenomena. Models are representations of the theory and hence a fundamental part of the research process. Whether the model needs to be formalised and described mathematically depends on whether the predictions of the theory can be worked out without formalisation.

Models can be used in two steps of research:

1. In generating hypotheses or predictions, that will suggest the observations of the real world that need to be made.
2. In assessing the extent to which our theory (as captured in the model) explains the real-world observations.

If the model and observations agree then there is nothing in the data to suggest the theory is not a good description of the real world. But of course we might have collected data that does not test the theory in ways that are interesting! An important part of research design is planning observations that do discriminate between models which are fit and unfit for their intended purpose.

If the model and observations do not agree then you can:

- Question the model structure and assumptions, and revise it
- Question the data: perhaps it is not really relevant to the model you have chosen
- Abandon the line of research.

Mathematical models

This chapter is about the mathematical models that are used in agricultural research. If the relationships and rules that make up the model are sufficiently well specified, then they can be written down mathematically and produce numerical results. In very many models the basic mathematical relationships and rules are simple (such statements as 'volume = mass/ density 'or 'yield is zero until after flowering'). Complex patterns of results often emerge because of the many interacting components, rather than because there are some complex mathematical ideas embedded in the model. This is important. **It means you do not have to be a mathematician, or even very good at using mathematics, to make effective use of models in your research.**

A **mathematical model** is a set of equations that represent interconnections in a system, and can be worked out either by hand or by using a computer. The equations are written in terms of mathematical objects that correspond directly to physical quantities. If these objects change as part of the phenomenon they are generally called **variables** while if they are fixed they are generally called **parameters.**

Typically a model will consist of formulae that link some responses or quantities of interest with inputs, or the things that affect them. For example, a simple model of soil moisture changes is illustrated in Figure 1.

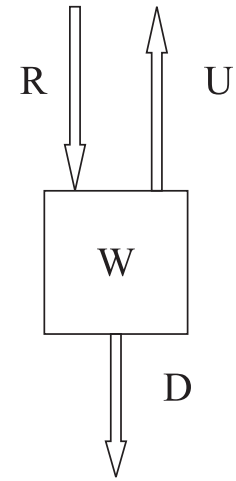


Figure 1. Simple model of soil moisture

The soil moisture (W) at time t is W_t. Rainfall is R_t, uptake by plants is U_t and drainage is D_t. The model can be written mathematically as:

$W_{t+1} = W_t + R_t - U_t - D_t$

If we know the initial conditions (W_0) and the values of R, U and D then we can calculate W at any time. The model is simplistic. It ignores soil evaporation. That will not be a problem if the model is being built for applications in which soil evaporation can be ignored, but would be a major deficiency in other cases. The model also requires inputs that might be hard to measure (U and D). For some purposes you might be able to predict U, by adding another part (some more components) to the model. For some purposes you could take:

$U_t = c.P_t$

where P_t is the potential evapotranspiration and c is a 'constant'. This model might well be useful for studying the effect of day-to-day changes in P on W. However it is still too simplistic for longer-term studies, as c will probably not be constant, but will change as the crop grows and matures. The value of c may also depend on W, with the plants able to take up less water when the soil is drier. It is easy to see how this process can quickly lead to models of ever-increasing complexity, even though each step involves simple and realistic relationships.

Part of the skill in modelling is in choosing the components to model, including the things which will be necessary but not putting in everything you can think of.

Conceptual and empirical models

Models can either be **empirical** (data-driven) or **theoretical** (theory-driven or conceptual). An **empirical** model is based mainly on data. It may be used in statistical analysis of study results and to predict within domains of 'similar' conditions to the empirical base. It does not explain a system. For example, a fertilizer response curve is an empirical model. It can be developed from observations on the yield of crops with different amounts of fertilizer, and used to predict the yield at any fertilizer level. However, it does not explain why the yield response is the way it is. An empirical model consists of one or more functions that capture the trend of the data. Although you cannot use an empirical model to explain a system, you can use such a model to predict behaviour. We use data to suggest the model, to estimate its parameters, and to test the model. An empirical model is not built on general laws and is a condensed representation of data. However many statistical or empirical models are built

on elements of an underlying theory, for example, we construct the input variables in a regression model based on a theoretical understanding of factors that should determine the response.

A conceptual, theoretical or 'process based' model includes a set of general laws or theoretical principles. If all the governing physical laws were well known and could be described by equations of mathematical physics, the model would be physically based. However, all existing theoretical models simplify the physical system and often include obviously empirical components. Thus the distinction between conceptual and empirical models is not clear-cut. And again, it is the modellers job to use something appropriate for the task, rather than to assume that one approach has more intrinsic value than another.

Roles of models

Models play several roles including:

- **Exploring** the implications of theory. It may not be possible to see the implications of theories that involve several interacting components without calculating what happens in different conditions. Used in this way, models provide insights and add creativity
- **Prediction** or forecasting tools help users make sensible educated guesses about future behaviour. These can be used in planning, scenario analysis and impact analysis
- **Explaining** observations and generating hypotheses
- **Training** so that learners can carry out 'virtual experiments', exploring the result of making changes.

In research models can help answer such questions as:

'Can I construct a theory that explains my observations?'

'Is my hypothesis credible?'

'What new phenomenon does my theory help to explain?

Used for prediction, models can answer such questions as:

'Given the model, what will happen in the future?'

'Given the model, what's going on between places where I have data?'

'What is the likelihood of a given event?'

How to model

You have three options if you decide to use simulation models in your work. You can use an already existing developed model, modify an existing model or develop a new model altogether.

Using an already developed model

Hundreds of models relevant to agricultural research have been developed and described and are available to you. A few have to be purchased. Many are available free to researchers and can be down loaded from web sites or obtained from the authors.

The advantages of using a model that someone else has developed include:

- **Time saving.** Some of the hard work has already been done
- **Recognition.** Some models have been widely used and described. Their value is already recognised so you will find it easy to justify their use
- **Support.** You will find documentation, examples and maybe technical assistance in using the models.

However there are also disadvantages, compared to the alternatives of developing your own models. These include:

- You may not find a model that actually describes the phenomena in which you are interested at the right level of simplification
- The available models may require inputs that are not available to you
- You may not fully understand how the model is constructed (the theory on which it is based)
- The model may not run on any computer available to you, or in the way you need for your research.

If you are considering using a model, then select it by:

1. Determining exactly what you want to do with it. You will only be able to decide if candidate models are suitable when your task is clear.
2. Searching literature and the Internet for references to models that tackle your problems, and asking experts in the field.
3. Evaluating each possible model against your requirements. If you end up with more than one candidate then choose the simplest.

Modifying an already existing model

You may well find that no available model meets your requirements but that some come close. Therefore it may be desirable to modify a model. Modifying it may mean anything from changing the way input files are handled to adding to or changing some of the underlying theory. Often modification will mean adding a description of further components and processes to address a specific situation.

If you plan to adapt or modify an existing model, all the points above about selecting it apply. In addition you will have to be able to:

- Get access to the original computer code and description of the theory behind it
- Understand them fully
- Know how to modify it for your needs.

The computing skills you need will probably be more than those you need to just run an existing model.

Some models are much easier to adapt than others. If they were originally designed and produced with adaptation in mind then the task may be straightforward. If they were not built to be adapted the task of modification may be all but impossible.

Adapting a model takes longer than using an already existing model. You need to go though all the steps in the modelling process that are discussed later in this chapter. This implies that the exercise becomes a major component of your research. It therefore demands that you have sufficient skills and are familiar with the language of the packages and software used.

Developing a new model

The third option is to develop your own model. Situations that necessitate developing models include those when:

- The outputs generated and inputs required are not catered for in the existing models
- Exiting models are too clumsy or complicated, or have a poor track record
- You are working in an area where no existing models can be found.

Given the novelty of most research, the last is likely to be the case.

Building and using your own models could be:

- Something that takes a few hours, if you are simply looking at a few interacting components and are familiar with a suitable computing environment

- Something that takes most of your 3 years as a PhD student!

More likely it will be somewhere between the two. The steps in developing a model are outlined below. The most critical are the first ones: **defining useful** and **realistic objectives.** You will probably be most successful if you start with simple objectives. Reduce the problem to its simplest objectives, and work on the simplest model that will meet those. This might be a model with no more than two interacting components and simple rules describing them. Yet even these models can give insights into your theory and observations that are not apparent until the model is formalised.

Steps in modelling

The steps involved in the modelling process are summarised in the flowchart (Figure 2). However, developing any useful model will be an iterative process – you will certainly have to return to early steps, for example, if you are looking again at the interactions in your model when it does not seem to give sensible predictions.

The model-building process can be as enlightening as the model itself, because it reveals what you know and what you don't know about the connections and causalities in the system you are studying. Thus modelling can suggest what might be fruitful paths for you to study and also help you to pursue those paths.

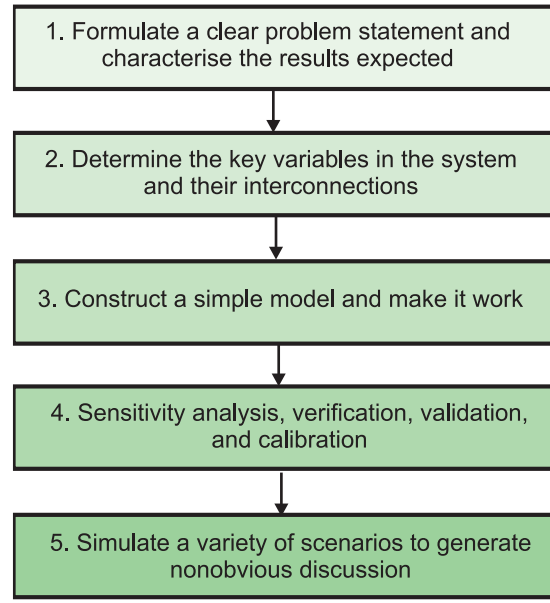


Figure 2. A summary of the steps involved in the modelling process

Formulate a clear problem statement and characterise the results expected

As with all other aspects of research, what you do depends on what you want to find out. Setting realistic and detailed objectives for your modelling will determine the whole nature of the task. It will help you decide on the following important characteristics of models:

1. Will the model need to be deterministic or stochastic?

In **deterministic** models the future state of the system is completely determined (in principle) by previous behaviour. In **stochastic** models the system is subject to unpredictable, random changes. These models involve probability and statistics. If you are interested in risks, your model will have to use stochastic components.

2. What timescale is appropriate?

The **timescale** of the processes in question determines the timescale of the models. Depending on the time taken for the processes under question to reach an equilibrium or to be felt, useful decisions on what to include/exclude in the model can be reached. For example, when looking forward 100 years, you need to ignore daily/monthly or seasonal variations of the

parameters in question. Such variations can be ignored in a long-term model but could be important in a short-time model. Examples of scales and typical times are:

- Metabolic (enzyme-catalyzed reactions; seconds to minutes)
- Epigenetic (short-term regulation of enzyme concentration; minutes to hours)
- Developmental (hours to years)
- Evolutionary (months to years).

3. Does the model need to be spatial?

All agriculture takes place in a spatial context, but only some problems require you to specifically describe spatial interactions. Think of the problem of modelling small farms. If you want to describe economic inputs and outputs of the farm you need to know that there are crops, animals and trees, but it may not matter where on the farm they are. If you want to model nutrient flow between tree and crop plots, then their location matters and the model you use will have to be explicit about that. Many of the management decisions made by small-scale farmers living in heterogeneous environments make use of spatial variability on their farms, such as growing different crops on different patches of land, abandoning part of their land, or focusing their efforts only on those patches with the highest returns to investment of labour or inputs. Most of the current models in agriculture do not handle spatial variability well, if at all. There is a clear need to develop existing models further, or to construct new ones, in order to address this limitation. Unfortunately, the structure of many existing models does not facilitate transformation to spatially explicit versions, as their linear nature restricts them to being run in sequence many times, in order to simulate each patch of land in turn. This makes it difficult to simulate simultaneous interactions between patches of land (e.g., soil, or water flow down a gradient). In circumstances where spatial variability is a key factor affecting the study it is advisable for you to explore using a model that takes this into consideration.

Determine the key variables in the system and their interconnections

In this step you need to determine the key variables in your study that will be represented by variables in the computer model. **Key variables** are the few most important, significant factors that affect the system and their relationships. The cause- and-effect connections in the real system will be represented by interconnections in the computer model. Adding more and more interconnections makes the model complex, though by design, models should be a simplification of the system under study. A determinant of model usefulness is therefore the ability of the modeller to leave out unimportant factors and capture the interactions among the important factors.

Note that a model is:

- Too complex when there are too many assumptions and relations to be understood
- Too simple when it excludes factors known to be important.

Constructing a model

Building a model is an interactive, trial and error process. A model is usually built up in steps of increasing complexity until it is capable of describing the aspects of the system of interest. **Note: It will never 'reproduce reality'.**

The appropriate tools you need to construct a model depend on the complexity of the model. The simplest tools may be **paper and pencil.** Others may use **spreadsheets,** while the

more complex models may require **dedicated modelling software** that uses its own language. The simplest mathematical model takes the form of **equations** show how the magnitude of one variable can be calculated from the others and **spreadsheets** like Excel are adequate for the task.

More complex computer simulations use special software that allows the building and testing of a model. There are software products available that make building and running some types of models very easy even if you know nothing about computer programming. Investigate such software as STELLA and ModelMaker before trying to write your own code in lower-level computing languages. They make the job of developing and running your own models very much simpler!

The development of the simple soil water model outlined in Figure 1 is shown here to give you an idea of what is involved. The model represented in Figure 1 is drawn in STELLA. In Figure 3a. STELLA uses four main types of building blocks:

Stocks. These are stores of 'stuff', represented by rectangles. They may describe water, money, people, biomass,... whatever you are modelling.

Flows. These are the movements of material into and out of stocks, represented by broad arrows. The arrow can be thought of as a pipe, with a tap on it to regulate the flow. Sources and sinks of the material are represented by 'clouds'.

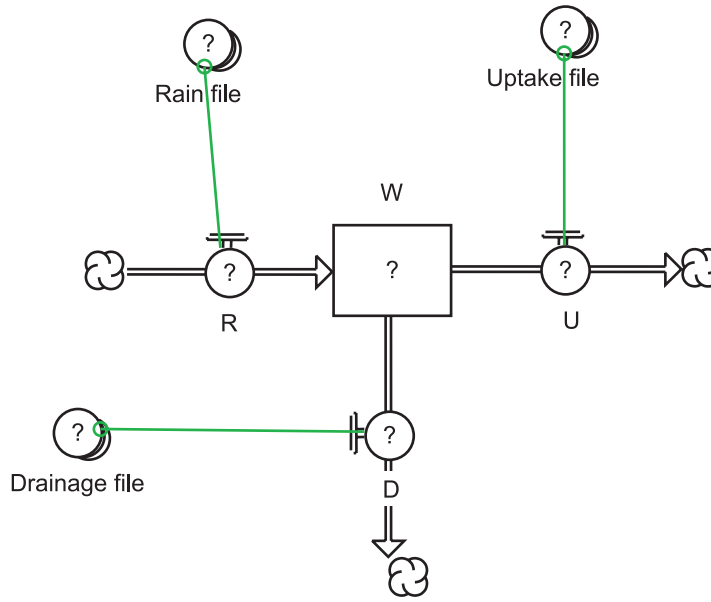


Figure 3a. Simple soil water model in STELLA

Converters. These are represented by circles. They hold values of constants and formulae used to convert one type of material to another.

Connectors. These narrow arrows show the logical connections between components in the model. The equations describing the model must be consistent with these connections.

The stock of soil water (W) has an inflow of rain (R) and outflows of uptake (U) and drainage (D). The actual values of these are read from data files. The model is completed by filling in a formula or other details in each location marked by '?'. The model can then be run.

In Figure 3b the uptake is now calculated as c.P, where P is the potential evapotranspiration (PeT), also read from a file. It should be clear from this that modifying the model requires little more than adding components to the diagram. The real challenge of course is deciding *how* to model uptake, not changing the computer code – this is why software such as STELLA is so important. The final step (Figure 3c) shown here displays two more changes that the modeller thought would help. The drainage is now calculated (because there was no

measured data available) and the uptake now depends on both the crop biomass and the soil water. The latter involves keeping track of the biomass growth, a second stock in the model. Many physiologists would be uncomfortable with a single 'type' of biomass, and start differentiating it into, say, roots, stems, leaf and grain. Then you need to add components that describe what the partitioning depends on. Similarly the soil scientist would like to have several soil layers, each with different hydraulic properties. The model can quickly become complex. The value of software such as STELLA is that it allows you, as researcher, to think about what constitutes a sensible model for you, rather than worrying about computer codes.

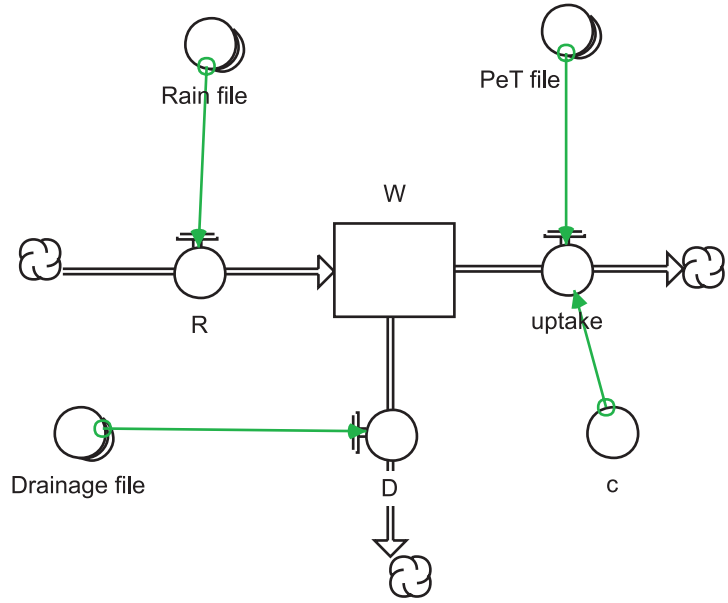


Figure 3b. Simple soil water model with uptake modelled as c.PeT

Sensitivity analysis, validation, verification and calibration

Sensitivity analysis

Through sensitivity analysis, you can gain a good overview of the most sensitive components of the model. Sensitivity analysis attempts to provide a measure of the sensitivity of other parameters or forcing functions, or sub-models to the stated variables of greatest interest in the model. It helps you to

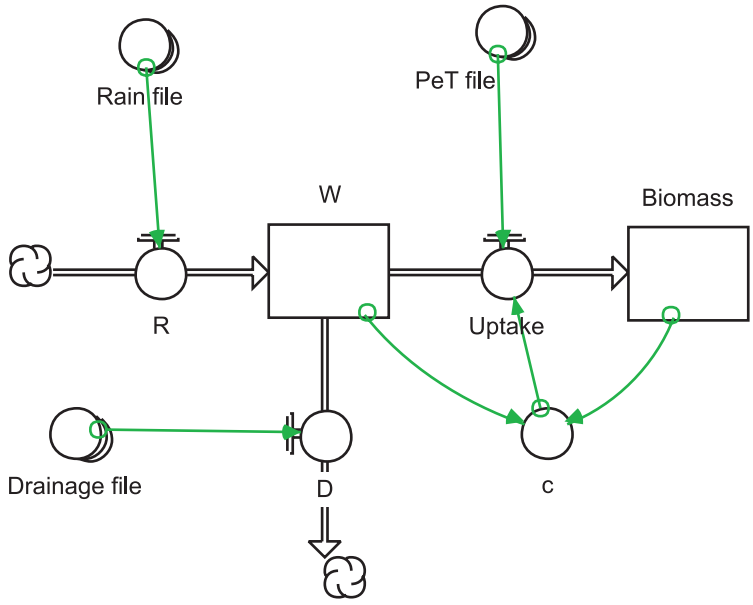


Figure 3c. Simple soil water model with uptake depending on both crop biomass and soil water

systematically explore the response of the model to changes in one or more parameters, to see how sensitive the overall model outcome is to a change in value. This **sensitivity** is always dependent on the **context** of the setting of other parameters, so you should be careful about the conclusions you draw. Some parameters only matter in particular types of circumstance. Others, however, seem to always matter, or to matter hardly at all. This type of model analysis is used to see which parameters should get priority in a measurement programme. You must be provided with affordable techniques for sensitivity analysis if you are to understand which relationships are meaningful in complicated models. This is equally true whether you are using an already developed model, modifying a model or developing one.

Validation, verification and calibration

In general, **verification** focuses on the internal consistency of a model, while **validation** is concerned with the correspondence between the model and the reality. **Calibration** checks that the data generated by the simulation matches real (observed) data, it can also be considered as tuning of existing parameters.

These steps can be among the most conceptually difficult. No model is universally 'valid' in the sense that it will give 'correct' predictions in all circumstances. There will always be discrepancies between observed and predicted values. These discrepancies can be made smaller by calibration and by making adjustments to the model. However this does not necessarily increase the usefulness of the model in either: explaining your observations of the real world, or making predictions about behaviour in the real world.

Simulate a variety of scenarios to generate non-obvious discussion

Simulation models have been used widely in Kenya to address various problems. Three examples are given to help you see how they can be used.

Soil fertility management in western Kenya: Dynamic simulation of productivity, profitability and sustainability at different resource endowment levels

A farm economic–ecological simulation model was designed to assess the long-term impact of existing soil management strategies, on-farm productivity, profitability and sustainability. The authors developed a model that links biophysical and economic processes at the farm scale. The model, which runs in time units of 1 year, describes soil management practices, nutrient availability, plant and livestock productivity, and farm economics. It concluded that low land and capital resources constrain the adoption of sustainable soil management practices on the majority of farms in the study area. Previously it had been assumed that low-input organic methods were suitable for the poorest farmers. For more details, see Shepherd and Soule (1998).

Modelling leaf phenology effects on growth and water use in an agroforestry system containing maize in the semi-arid Central Kenya using WaNuLCAS.

The three tree species under study were *Grevillea robusta* (evergreen), *Alnus acuminata* (semi-decidous) and *Paulownia fortunei* (deciduous). The inputs included climate data, soil data, calendar of events, crop and tree parameters, agroforestry zones and layers, and leafing phenologies. The scenario outputs included soil water balance, tree and crop biomass and stem diameter. WaNuLCAS model simulations demonstrated that altering leaf phenology from evergreen through semi-deciduous to deciduous decreased tree water uptake and interception losses but increased crop water uptake, and drainage rates in all the species. It was

therefore concluded that deciduous tree species would compete less with crops and be more advantageous in increasing stream flow than evergreen trees. Phenology had not previously been a major consideration in determining tree selection For more details, see Muthuri (2003).

Modelling the benefits of soil water conservation using PARCH; A case study from a semi-arid region of Kenya.

The PARCH model was used to simulate maize grain yield under three soil/water conservation scenarios: 1. a typical situation where 30% of rainfall above a 15 mm threshold is lost as runoff, 2. runoff control, where all rainfall infiltrates, and 3. runoff harvesting, which results in 60% extra 'rainfall' for rains above 15 mm. The study showed that runoff control and runoff harvesting produced significant maize yield increases in both the short and the long rains. Previously runoff control was justified more for erosion benefits than increased crop production. For more details, see Stephens and Hess (1999).

Conclusions

The success of models developed by physicists and chemists has led to the rapid development of modern technology, the conquest of many diseases resulting in increased life expectancy, and the improvement of human lives on earth. But, no matter how successful a model has been, scientists realise there may be aspects of the world that the model fails to explain, or worse, predicts incorrectly. Nevertheless, creating and using models is one of the most powerful tools ever developed. But, there is a need to revise and improve models as new information is discovered.

Further resource material and references

There are many books, journals and articles on models. Most tend to be specialised and specific to certain models or application of models in specific areas of specialisations. To understand some basics on what models are, and how you can build a model, three books are listed below particularly useful.

Appendix I. The Craft of Research. Paul L. Woomer. PowerPoint on CD.

Appendix II. ICRAF. 2003. *Genstat Discovery Edition and Other Resources*. World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Anon. 2003. Why the analysis of wide range of physical phenomena leads to consistent and successful results when applying the BSM concept and models? http://www.helical-structures.org/Applications/why_successful.htm

Ford, A. 1999. *Modelling the Environment. An introduction to system dynamics modelling of Environmental Systems*. Island Press, California, USA. 401 pp.

Jorgensen, S.E. 1994. *Fundamentals of ecological modelling*. Elsevier, London, UK. 628 pp.

Matthews, B.R. and Stephens, W. 2002. *Crop–Soil Simulation Models: Applications in Developing Countries*. CAB International, Wallingford, UK.

Muthuri, C.W. 2003. *Impact of Agroforestry on crop performance and water resources in semi-arid central Kenya*. PhD Thesis. Jomo Kenyatta University of Agriculture and Technology (JKUAT). 289 pp.

van Noordwijk, M. and Lusiana, B. 2000. WaNuLCAS version 2.0: Background on a model of water, nutrient and light capture in agroforestry systems. International Centre for Research in Agroforestry (ICRAF), Bogor, Indonesia, 186 pp.

Shepherd, K.D. and Soule, M.J. 1998. Soil fertility management in Western Kenya: dynamic simulation of productivity, profitability and sustainability at different resource endowment levels. *Agriculture, Ecosystem and Environment* 71: 131–145.

Soto, R. 2003. Introducing system thinking in high school. The connector (Connecting system thinkers around the world) 1(5). http://www.hps-inc.com/hps/zine/sepoct03/jake.html

Stephens, W. and Hess, T.M. 1999. Modelling the benefits of soil water conservation using the PARCH model—a case study from a semi-arid region of Kenya. *Journal of Arid Environments* 41: 335–344.

Internet resources

Ecological models http://www.wiz.uni-kassel.de/ecobas.html

CERES crop models http://www-bioclim.grignon.inra.fr/ecobilan/cerca/ceres.html

FALLOW model at http://www.icraf.cgiar.org/sea/AgroModels/FALLOW/

FLORES model at http://www.cifor.cgiar.org/flores/ An example of model building in participatory research

PARCHED-THIRST at http://www.cluwrr.ncl.ac.uk/projects/tanzania/modelling.html

WaNuLCAS model at http://www.icraf.cgiar.org/sea/AgroModels/WaNulCAS/

STELLA software; High performance Systems Inc at http://www.hps-inc.com/

Powersim software; The business simulating company www.powersim.com/

Vensim PLE. Vantana Systems Inc. www.vensim.com/

Management Unit of the North sea Mathematical Models (MUMM) (2003) http://www.mumm.ac.be/EN/Models/Development/Ecosystem/how.php

Model Maker: available from www.modelkenetix.modelmaker/index.htm

4.9 Where is the participation?

Richard Coe

- **Effective projects will involve participation of stakeholders in all stages of planning, implementation and evaluation**
- **Participation in a research study, both who and how, should be determined by the objectives of the study**
- **Participation of farmers or others in a research study is no reason to forget the elements of research design that will allow you to reach valid conclusions**

'One hour of life, crowded to the full with glorious action, and filled with noble risks, is worth whole years of those mean observances of paltry decorum, in which men steal through existence, like sluggish waters through a marsh, without either honour or observation.'

Sir Walter Scott

Part 2 of the book made a strong case for 'participation' – involving those who will use the research in the process. So why has Part 4 of this book, which focuses on research methods, not included a specific section on these methods?

The answer is simple: appropriate method and levels of participation are needed in ***all*** stages of a project. The subject of this book is sound research, and the principles of methods needed to do good research are much the same whoever is participating. 'Participatory methods' for both the social and natural sciences are widely discussed and described in the literature and they have been referenced in the appropriate chapters. Thus, just as we do not describe all the methods available for experimental design, or analysing data or building models, so this book does not provide a comprehensive review of the research methods which are available when collecting and analysing data in a participatory research process. There are specialised texts to help you with this.

This answer will not satisfy everyone, so it is elaborated below. In the discussion I distinguish between participation in a ***project*** and participation in the ***research studies*** that make up the project. The role of participation at the two levels can be different, and the extent to which, as a student, you can have some influence over them is also different.

Projects

Think of the cowpea project described in **Chapter 2.3**. Pests were found to be an important constraint in cowpea production, so the project aimed to find ways of overcoming them. However it is conceived, a project to tackle this problem will have elements of refining understanding of the problem, devising and testing possible solutions, promoting widespread use of the solutions and assessment of their impact, with iteration and cycling around these elements. It therefore seems natural that farmers should be involved in all the stages –

- Who better understands the occurrence and impacts of cowpea pests than the farmers growing the crop?

- Who can test and evaluate solutions, but farmers who will have to use them?
- Who can evaluate their impacts, but the farmers who feel them?

These are the pragmatic reasons for participation in problem-solving projects.

Other reasons are also often given, reasons which might be described as ideological, some discussed in **Part 2**. Development workers all over the world believe that 'development' involves giving people more control over their lives and resources – indeed for some this is the definition of development. It is a right of farmers not to be *told* that their problem is cowpea pests, and told what to do about it, but facilitated in understanding for themselves their problems and solutions which suit them. Projects that take this approach are more likely to lead to sustainable solutions, that continue to have impact past the end of the project and the departure of researchers.

So, if the cowpea project leaders are aware and agree with these arguments they will set up a project which involves the farmers in all stages, with farmers working collegially with researchers, each bringing their own expertise and knowledge to the table. But try to do this and many further questions arise. For example:

1. Just who should participate? Farmers, or maybe others with an interest, such as cowpea consumers and traders. Almost certainly the interests of all parties will not coincide.
2. Who decided that cowpea pests was the problem to work on? If you ask farmers their problems you are more likely to get responses such as 'Paying school fees' or 'Getting a job', rather than 'Pests on the cowpeas'.
3. The participatory approach requires intensive engagement in villages, meaning the project can only involve a few of them. But the problem covers large areas. How can you get the participation of all cowpea farmers?

There are no simple answers to these types of questions. Each project has to find ways that are best suited to the circumstances, but this has to be done knowing and understanding the many options and approaches available.

Point 3 above is one of the main reasons why projects need a sound research component. If your objectives only stretch to helping in those villages or households with which you are immediately engaged, then maybe you do not have to pay much attention to research methods. However there are few instances in which this is the case. Every project wants to generate information that will be useful beyond its own bounds. The only way to do this with known reliability is to use well planned research methods.

As a student joining a project to undertake thesis research you may not have been involved in planning the overall project and the approach to participation adopted. However you need to understand what the approach is and the reasoning behind it. And you must be prepared to challenge it if necessary.

Research studies

Most projects will involve a number of specific research studies which contribute to its overall strategy. How should participation be built into these, and what participatory methods are appropriate? The answer is the same answer that we give to most other research methods questions: it depends on objectives. If the objectives are specified clearly enough then they should guide you. And, referring back to the previous section, there should have been appropriate participation in setting the objectives for each component study.

A single project may well have component studies with different levels of participation, which are mutually agreed by all concerned. An example of a project introducing high-value

trees in a mountainous region identified (among others) three component studies:

1. Farmers wanted to plant a cover crop between newly established trees but did not know whether beans or sweet potato would be most suitable. They agreed they could investigate this themselves without involving a researcher.
2. They had consistent difficulty germinating one species and asked the researcher to investigate. It turned out to be a problem of fungal attack which could be controlled by keeping moisture levels low. Farmers got involved again with testing ways of managing the moisture level in the nursery.
3. They wanted to know how well different species were adapted to different altitude zones. This they agreed had to be a joint effort, with farmers growing the trees but the researcher coordinating across altitudes to make sure comparable methods were used and results compiled.

Don't make the mistake of getting locked into a single mode of participation for all studies! Franzel and Coe (2003), for example, explain how different objectives in testing agricultural technologies can lead to differing balances of farmer and researcher involvement in design, implementation and assessment of the trial.

If your study will require an experiment, there are many ways in which stakeholders may participate. They could have been involved in the whole project, setting objectives, approaches and priorities, including identifying the need for that experiment. They may set objectives for that particular experiment, define details such as treatments and management, choose methods of assessment and evaluation, and plan for taking the work to the next step. Some objectives require little more from the researcher than facilitating farmers evaluating alternatives that they design, using criteria that they choose. In the cowpea project, this might be the case if the objective is simply for farmers in participating villages to determine if the new varieties are of value to them. However if the objectives require you, as a researcher, to come up with generalisable and defendable conclusions, then you will need to pay attention to the experimental design.

Chapter 4.3 on design of experiments described the reasons why experiments are important, the elements of the design that will lead to valid results, and ways of making the study efficient. These do not depend on who is participating or why. They depend on the logic of replicability and reliability of the results. Thus it will be important to ensure that the experimental design takes account of this, while not compromising on participation. For example, just what are the treatments being compared? The new varieties will be grown and compared with existing ones. Researchers and farmers together can agree on which control varieties should be included. Not every farmer needs to test all the new varieties or all the controls – that sort of 'balance' is *not* a requirement of a sound experiment. But the design will have to include sufficient replication of the various comparisons that are important – that *is* a requirement.

If some farmers grow cowpea monocropped and some intercropped, then comparing the new varieties under these two conditions may be added to the objectives. This will increase the range of treatments needed. Your approach to participation may still have farmers select the actual treatment combinations they test. What do you do if no one wants to test some of the treatments? Find out why. Maybe farmers know something you don't and recognise that those treatments cannot work. Or maybe after discussion some farmers will decide that testing them is a good idea. Or maybe you have to add some of you own plots to look at those treatments. All these things can add to the practical complexity of carrying out the

experiment and analysing the data. But they do not alter the underlying requirements for a good experimental design.

The same is true for surveys, the other way of collecting data. Here it is maybe even more important to recognise the elements of sound design when using participatory methods. Many of the tools used in participatory research are actually survey tools, so think about their use in terms of survey principles. For example, focus group discussions can be very valuable in understanding local conditions, problems and opportunities, particularly when linked to such techniques as resource mapping and wealth ranking. They give real insights into the villages with which you are working. And they give the participants themselves insights into their own situation. If that is your objective, then you can be flexible in how and where the tools are used But do you want to know how broadly applicable those results are, and the extent to which they are representative of a larger population? If so, then use sampling techniques to select the villages in which you work, and use the ideas of survey management to make sure that information is really comparable across different villages.

The final point to make about the use of participation in a research study is 'Beware of packages!'. There are many guides to participatory research that present a packaged set of tools and processes, based on what the author has found to work. But that author's objectives and circumstances will not be the same as yours, so do not expect to be able to follow the same steps and use exactly the same tools. You have to learn to pick and choose the methods and tools that meet your objectives. As a simple example, we have found that participatory tools for matrix ranking to assess alternatives, based on the traditional mbao game, can be used effectively to evaluate an on-station, researcher-designed experiment.

Resource material and references

Appendix 3. Designing Research Around Client Needs. Paul L. Woomer. PowerPoint on CD

Appendix 11. ICRAF. 2003. *Genstat Discovery Edition and Other Resources*. World Agroforestry Centre (ICRAF), Nairobi, Kenya. On CD.

Chambers, Robert 1992. *Rural Appraisal: Rapid, Relaxed and Participatory*. Institute of Development Studies (IDS) Discussion Paper 311, UK.

Franzel, S. and Coe, R. 2003. The balance between researcher and farmer involvement in technology testing. In: *Designing Participatory On-farm Experiments: Resources for Trainers*. Coe, R., Franzel, S., Beniest, J. and Barahona, C. (Eds.), pp 52–63. World Agroforestry Centre (ICRAF), Nairobi, Kenya.

There are many African universities involved with participatory research and you are encouraged to consult their publications. Two university institutes in the region with particular strengths in, and experience with participatory research include:

Centre for Applied Social Sciences, University of Zimbabwe, Zimbabwe
http://www.cassuz.org.zw/

University of Western Cape, South Africa
http://www.uwc.ac.za/

The online 'Participation reading room' at the Institute of Development Studies (IDS) has a large amount of material on participation in many different contexts
http://www.ids.ac.uk/ids/particip/information/readrm.html

The World Bank also maintains extensive online resources on participation
http://www.worldbank.org/participation/participation/participation.htm

PLA notes is one of the leading periodicals on particpatory work, available from the International Institute for Environment and Development (IIED), UK
http://www.iied.org/sarl/pla_notes/whatispla.html

The PRGA program of the CGIAR is continually developing new material
http://www.prgaprogram.org/

IIRR is an example of an NGO producing guides on doing participatory work
http://www.iirr.org/

If all else fails, resort to ELDIS, the most comprehensive online development library with many documents on participation.
http://www.eldis.org/

5 Where to from here?

Kay Muir-Leresche

- **The more people able and determined to contribute to sustainable development the better**
- **Your individual ideas and actions do count. You can make a difference**
- **Creativity and adaptability are essential criteria for successful economies in a rapidly changing and global environment**
- **You can contribute to both the poor and to your own advancement with imagination**
- **Experience and a track record are important for getting jobs — you may need to start off in a menial position or doing voluntary work to establish your credentials**
- **It is possible to be an entrepreneur even without capital**
- **Blend modern and traditional, indigenous and conventional**
- **Be proud of your heritage, understand the limitations and grasp the opportunities**

'...the initiation of all wise and noble things comes...,generally at first from some one individual.'

John Stuart Mill
Representative Government

'...there will be no injustice in compelling the philosophers who grow up in our state to have a care for the others.'

Plato
The Republic VII

Where to from here?

You have finished your thesis, you have had it examined and you have undertaken your corrections — it is bound and you have a copy which you proudly present to your family.

What now?

Remember when you started out how unsure you were – 'How will I ever manage that?', you thought? Well you have – and you have grown in the process. You have developed skills and most important of all, you have gained confidence. After the strain of producing your thesis, you may be feeling a bit flat and uninspired. You probably need to renew your enthusiasm and burning desire to contribute.

Confidence, the ability to use your own initiative and the inspiration and determination to make a difference, are the most valuable of all resources any country can have.

The more people there are determined and able to contribute to sustainable development, the better. You only have to look around Africa, or the world, to see that it is not necessarily rich mineral resources, nor rich agricultural land, nor rich coastal waters that make countries better able to provide for their poor. It is their human capital. It is their commitment to success and their ability to respond quickly – to be creative and adaptable – so that they can take advantage of changing technology, institutions and social relations. It is essential for us to start to take charge of our destiny; to succeed in developing our countries. We Africans want our children to grow up in an environment where they are able to chart their own course and do not feel hopeless. We need to be able to move away from corrupt practices that obtain short-term advantages. We need to earn our incomes by providing goods and services which in turn will develop Africa. We need to be able to

take control of our development at the personal, village, national, and regional levels.

Finding work that provides you with money and prestige are common goals. They are important. We need that money to live, to repay our families who invested in our education and to provide for our futures. Social standing can be important to many people – but remember that fashions change and what is prestigious today may not be in 10 years' time. Happiness, however, is not limited to wealth and fame. There is considerable personal satisfaction from contributing to society. If you can make a lasting difference to the lives of the poor, to the development of your country, or even to one student or one farmer or one village, you will be able to look back when you are 80 and say – Yes, I did make a difference!!

These goals do not have to be mutually exclusive.

Example 1

Nyasha is hired by Norsk Agricultural Chemical Co. to promote the sale of fertilizers and pesticides. She earns her salary by selling to conventional markets and using the established recommendations. Perhaps she remembers that when she was doing her graduate research, many small-scale farmers could not afford to apply fertilizer at the recommended rate. She has heard of someone who has adapted the established recommendations to more closely suit small-scale, poor farmers. So in her spare time, she contacts them and then draws up a marketing strategy that would provide farmers with access to this new information. She has to persuade the company that, although these recommendations are for lower fertilizer use, they will make it possible to sell to many more farmers.

In this example the sales agent used her initiative and commitment to change things for the better. She also advanced her career. You can all do this. In every job you do, it is possible to make the world a little better for the future. You need to believe in your own power. You need to learn to be a self-starter and to be prepared to put in that extra effort.

You need to take an ethical stand. Do not allow your valuable skills to be used to further corruption and the cheating of your fellow citizens. Do not contribute to the degradation of the environment and the impoverishment of future generations. We owe it to our children to leave a better world than we found, and you can make a difference.

Employment

The formal sector

Remember that employment does not necessarily mean working for someone in return for a salary. Employment means using your skills and labour to produce output that will have financial and other rewards. Professional jobs for new graduates in Africa have become increasingly difficult to find, despite the considerable shortage of skills. This is because for some 50 years, governments employed new graduates. They would obtain practical experience and learn to operate in the working world that gave them credibility and led to formal employment in the private sector. Decentralisation, declining government budgets and reduced investment in research, extension, and education have all contributed to shrinking these opportunities in most African countries. At the same time, the private sector is reluctant to hire untested graduates. In most countries very strict legislation makes it difficult for employers to release staff once they have been hired and as a result they are very risk-averse in their employment policies. You are required to be much more innovative than your parents in seeking employment.

You need to get together some evidence of your ability. Take a copy of your thesis and of a few other projects or papers you have produced. Ensure that you include the extra-curricula activities with which you have been involved and any positions of leadership or trust which you may have held. Speak to the people you are going to use as referees and be sure they are happy to do this. Provide them with a copy of your CV so that it is easier for them to write the reference.

Find out about the company before you go for an interview. See where you think you would fit in. At the interview you should not be arrogant but you must make an opportunity to be able to tell them how you think you could contribute to their organisation. For example, if the job involves selling tractors, you might mention your contacts from your home area who may be interested clients – or mention your experience working in a garage during one of your vacations. If it is project management and budgeting, you could mention your role in the university agricultural student society. If you don't have anything specific you could offer, at least be sure you understand what the organisation does and show that you have thought through how you could play a role within it.

Something else

If you are unsuccessful in obtaining formal sector employment, you should seriously consider voluntary service as a stepping stone. Most prospective employers would be prepared to provide you with basic transport and food costs. If you cannot find a company to hire you even on these terms, then prepare a research proposal and contact relevant NGOs, government research departments or even churches. Do not be ambitious for a high financial reward even when you are contacting an international agency. Remember this first 'job' is more to establish your credibility and gain experience than to provide an income. Impress the prospective benefactor with the fact that you are prepared to sacrifice in order to get ahead in the future and to contribute to your society. You need to realise that the world does not owe you a living and that you have to be creative in getting that first job. Once you have experience, if you prove yourself, it will be much easier to move up the ladder.

For many African students this is difficult. Their families have invested resources in the graduate's education and now they expect that person to start to contribute to the family. Prepare your family. Show them your strategy ahead of time and I am sure you will find them much more understanding.

Even if no-one is prepared to take you on, even as a volunteer, you may then have to go and take a much more menial job. Look at it positively as a stepping stone and be constantly on the look-out for how you can contribute to the success of the organisation for whom you are working. It is surprising how many highly successful people have started in very menial positions.

Increasingly in Africa the best way to get ahead is to become an entrepreneur yourself. How you go about this will depend on the contacts and resources you have. If you are able to raise capital then you can be more adventurous. If you cannot raise any capital then start very small. Identify a need and provide for it even in a very small way.

Example 2

Tapiwa realised that there would be no bread available in the following year. He knew that urban workers would need to have convenience food that they could afford. He went to his aunt in the rural areas and asked her to provide him with some sweet potatoes and promised to repay them when he harvested his own crop. He went and read up all the literature on

sweet potatoes and learned what he could about their preferred soil types, mineral requirements and the most ideal moisture conditions. He could not afford fertilizer but he approached the people in his street and asked them if they would put all their vegetables and other wet refuse into bags for him. He would collect it and this would reduce the unpleasantness of such refuse left out on the road for days. He also collected newspapers and on a vacant lot he made a compost pit. As a result he had a bumper harvest of high-quality sweet potatoes that fetched a high price because of the need he had identified. In due course he became a successful market gardener, bought his own plot and was able to employ workers.

Agricultural research and our commitment to sustainable development

Most of students with post-graduate degrees will eventually go into work that involves research. The ultimate goal of research is to search for the truth. Thus, we make a moral commitment when we undertake research and we need to honour that. We must avoid any fabrication or falsification of information and data. We must be sure that we set the highest standards for ourselves and that we maintain our integrity. This will require our research to be as objective as possible. It will mean that we need to closely supervise the collection and entry of our data. Most important of all, we must remember that the work we produce will be used to affect the lives of people who live at the margin. A small error can tip them into very serious poverty or even starvation. At the same time we need to respect their abilities, their privacy and their needs. We need to listen to them and to try to establish research, policies, and implementation strategies that empower the disadvantaged. The results may be slower but they will be more sustainable.

We need to find ways in which we can adapt some of the modern technologies so that they can be used despite the constraints facing the users. We need to adapt traditional norms and values so that they can accommodate new technology. There is much scope for blending modern and traditional, conventional and indigenous and of finding ways to commercialise, improve and extend the use of traditional commodities. We need to find better ways to use our resources so that we do not endanger our environment. Be proud of your heritage, understand the limitations but grasp the new opportunities. Graduates must take a pride in creativity and in their ability to make something different by using both the old and the new. Technology is changing constantly and global competitiveness requires the ability to innovate rapidly (Porter and van der Linde, 1996).

Resource material and references

Foster, Michael B. 1942. *Masters of Political Thought: Plato to Machiavelli.* Harrap & Co., London, UK.

Lancaster, Lane W. 1959. *Masters of Political Thought: Hegel to Dewey.* Harrap & Co., London, UK.

Porter, M. and van der Linde, C. 1996. Ending the stalemate. In: *Business and the Environment*, Richard Welford (Ed). Earthscan, London, UK.

Contributors

Gerald W. Chege is a Kenyan with a PhD in Parallel Computing from York University, UK. He is presently Assistant Professor and Coordinator of the Information Systems and Technology Department, United States International University, Nairobi. His main area of interests are: computer networks, database technology; systems development, and Internet technology.

Richard (Ric) Coe is an Applied Statistician from the United Kingdom. He gained an MSc in Biometry from the University of Reading, where he continued as a lecturer for 10 years. During that time he was involved in a number of training and research projects in Africa and Asia. In 1990 he moved to the World Agroforestry Centre (ICRAF) in Nairobi, Kenya. There he is Head of the Research Support Unit that provides technical support and training in research planning and design, data management and analysis to all ICRAF projects and partners. His interests are in making research for development as effective as possible through the use of sound methodology, and increasing capacity in Africa to do this. He has taught courses at several universities in the region and has worked with hundreds of graduate students on their research projects.

Tony Greenfield, a graduate in Statistics from London University with a PhD in Experimental Design from Sheffield Hallam University, was formerly Head of Process Computing and Statistics at the British Iron and Steel Research Association, Sheffield, and Professor of Medical Computing and Statistics at Queen's University, Belfast. He is a Visiting Professor to the Industrial Statistics Research Unit (ISRU), at the University of Newcastle-upon-Tyne and is past President of European Network for Business and Industrial Statistics (ENBIS). While at Queen's University he established a course in research methods for the medical faculty. His publications include *Research Methods: Guidance for Post-Graduates* (Editor and co-author), first published by Edward Arnold in June 1996, second edition in June 2002. This book is used in some English universities in courses for postgraduates who intend to proceed to research degrees.

Thomas Gumbricht is a Swedish Hydrologist working with the World Agroforestry Centre. He holds a PhD in Land Improvement and Drainage from the Royal Institute of Technology (KTH), Sweden, and prior to his arrival in Kenya was Head of Geoinformatics at the Department of Earth Sciences, Uppsala University, Sweden. His main interests are systems ecology and hydrology; using geoinformatics as a platform for understanding and modelling processes on a landscape scale.

Sue Hainsworth has been editing all her working life. After graduating in Agricultural Sciences from Nottingham University she edited the Tropical Pest Management Journal and wrote the first three titles in the Tropical Pest Management Manual series on bananas, groundnuts, and rice. After time in Rome with the Food and Agriculture Organization of the United Nations (FAO) and the International Plant Genetics Resource Institute (IPGRI, then IBPGR) in 1983 she joined the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) and rose to become Manager, Publications before leaving to start her own Editorial and Publishing Services in 1998.

Erica Keogh is a Zimbabwean, holding an MSc in Statistics (University of Zimbabwe, 1987). She has been employed as a lecturer at the University of Zimbabwe since 1980 but is currently on long leave while she engages in a long-term consultancy with the UK's Department for International Development (DFID) monitoring their Humanitarian Relief Programme in Zimbabwe. Since

the early 1990's she has become increasingly involved in applications of statistics and has had extensive experience in the design and implementation of surveys in both rural and urban areas, focussing mainly on issues of poverty and related aspects of social change.

Eric McGaw, an American national, has lived and worked in the developing world for over 30 years. He graduated from Rockford College with a degree in Fine Arts, and completed post-graduate work in Education at Boston State College, USA. After serving in the Peace Corps in El Salvador, he worked as a university professor, a deep sea diver, a freelance writer and editor, and a communication specialist in Colombia, Brunei, Singapore, the Philippines and India. He has travelled widely throughout Latin America, Asia and Africa. Currently, he is employed as Head of Communications at ICRISAT located near Hyderabad, India.

Kay Muir-Leresche is a Zimbabwean with 23 years' developing-country experience in agricultural and natural resources higher education, research, training, and policy analysis. Her main focus since 1981 was as an Economics Lecturer in the Faculty of Agriculture, University of Zimbabwe. When she left in 2002 she held the Professorial Chair in the Department of Agricultural Economics and Extension with major responsibility for the supervision of doctoral candidates. She taught both undergraduate and graduate programmes and supervised student dissertations and she served on the Faculty Higher Degrees Committee for 15 years.

Peter K. Muraya is a Kenyan with a BSc in Electronic Computer Systems Engineering from Loughborough University, UK. Presently he is Data Management Specialist with the responsibility of leading the World Agroforestry Centre's initiative to bring management of research data to agreed standards in all regions and projects. He was earlier involved with the development of simulation models for agroforestry systems. His main area of interest is in conceptualising, design and implementation of data management methods and software tools.

Liliosa Maveneka has a BSc in Mathematics and Botany and an MSc in Agricultural Economics. She also is an Associate Member of the Institute of Chartered Secretaries and Administrators. She has worked as a Registrar in the Faculty of Agriculture and as a Senior Administrator for the University of Zimbabwe for 15 years. She is a consultant working on HIV/AIDS impacts, in issues related to water allocation and pricing and in providing assistance in accessing Internet data to post-graduate students and researchers.

Catherine Wangari Muthuri is a Lecturer in the Department of Botany, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Kenya, where she earlier gained an MSc in Botany (Plant Physiology). Catherine submitted her PhD thesis on the 'Impact of agroforestry on crop performance and water resources in semi-arid Central Kenya' for examination in September 2003. She carried out her research work for 3½ years at the World Agroforestry Centre. For the past 10 years, Catherine has been involved in teaching, administration and research at JKUAT. Her research interests include environmental plant physiology in agroforestry and non-agroforestry systems with particular emphasis on drought stress and the application of agroforestry models in research.

Joseph Opio-Odongo is a Ugandan holding a PhD in Rural Sociology from Cornell University, USA. He is one of the United Nations Development Programme's (UNDP's) Environmental Policy Specialists, out-posted to Nairobi to provide technical backstopping to UNDP Country Offices in sub-Saharan Africa on policy and programme development. He previously served as a Sustain-

able Development Advisor at the UNDP Country Office in Uganda after some years of teaching at Makerere University in Uganda and Ahmadu Bello University in Nigeria. His research and teaching experience has been mainly in the fields of agricultural and rural development. His research and development interests include policy analysis, empowerment of civil society, sustainable development, organisational development, science and technology policy, lobbying and advocacy, and the codification and application of indigenous knowledge and technology.

Bharati K. Patel has been working in the Food Security Division of The Rockefeller Foundation in Africa for the past 10 years. As an Associate Director she ran the Forum on Agricultural Resource Husbandry a competitive grants programme designed to encourage and support research on agricultural resources. The programme supported the staff in ten Faculties of Agriculture in Kenya, Malawi, Mozambique, Uganda, and Zimbabwe in their training of graduate students. A Zambian with a primary degree in Botany and a PhD in Nematology from the Waite Institute in Australia, Bharati also worked for the Zambian Agricultural Research System for 20 years where she rose to become the first woman Director of Agricultural Research in Africa. She also worked in ICRISAT prior to her assignment with The Rockefeller Foundation.

Aleya Pillai is an Indian, with a Commercial Arts Diploma from Jawaharlal Nehru Technological University, Hyderabad, India. She worked as Art Director from 1984 to 2001 with various Indian advertising agencies, the last one being Mindset in collaboration with Saatchi & Saatchi. During that time she designed several award-winning annual reports, calendars, press ads and logos. She is no newcomer to cartoons, having used them to get across scientific concepts in the past. Aleya also enjoys cartooning and oil painting in her spare time.

Jane Poole is a British national and holds an MSc in Biometrics (Applied Statistics) from Reading University, UK. She has recently returned to the UK after 6 years of working in Africa, where she provided biometrics support to scientists and research students at the World Agroforestry Centre and CAB International (Africa Regional Centre), both based in Nairobi, Kenya. Jane currently works at the UK Forest Research Agency with scientists covering a wide range of disciplines: from forestry, entomology, and pathology to ecology and environmental research. Jane has wide experience in experimental design and analysis and in small and large-scale biological and socio-economic surveys. She enjoys working with students and scientists from many disciplines, learning about their research and working with them as a member of the research team throughout their projects.

Sue J. Richardson-Kageler is the Biometrician in the Faculty of Agriculture, University of Zimbabwe, where she previously taught in the Statistics Department. One of the major components of her job is advising undergradutes and postgraduates on the design of their experiments. A Zimbabwean by nationality, Sue has a DPhil in Ecology which examined the changes in woody plant diversity brought about by large herbivores and an MSc in Biological Computation from the University of York.

Jayne Stack is a Senior Lecturer in the Department of Agricultural Economics and Extension, University of Zimbabwe and has more than 20 years experience in development training, development programmes and research in Africa and Asia. She has taught research methods at undergraduate and postgraduate level and contributed to the development of distance-learning courses in research methods and data analysis for Imperial College, London. Jayne has a wide interest in development issues ranging from crop marketing to agricultural policy reform, house-

hold food security and livelihood analysis. Her research work aims to make a difference in the lives of the poor and to contribute information that will enhance livelihood security of vulnerable households.

Paul L. Woomer is a researcher working with the Sustainable Agriculture Centre for Research Extension and Development in Africa (SACRED-Africa), a Kenya-based NGO. One of his major interests is the adaptive research process where different potentially useful technologies are compared, combined and refined to suit the needs of individual farmers. He has written or edited four books and published over 90 papers or chapters in international journals and multi-authored books. Paul was raised in the Hawaiian Islands where he developed a keen interest in tropical crops and ecology. He attended the University of Hawaii, where he obtained a BSc in Agronomy and a MSc and PhD in Soil Science. He previously worked with NifTAL-MIRCEN, TSBF-UNESCO, The Alternatives to Slash and Burn Consortium and the University of Nairobi. Paul has lived in Kenya since 1990 and has visited or worked in 18 different African countries.

Acronyms and abbreviations

ACSS	African Crop Science Society
ADDS	African Data Dissemination Service
ACT	almanac characterisation tool
AFRENA	Agroforestry Research Network for Africa
AHI	African Highlands Initiative
AI	appreciative inquiry
AKIS	Agricultural Knowledge and Information System (World Bank)
ANOVA	analysis of variance
AVHRR	Advanced Very High Resolution Radiometer
CARPE	Central African Regional Program for the Environment
CBO	community-based organisation
CD	compact disk
CGIAR	Consultative Group on International Agricultural Research
CIESIN	Center of International Earth Science Information Network
COSOFAP	Consortium for scaling up options for increased farm productivity in Western Kenya
CRSP	Collaborative Research Support Program (USAID)
CRU	Climate Research Unit (University of East Anglia, UK)
CTA	Technical Centre for Agricultural and Rural Cooperation (the Netherlands)
DCW	Digital Chart of the World
DEPHA	Data Exchange Platform for the Horn of Africa
DEM	digital elevation model
DFID	Department for International Development (UK)
DMA	Defense Mapping Agency
DRASTIC	depth of groundwater, recharge, aquifer media, topography, impact of root zone, conductivity
DRC	domestic resource cost
DSMW	Digital Soil Map of the World (FAO)
DSS	decision-support system
ENBIS	European Network for Business and Industrial Statistics
ESA	European Space Agency
ESRI	Environmental Systems Research Institute, Inc.
ETM	Enhanced Thematic Mapper
FAO	Food and Agriculture Organization of the United Nations
FEWS	Famine Early Warning System (USAID)
GDP	gross domestic product
GIS	geographic information system
GPS	global positioning system
GUI	graphical user interface
IARC	international agricultural research centre
IBPGR	International Board on Plant Genetic Resources (now IPGRI)
ICIPE	International Centre for Insect Physiology and Ecology
ICRAF	World Agroforestry Centre

ICRISAT	International Crops Research Institute for the Semi-Arid Tropics
IDS	Institute of Development Studies (UK)
IDW	inverse distance weight
IPCC	Intergovernmental Panel on Climate Change
IPM	integrated pest management
INASP	International Network for the Availability of Scientific Publications (UK)
IPGRI	International Plant Genetic Resources Institute
ISSER	Institute of Statistical, Social and Economic Research
KARI	Kenya Agricultural Research Institute
KEFRI	Kenya Forestry Research Institute
LFM	logical framework matrix
MAP	methods of active participation
MCE	multi-criteria evaluation
MODIS	Moderate Resolution Imaging Spectroradiometer
NARES	National Research and Extension Services
NARO	National Agricultural Research Organisation (Uganda)
NARS	national agricultural research systems
NDVI	Normalised Difference Vegetation Index
NGO	non-governmental organisation
OECD	Organization for Economic Cooperation and Development
PAR	participatory action research
PI	principal investigator
PRA	participatory rural appraisal
R&D	research and development
RELMA	Swedish-supported Regional Land Management Unit
ROM	read-only memory
RS	remote sensing
SACRED	Sustainable Agriculture Centre for Research Extension and Development in Africa
SAS	statistical analysis system
SPSS	statistical program for social sciences
SRTM	Shuttle Radar Topography Mission
SSA	sub-Saharan Africa
TEEAL	The Essential Electronic Agriculture Library
TM	Thematic Mapper
TSBF	Tropical Soil Biology and Fertility
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNSO	United Nations Statistical Organization
USAID	United States Agency for International Development
USGS	United States Geological Survey
VCD	visual compact disk
WARDA	West Africa Rice Development Association
WBS	work breakdown structure
WFP	World Food Programme
WRI	World Resources Institute

Appendices on the CD

Appendix 1. The Craft of Research. Paul L. Woomer. PowerPoint

Appendix 2. Innovation, Problem Solving and Operational Research Strategies. Paul L. Woomer. PowerPoint

Appendix 3. Designing Research Around Client Needs. Paul L. Woomer. PowerPoint

Appendix 4. Preparing and Refining a Research Proposal. Paul L. Woomer. PowerPoint

Appendix 5. Stapleton, P., Youdeowei, A., Mukanyange, J. and van Houten, H. 1995. *Scientific Writing for Agricultural Research Scientists*. WARDA/CTA, Ede, The Netherlands.

Appendix 6. Publication as an Output of Science. Adipala Ekwamu. PowerPoint

Appendix 7. The Art and Ups and Downs of Scientific Publication. Adipala Ekwamu. PowerPoint

Appendix 8. Presentations and Style – Tips on Photography and Writing. Eric McGaw.

Appendix 9. Muraya, P., Garlick, G. and Coe, R. 2003. *Research Data Management*. World Agroforestry Centre (ICRAF), Nairobi, Kenya.

Appendix 10. Coe, R., Stern, R., Allan, E., Beniest, J. and Awimbo, J. 2002. *Data Analysis of Agroforestry Experiments*. World Agroforestry Centre, Nairobi (ICRAF) Kenya.

Appendix 11. ICRAF. 2003. *Genstat Discovery Edition and Other Resources*. World Agroforestry Centre (ICRAF), Nairobi, Kenya.

The African Crop Science Society
P.O. Box 7062, Kampala, Uganda

ISBN 9970-866-00-1