Part 4

Data management and analysis

4.1 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 rows and columns of numbers in your numerous data files. This is important both 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.

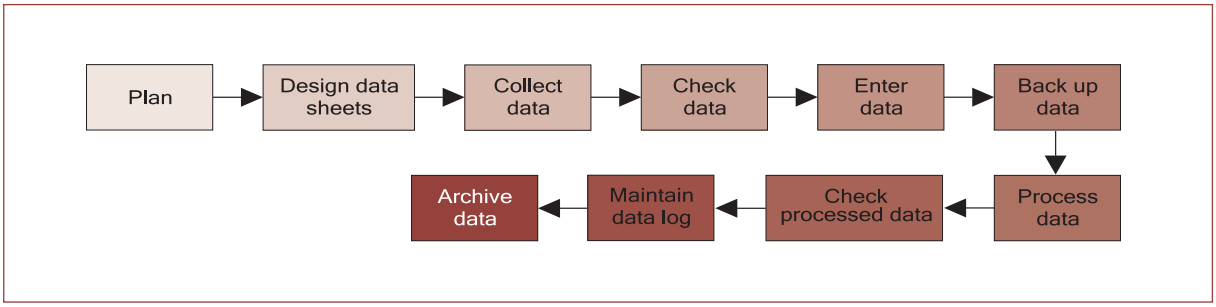


Figure 1. Data management processes

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.

As a student, your research will probably be part of a larger study. Your data management will therefor influence and depend on the data management in the rest of the project. For example, you may need data collected by others, and they may need yours. Muraya and Chege (2007) discuss some of these larger scale data management issues.

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 following sections assume that most of your data is numeric, or can be coded as such. Not all research data is of this type. For example, in some social research you may collect narratives – free-ranging stories. These will be recorded as audio or textual transcriptions. The tools described here may not be immediately applicable to this type of data. However the principles are much the same.

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/Strigaresearch/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.

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

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 subplot. The solution is simple: create a new sub-plot column and fill it accordingly. The other 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. This type of data layout is also a problem if you transfer the data to a statistical computing system such as Genstat or SPSS, as these applications will take the blanks as missing 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

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

Program	Domestication of Agroforestry Trees					
Statistical Design	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

the cells 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.

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.

Adding comments to cells

Unusual instances occur in data capture and subsequent entry. For example, when no value is

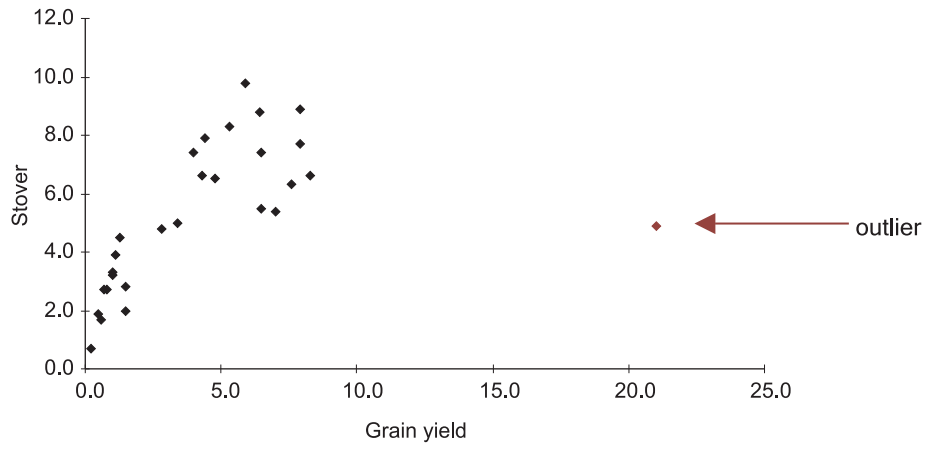


Figure 3. Scatter plot example showing data outlier (for details see Appendix 11)

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.

C	D	E	F	G	H
species	rcd	height	branch	Crown_0	Crown_90
A. polycantha	12.7,13.3	438	23	673	730
A. indica	15.1	415	17	374	354
A. nilotica	11.1				
Albizia lebeck	21.1				
Control					
A. indica	15.1	470	19	420	395
control					
Albizia lebeck	12	300	12	394	322
A. polycantha	DEAD				
A. nilotica	10.1	343	22	420	401
A. nilotica	10	330	23	443	402
Albizia lebeck					
Control					
A. indica	11	410	21	415	440
A. polycantha	14.25	635	23	852	880
Control					
A. nilotica	12.5	373	23	602	500
A. polyantha	25.8	630	25	920	750
A. indica	18.5	404	22	420	370
Albizia lebeck	198	465	10	352	340

Auditing

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

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.

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 CD, DVD or an external hard drive. 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 11), 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.

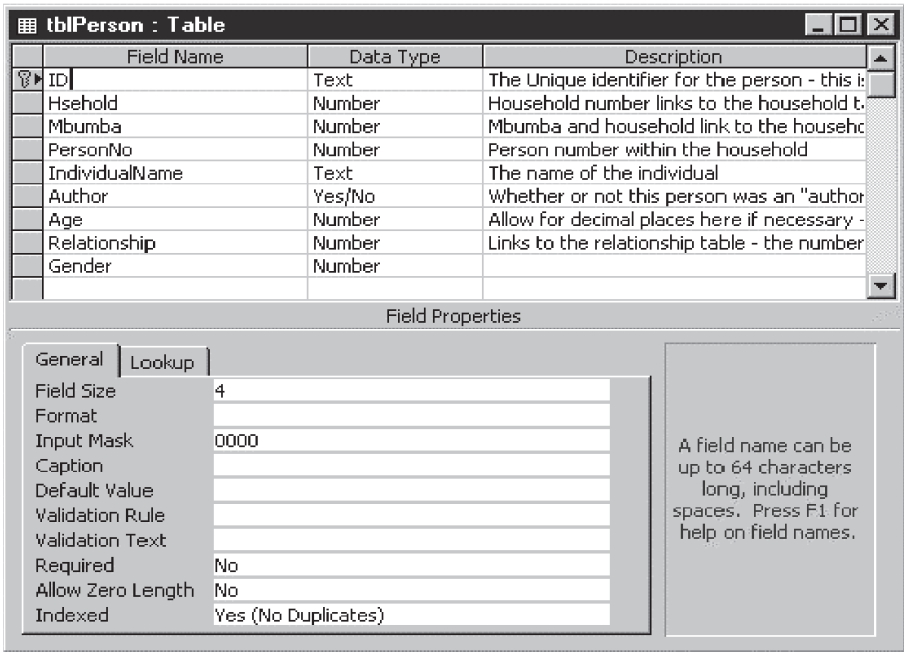
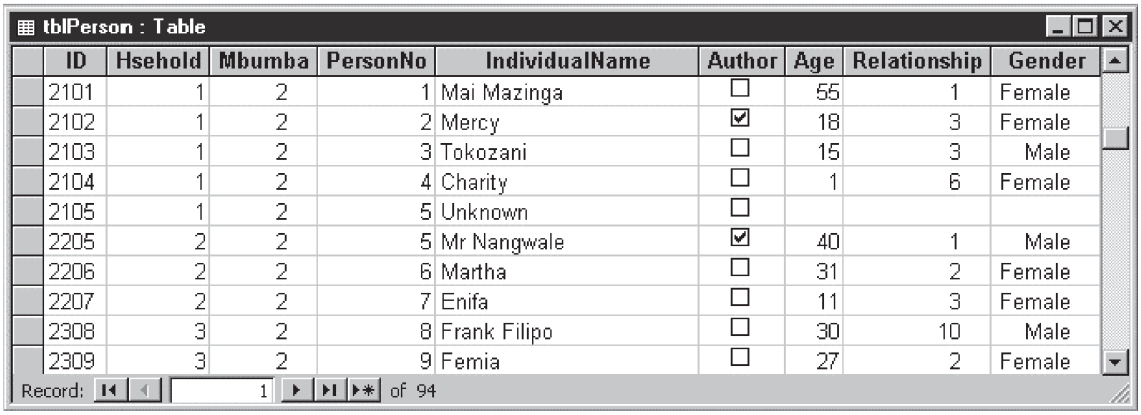


Figure 6. Table design in MS-Access for a person-level table

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	Mbumba	PersonNo	IndividualName	Author	Age	Relationship	Gender
2101	1	2	1	Mai Mazinga	☐	55	1	Female
2102	1	2	2	Mercy	☑	18	3	Female
2103	1	2	3	Tokozani	☐	15	3	Male
2104	1	2	4	Charity	☐	1	6	Female
2105	1	2	5	Unknown	☐			
2205	2	2	5	Mr Nangwale	☑	40	1	Male
2206	2	2	6	Martha	☐	31	2	Female
2207	2	2	7	Enifa	☐	11	3	Female
2308	3	2	8	Frank Filipo	☐	30	10	Male
2309	3	2	9	Femia	☐	27	2	Female

Record: 1 of 94

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 crosschecking 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 CDs or DVDs as distribution media is of course an option.

To show the seriousness of data archiving and its place in research, it is now possible to publish peer-reviewed data papers (http://www.esapubs.org/archive/).

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.

Ecological Archives. A Publication of the Ecological Society of America. http://www.esapubs.org/archive/

Coe, R. 2001. *Audit Trail in Research Data Management*. (Draft Report), World Agroforestry Centre (ICRAF), Nairobi, Kenya. Some useful data management documents available at http://www.ilri.org/rmg/

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

Muraya P and Chege G. 2007. *Handing Research Data*. Nairobi, World Agroforestry Centre.

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).

4.2 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.

A different intercropping experiment may have different overall objectives, and hence analysis objectives. For example, the quantity known as the 'land equivalent ratio' (LER) is often used in analysis of intercropping, but is not relevant in this particular example.

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 4.1**
- Data must have been checked – see **Chapter 4.1** – 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, 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? In the pumpkin experiment, plant height was measured on a sample of 10 maize plants in each plot. This 'plant-level' data needs summarising to the 'plot-level' before the analysis proceeds. In this case, simply find the average plant height per plot. 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.

Compatibility means that you can move the information from one application 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 tool 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 system such as Genstat or R, can be copied and pasted into your project report.

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).

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 **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.

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.

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 Eb 4 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 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?

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**

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**

Maybe 'village' is only recorded as part of the logistics of data collection, and need not appear in your summary tables. On the other hand, if large differences between villages appear, that is part of the pattern and needs to be recognised in the analysis, and ideally explained and understood. **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.

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

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

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

Crop	Weeding 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**

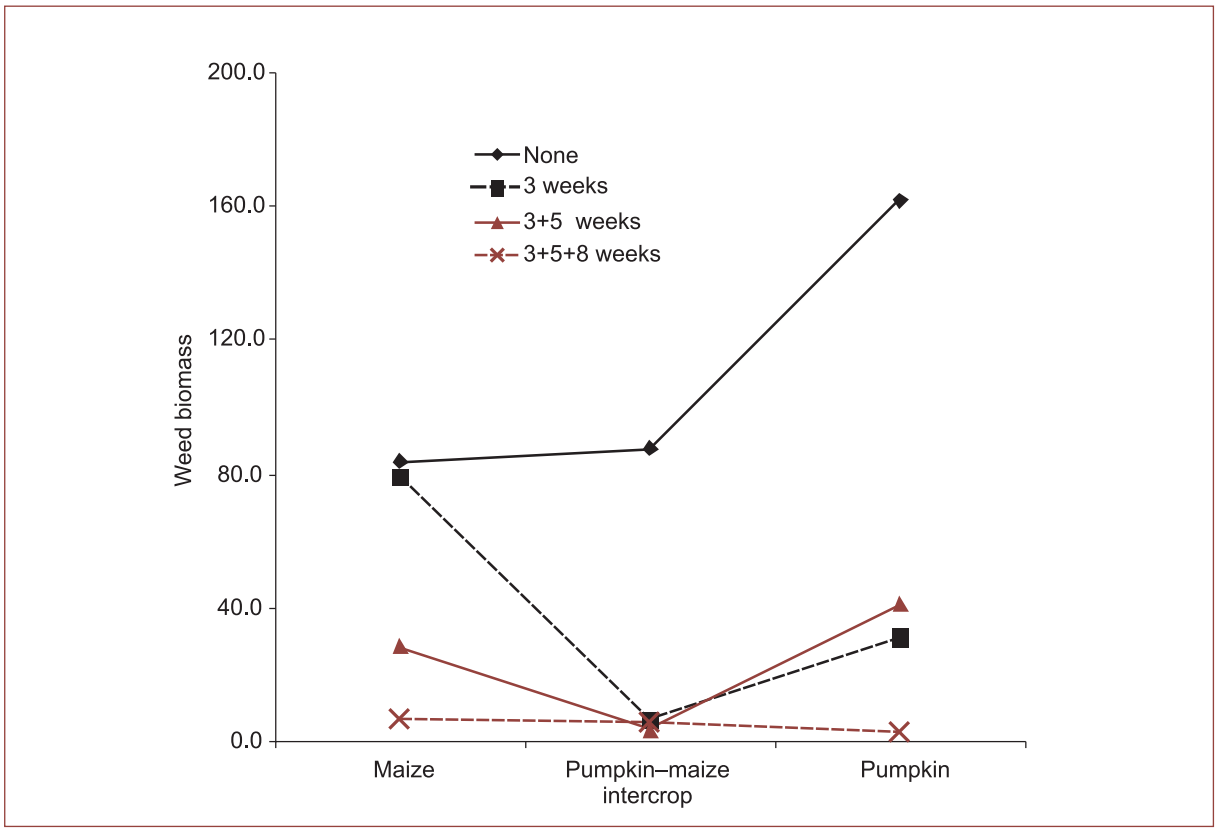


Figure 1. Graph of data in Table 4

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 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

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

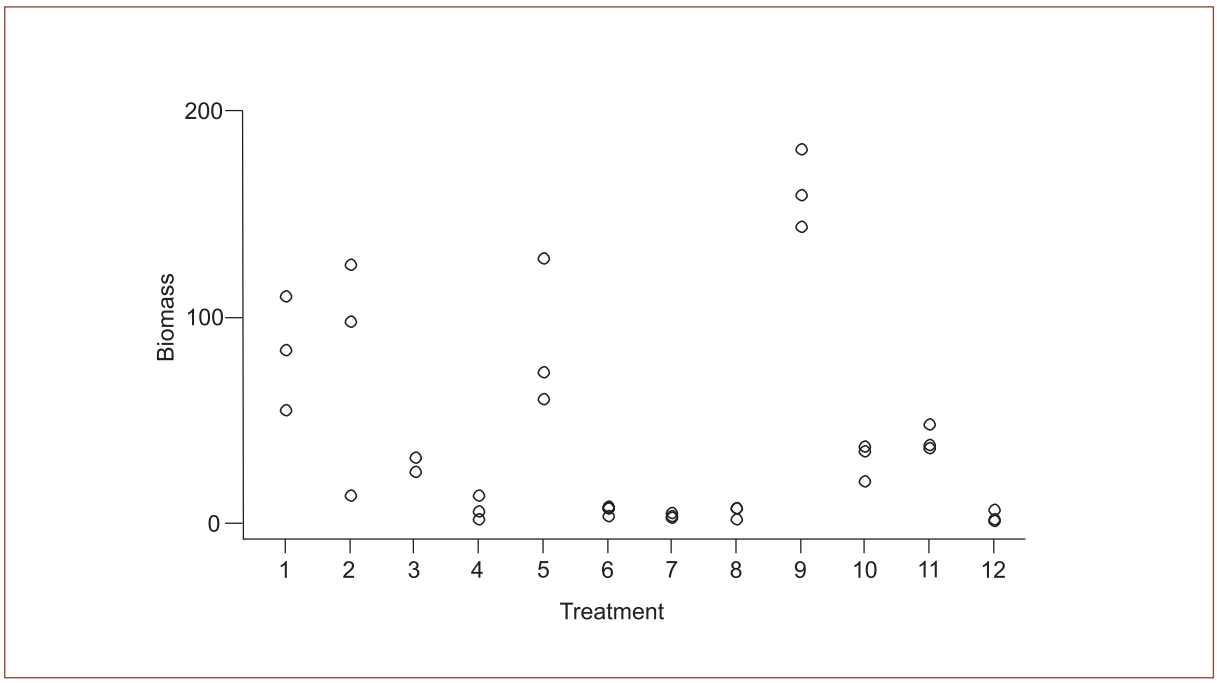


Figure 2. Dotplot of the weed biomass for the pumpkin-maize intercrop experiment showing each of the treatments

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 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

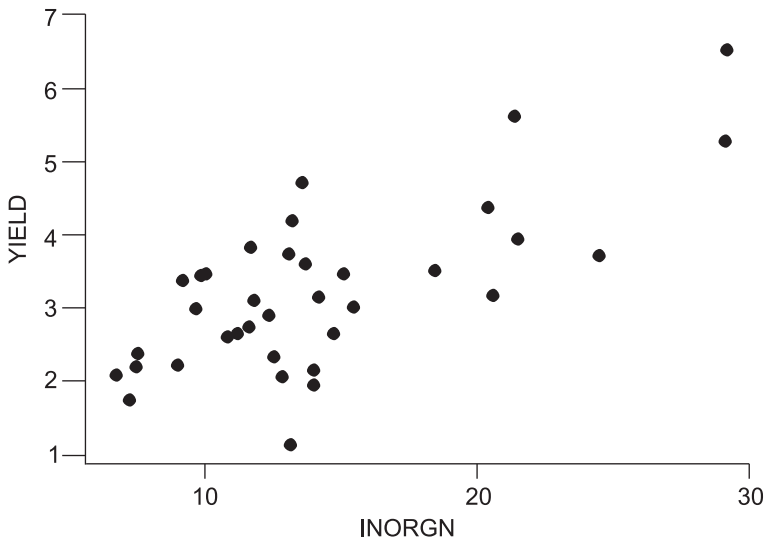


Figure 3. Relationship between soil inorganic nitrogen and crop yield

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 book on analysis of designed experiments. Some examples are listed at the end of this chapter.

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.

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

a.	Treatment	Mean weed biomass (g/m^2)[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/m^2)		
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

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 – those letters attached to the means – 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.

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: Part 1 | **Part 2** | **Part 3** | **Part 4** Coe, R., Stern, R., Allan, E., Beniest, J. and Awimbo, J. 2002. *Data Analysis of Agroforestry Experiments*. World Agroforestry Centre (ICRAF), Nairobi, Kenya.

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 Methods Group of the World Agroforestry Centre (ICRAF) have some training materials and other guides on analyses. http://www.ilri.org/rmg/

4.3 Econometric models for varied contexts

Chris Sukume

- **Context has a significant effect on the relationship between variables**
- **When using a model from elsewhere, adapt it to your context**
- **There are powerful free tools for analysis available – SPSS may not be the answer**
- **Consider what your model results mean for rural development**
- **Answer the so what question … what can we do, or know, now that we didn't before the research?**

Introduction

This chapter is about the quantitative analysis, modeling and interpretation of data related to the economics of farms and the environment. It assumes that graduate students have taken courses in basic statistics and intermediate econometrics courses prior to undertaking research. However, most econometrics courses glaze over the practical aspects of using econometrics tools. The purpose of this chapter is to introduce students to the practical suggestions to help them in econometric model building, estimation, interpretation and reporting of econometric results.

The traditional approach to practical econometric analysis starts with a theoretical model of the population data generating process to be estimated using a sample of data. Thus we assume a particular structure of the data generating process based on theory developed in similar studies elsewhere. However, in developing countries, there is considerable variation in contexts – pastoral semi-nomadic communities of West and East African Savannah; coastal plantation dependant communities of the East African coast, mixed farming smallholder farmers of Central and Southern Africa, and various other farming systems. Context has a significant effect on the relationships between variables. Thus it may not be wise to apply theoretical models developed in one context to another without assessing the implications and making some modification. You need to think carefully about how the situation you are analyzing differs from that where the model was first applied. A component of your research might be to check the validity of standard models and produce suitable refinements for your context. Research in rural Africa needs to be approached with an open mind – involving re-examination of imported theoretical constructs in the light of the community you are working with. This necessarily involves also using the sample information to inform model specification even though this runs counter to the traditional approach to econometric modeling. In statistical terms, this means paying a lot of attention to the 'exploratory analysis' of data before formal modeling.

Modern approaches to econometric modeling provide for a role for both existing theory and data in specifying models. As Kumar (2007) argued 'the overall credibility of statistical evidence depends not only on the statistical evidence we provide *under the assumed model* but also on *the credibility of our assumed model*.' It is here that the description of the data generation process (i.e. theory in current context) and exploratory data analysis (i.e. sample data) play an important role.

Due to the paucity of secondary data at farm or community level, most agricultural and rural development research in Africa relies on primary data collected through surveys. Econometrics is one of the tools that are available for analyzing relationships between variables generated by surveys. Most regression analyses of survey type data involve either multiple linear regression or discrete dependent variable estimation. Multiple linear regression measures the relationship between a continuous dependent variable and one or more continuous and/or binary variables (qualitative variables). Examples of these include yield or area response models, determinants of income or wealth of households and determinants of sales, among others. Discrete dependent variable regressions model the probability of a discrete set of possible outcomes as a linear function of one or more continuous and/or binary variables. Examples of these include determinants of loan default and technology adoption studies.

Theory and Specification

Modern econometric modeling allows for the possible existence of a variety of competing theories constituting candidate data generating process or specification. The strategy is to identify all competing theories – variables affecting the key dependent variable as well as the nature of the effects – and to bring this into a hybrid model. This model will then be subjected to a number of diagnostic tests to weed out theories that do not conform to the target population. This strategy is described in Mukherjee, White and Wuyts(1998) as general-to-specific model building.

What does the above strategy suggest to us as a start to specifying our model? I suggest the following stages:

- A rigorous review of existing literature to find out which variables have proved in other, preferably similar, contexts that they affect the dependent variable under analysis
- A close look at qualitative information gathered during the survey exercise for context factors that may have an effect on the dependent variables. Most surveys are preceded by rapid rural survey exercises that look at the institutional, environmental and socioeconomic environment under which the target community operates (note: these could form the basis for new theories)
- A look at the nature of the effect of the variables identified above to account for the possibility of interaction effects. In a linear model representing the variation in a dependent variable Y as a linear function of several explanatory variables, interaction between two explanatory variables X and W can be represented by their product - that is, by the variable created by multiplying them together. The interaction effects to be included in the general specification however need to make theoretical or logical sense to avoid having excess variables of dubious use
- Take into account theoretical restrictions on the nature of the effects in selection of functional forms. For example, there is a difference in the effect on yield of preventive inputs such as pesticides compared to productive inputs such as fertilizers in production function estimations.

Care should be exercised in selecting the independent variables. A frequently encountered error is to specify identities. For example, specifying total household income as a function of income from livestock, income from cropping, off-farm income, and other variables. Also to avoid are models containing dominant variables such as using harvested area in addition to other factors (X) as explanatory variables for output. The risk is that it is likely the other variables (X) have an effect on area leading to problems of multicollinearity or the area variable

will overshadow the impact of the X variables in the model. This would be counterproductive since it is the effect of the X variables – the primary causal factors – we wish to investigate.

Data and Model Specification

Once the full array of theoretically plausible variables, interaction effects and functional forms are assembled, the next stage of the general-to-specific modeling begins. This involves a close look at the data to uncover hidden messages in the data that can inform model building.

Why do we need to look closely at variables? The answer lies partly in the definition of regression. Regression attempts to explain the total variation in the dependent variation by breaking it down into that explained by variation in independent variables and the residual variation. Thus we need to look at the nature of the variation in the dependent variable and compare this to variation in the independent variables. According to Mukherjee, *et al.*, '…in general, regressions between variables that differ significantly in shape do not perform well'. Ideally the shapes of the distributions should not depart much from the normal distribution. However, most variables in social sciences are skewed (Mukherjee *et al.*) making the normality assumption invalid. This is also of concern since most inferences depend on the normality assumption.

Another reason why looking closely at data is important is to help guess the nature of relationships between variables. This is normally done using correlations, box-plots and pair-wise scatter diagrams. To avoid the data mining accusation though, the interpretation of these apparent relations has to be done as part of a dialogue process between theory and data. That is, look at the structure of the variables (data); try to find plausible explanations given the context of the research (theory) and; adjust the model accordingly. Histogram plots may show that the univariate distribution is bi- or triple-modal. This tends to suggest that the sample includes more than one category of individuals that should be considered separately. For example, productivity may differ because of gender of household head or ownership of draft-power.

There are a number of tools that have been developed under the banner of exploratory data analysis that help us to interrogate the data from different angles. These tools are available in all good statistical analysis and econometric software systems. Note there are powerful free tools available. Costly old favorites like SPSS may not be the answer!

Univariate Tools

- Histograms: These are frequency distributions of the variable over the sample range. These would give an idea on deviation of the distribution from normality either as skewed or as poly-modal
- Mean, Variance, and other summaries: These are descriptive statistics that are provided by most statistics packages as part of summary statistics. Include summaries such as median and interquartile range which are less influenced by outliers
- Outlier Identification: Looking for odd observations which seem out of line with general patterns. They may represent errors, or they may represent good data which comes from a different population. Or maybe your expectations of what the data should look like need updating!

The usefulness of the above analyses is as follows:

- If distribution is poly-modal, investigate the possibility of influence of some categorical variables on the variable using the bivariate methods discussed below
- If data has some extreme outlier observation(s), try to find out why this might be so. If there is no plausible reason consider leaving out these from the analysis
- In the case of skewed distributions, there are a number of transformations that can be made to the data to remove skewness. For a variable(Y), transformation Y^3 reduces extreme negative skewness, while Y^2 reduces mild negative skewness. On the other hand, the transformation Log(Y) reduces mild positive skewness while -1/Y reduces extreme positive skewness.

Bivariate Analysis

Bivariate can be used to indicate the presence of outliers and non-linearities in pairwise relationships. These include:

- Scatter plots: the dependent against each independent variable (e.g., yield, fertilizer) as well as pair-wise among the independent variables. Dependent-independent scatter plots can reveal whether or not there is any apparent relationship as well as the pattern of relationships (i.e. linear or non linear) between the two. Pair-wise scatter plots between independent variables can identify the likelihood of possible collinearity between the regressors
- A Correlation Matrix between all variables in the model is the counterpart of the scatter over. Among the independent variables, correlation coefficients greater than 90% tend to indicate possibility of excess multicollinearity in a model which may call for the dropping one or more such variables from the model. Correlation coefficients on their own may, however, give misleading results due to the sensitivity of the measure to outliers
- Comparative box-plots allow one to look at summary statistics of different sub-samples of the variable. The summary statistics include the median, lower and upper quartiles, and the maximum and minimum values. Sub-samples could be for example sex of household head or household income category. Such comparative analyses would help identify the influence of categorical variables – income class or sex on the structure of the distribution. Thus it enriches the meaning of the summary statistics through a visual exposition.

Estimating and Testing Alternative Specifications

After developing a general model specification that encompasses all competing theories and conditioned by exploratory data analyses, the actual econometric estimations begin.

The first task is to ascertain that the general specification does not leave out any other relevant variables by looking at the residual variation. This involves examining the residuals from an estimation of the general model to ascertain whether there still remain any patterns to the residuals which imply misspecification. These include investigation of serial correlation, normality and heteroscedasticity.

The second task involves investigating alternative specifications of the model to find the most suitable form. This will involve balancing consideration of simplicity, conformity of the model to expected patters and power of the model to reproduce the data. Any of numerous books on regression modeling and economictric analysis will take you through the many approaches and tools available.

After arriving at the 'best' model one needs to ask how sensitive the results of the model estimates are to minor changes in specification. That is, how stable is the model? Evidence from a model that gives very varied estimates when one minor variable is added or removed from the model, is not reliable evidence and may cast doubt on the specification of the model. Very often this happens if functional form is not correct. Most often the model estimation focuses on the effects of select key variables on the dependent variable. We would like the model estimates of these key variables to remain as stable as possible with the inclusion or exclusion of fringe regressors.

Reporting Econometric Results

Finally, once you have thoroughly explored and tested your models and results, the following reporting and interpretation strategies would give a rich treatment of the econometric research findings.

Descriptive Analysis

Most dissertations using econometrics have a chapter summarizing key variables from the survey. However, these are usually just that – summaries. The exploratory data analysis suggested above can form the basis of a more meaningful interrogation of the data bringing

more useful information to aid in the formulation of the models as well as providing information to help explain the econometric estimates.

Regression Results

A standard way to report regression results is to present in tables all variables, coefficients, standard errors, P-value, indication of significance (eg. levels *= .01,** = .05,*** =.10), F-value and adjusted R^2 for the preferred model. What would enrich the reporting is to also include results of specifications based on the main competing specifications. This would communicate the differences in the current model and other models perhaps estimated elsewhere and would form the basis for a discussion on why these differences exist.

The results section needs to report the following:

- Overall assessment of model fit including the coefficient of determination, the adjusted R^2
- Next, the section needs to report on the significance (including non-significant variables and why they may be insignificant), direction and magnitude of each variable on its effect on the dependent variable. For example, 'regression analysis indicated that nitrogen (0.5) and weeding days (0.35) were significantly associated (95% level) with an increase in yield'
- Discuss the stability of the estimated model to changes in minor variables. Many reports tend to leave this out which is a weakness.

Implications of model results for key rural development objectives

What is the meaning of the results for different classes of household? Are there any differences in the effect of some key variables on the dependent variable between categories of households? Why? What do these differences imply for policy or programming?

One way of using the estimated results to answer the above questions is to simulate changes in the dependent variable caused by shifts in various combinations of dependent variables. An excellent example of such simulations is provided in a study by Thomas Walker *et al.* on determinants of poverty in Mozambique. Table 1 is an extract from their paper. After estimating a model of the determinants of poverty (measured as the square of the poverty gap), Walker *et al.* followed it up with simulations of the impact on poverty (using the percentage change in the poverty gap index) of concrete changes in key policy variables such as a given change in education level or the increase in intensity of production of agricultural commodities. That way they were able to make meaningful conclusions on how different types of agricultural development initiatives would contribute to the millennium development goals. That is, the study goes beyond just determining whether or not certain variable have an effect on poverty, but outlines the expected relative impacts on poverty of concrete changes. Notice that how suitable your model is for this sort of interpretation needs thinking about carefully. If 'independent' variables are actually related to each other, but that has not been built in to the model (as is the case in multiple regression models), then the predicted effects of changing one variable can be misleading.

Some Issues in Reporting Binary Choice Model Results

The use of qualitative dependent variable models in general and binary choice models in particular are currently on the increase in agricultural econometric studies. Of these, studies looking at determinants of technology adoption and credit default have constituted the majority of theses and published articles.[1] In the adoption context, a farmer either adopts (Y=1) or does not adopt (Y=0) a given technology, recognizing that what constitutes 'adoption' needs careful and context-specific definition. The decision to adopt or not to adopt (the dependent variable) is influenced by a number of factors including the characteristics of the farmer, the profitability of

1 A simple web search using 'agriculture' and 'logit or probit' would testify to this.

Table 1. Changes in the Severity of Rural Income Poverty by Scenario

Scenario No.	General Description	Specific Description	Change in the squared poverty gap index (in %)
1	Education	Shift upwards in one educational category, i.e., illiteracy to 1-2 years, 1-2 years to 3-4 years, and 3-4 years to 5 or more years	-7.0
2	Education	All household heads with some schooling attain highest educational level of 5 or more years	-9.3
3	Farm size	Households in the next to largest farm size category move to the largest category	-7.0
4	Farm size	Households in the smallest farm size category (0-0.75 ha) move to the next group (0.75-1.75 ha)	-2.5
5	Fields	Similar to Scenario 1, households move up to the next field number category	-6.5
6	Farm size + fields	Scenario 3 plus all households in farm size 4 (>5 ha) operate 5 or more fields	-13.8
7	Local crop	Similar to Scenario 1, increase the number of potential crops that are cultivated in the community by one category	-9.3
8	Intensification: coconuts	Households with 1-19 coconut trees move to the next level of 20 or more	-3.1
9	Intensification: cattle	Households with 1-9 head move to the next level of 10 or more	-1.1
10	Intensification: chickens	Households with 1-29 chickens move to the next level of 30 or more	-11.5
11	Intensification: tobacco	Tobacco cultivation reaches full adoption in the 8 districts where tobacco is most widely cultivated	-2.4
12	Demographic	Incidence of widow-headed households is halved; i.e., 50% of widow-headed households are changed to male-headed households	-1.0
13	Demographic	One more young child (ages 0-4) to households with one or more children in the 0-4 and 5-14 age groups	+16.7

the technology, accessibility of the technology, among other factors (the independent variables). The binary response models estimate the probability of a farmer having adopted a technology (Probability that Y=1) as a function the independent variables. Two functional forms – the logit and probit – are generally used to model this relationship. Interpretation of the results centre on the how the independent variables affect the predicted probability that Y=1 (in short P(Y=1)).

However, despite the rigorous theoretical exposition at the beginnings of such studies, the interpretation of the results is often rather limited. Students mainly emphasize the identification of factors that affect the probability of the dependent variable. For example, that education, experience, and wealth have a significant positive effect on technology adoption while sex of the farmer and ethnicity has no significant effect. A few students go further and

present marginal effects of independent variables on the probability of the dependent variable holding other variables at the their mean values and stop their analyses. However, as outlined in Long, Chap 3, a number of factors limit the marginal effects at the means:

- Due to non-linearity of the probit and logit models it is difficult to translate the marginal effect into discrete changes in the predicted probability
- Averaging dummy independent variables does not make intuitive sense as 0.5 does not represent any observed value
- For highly skewed independent variables the means do not represent characteristics of the 'average' sample member.

Scott Long suggests a number of reporting strategies that can potentially enrich the interpretation of binary choice models. These are briefly outlined below.

Setting the Basis Values for Independent Variables in Measuring Effects

Due to the highly non-linear nature of logit and probit models, the effects of any particular independent variable on predicted P(Y=1) will be dependent on the levels of other independent variables. The convention is to investigate the effects of each independent variable using the 'typical' or 'average' sample member as the base case, hence the wide use of arithmetic means as the basis. However, in the case of skewed distributions of independent variables the median may provide a more representative sample member.

More often than not, just focusing on effects based on typical households may not provide a more complete description of the effects. Using multiple base valuation points may give a richer range of effects. The idea is to define combinations of characteristics that typify particular subgroups of people then compute and compare effects. For example, one may want to investigate how the effects of years of education on predicted probability of adopting technology varies with level of wealth and gender of household head. In such a case the effects of years of education on predicted probability is evaluated under six alternative settings of independent variables described in the following table (Table 2).

Table 2: Scenarios for Assessing Gendered Impacts Under Different Wealth Classes

	Female Headed Household	Male Headed Household
Poor	Set gender dummy variable to female; wealth index to average of the poorest third of the sample and all other variables to their means(or median)	Set gender dummy variable to male; wealth index to average of the poorest third of the sample and all other variables to their means(or median)
Average	Set gender dummy variable to female; wealth index to average of the sample and all other variables to their means (or median)	Set gender dummy variable to male; wealth index to average of the sample and all other variables to their means(or median)
Rich	Set gender dummy variable to female; wealth index to average of the richest third of the sample and all other variables to their means(or median)	Set gender dummy variable to male; wealth index to average of the richest third of the sample and all other variables to their means(or median)

Using Plots of Effects over the Range of An Independent Variable

In any analysis there are some key variables that are central to the investigation and hence whose effects on the probability need to be better understood. However, due to the complexity of the response function some of the effects are not apparent from just looking at estimated parameters. In these cases graphing the effects of such variables under differing levels of other

variables would help illuminate the behaviour of the effects.

The hypothetical graphs below illustrate these. They show that the probability of adoption decreases with age for either gender and for all wealth levels. However, the graphics give us extra information on the effect of age on adoption of technology, specifically that:

- Female farmers at all ages are more likely to adopt compared to male farmers though the difference on probability of adopting decreases with age of the farmers; and
- The rate of decrease does not change with level of wealth of the farmer even though richer farmers are more likely to adopt at whatever age.

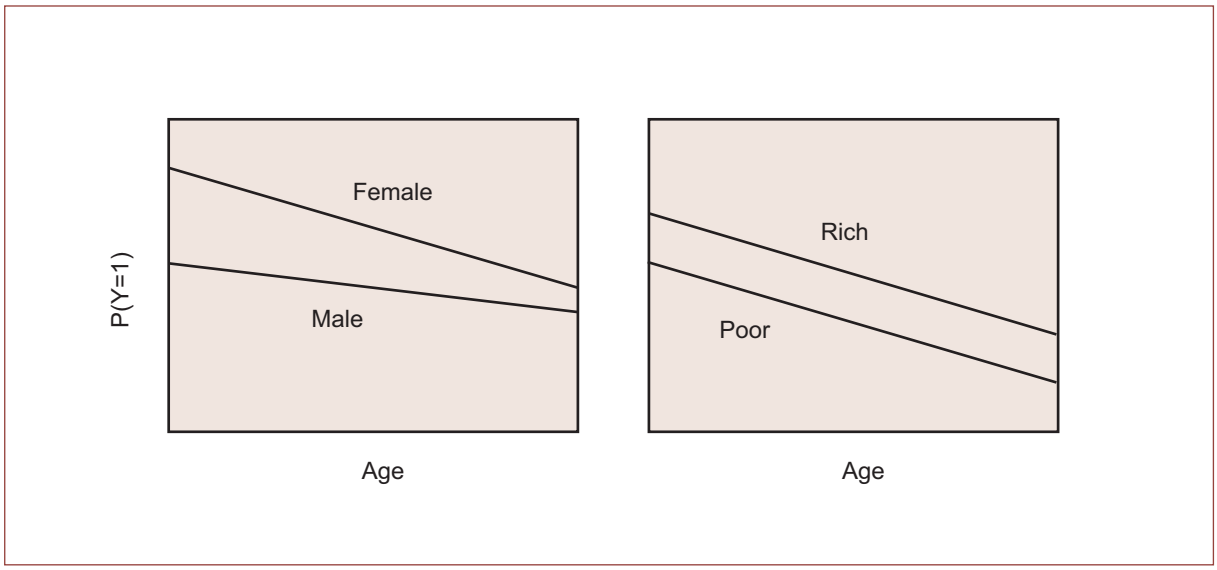


Figure 1. Effect of Gender and Wealth on Effect of Age on Adoption

Interpreting results

This does not mean you 'understand which effects are significant' but 'understand and communicate what you now know about the problem'. You should be able to:

- Meet the objectives of the study
- Clearly state what is the substantive new knowledge which as been generated
- Show how this new information and understanding builds on what was there before. Does it:
  - add more examples of something previously known?
  - mean that general rules or principles can be stated with more confidence?
  - allow predictions to be made for new and important situations?
  - mean that current understanding or theory has to be substantially modified?
- Use the quantitative information you have generated to make quantitative predictions about the larger picture
- The ultimate goal of the research is a development objective. Explain how your results help you towards that objective, and what the next steps will be
- Your survey and its analysis cost thousands of dollars. Explain why this was a good investment
- Answer the 'So what? question. What can we now do which we could not do before you did your survey?

References

Bennet, J. and Awimbo J. (Eds) 2002. 'Data Analysis for Agroforestry Experiments: Lecture Notes' World Agroforestry Center, 2002, Nairobi.

Coe, R. 2001. 'Statistics in Survey Analysis' and 'Steps in Survey Analysis', ICRAF, Nairobi, Kenya

Kennedy, P. 1992. 'A Guide to Econometrics', Massachusetts Institute of Technology Press, Cambridge, Massachusetts.

Kumar, T. K. 2007. 'Some Basic Issues in Statistical Modeling in Social Sciences'. *Economic and Political Weekly, India* 3027-3035 21 July 2007.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniques in the Social Sciences Series. Sage Publications, New Delhi.

Mukherjee, C., White, H. and Wuyts, M. 1998. *Econometrics and Data Analysis for Developing Countries*, Routledge, London

Statistical Services Centre. 2001. 'Modern Methods of Analysis'. University of Reading, Biometrics Advisory and Support Service to DFID.

Walker, T., Boughton, D., Tschirley, D., Pitoro, R., and Tomo., A. 2006. 'Using Rural Household Income Survey Data to Inform Poverty Analysis: An Example from Mozambique'. Paper presented at the International Association of Agricultural Economists Conference, Brisbane, Australia, August 2006.

4.4 Mathematical 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. Models are not restricted to science. A religious belief, philosophy or stereotypes are models. Indeed we use models unconsciously in all decision making from deciding when to change lane when driving to choosing careers. 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 representations 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, revise and retest it
- Question the data: perhaps it is not really relevant to the model you have chosen
- Abandon the line of research.

Note that we should not persist with a model for too long when the 'real world' evidence is that it should be rejected.

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.

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.

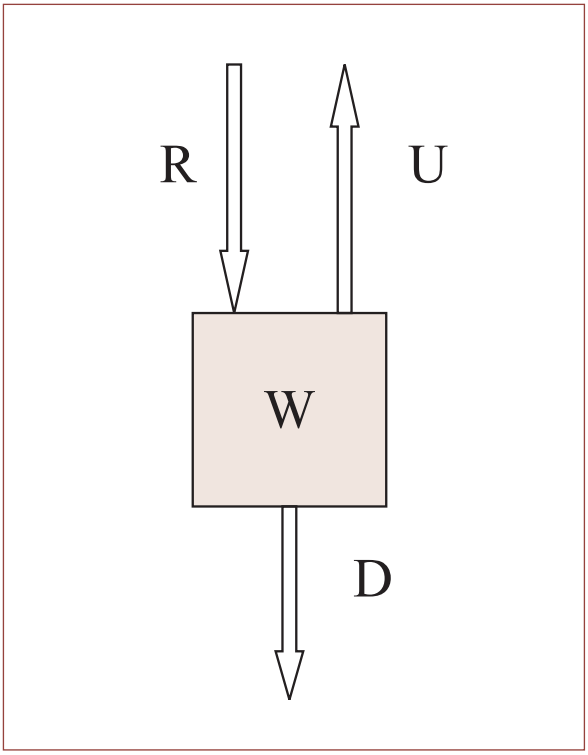


Figure 1. Simple model of soil moisture

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.

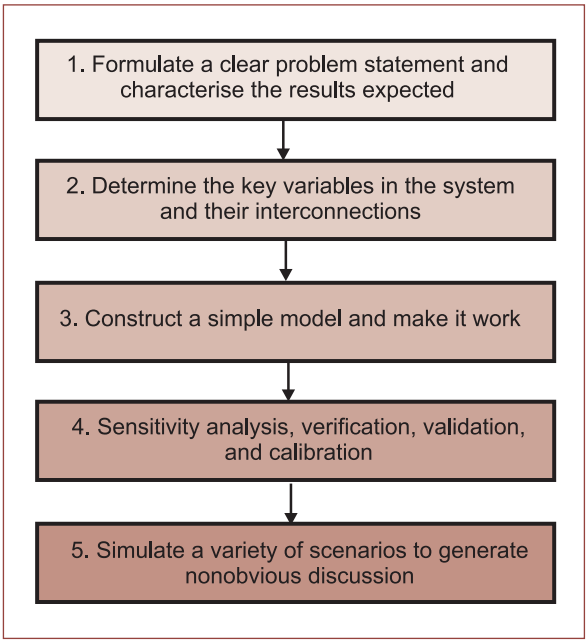


Figure 2. A summary of the steps involved in the modelling process

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'.

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.

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 systematically explore the response of the model to changes in one or

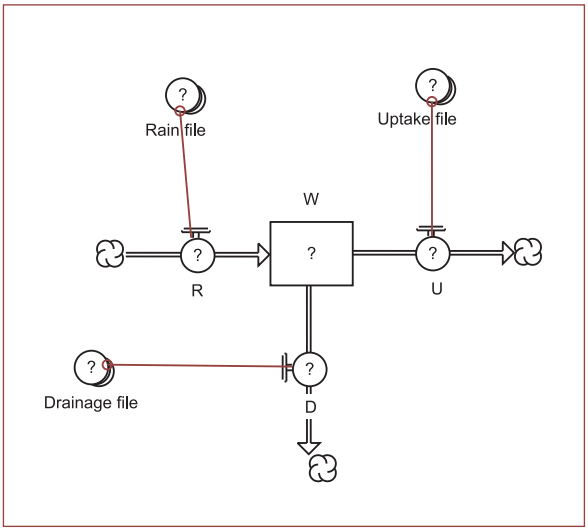


Figure 3a. Simple soil water model in STELLA

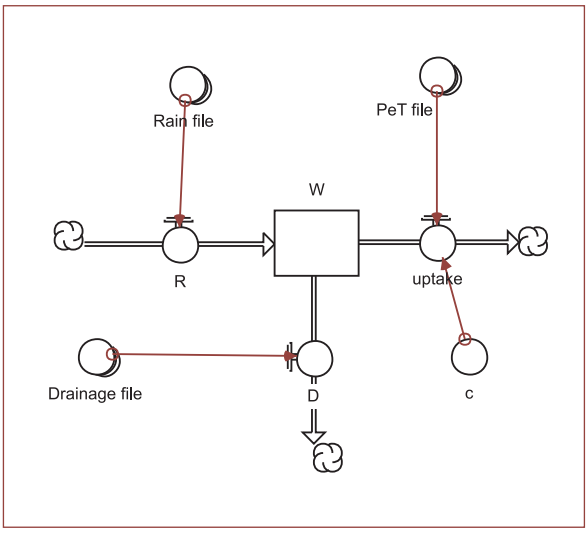


Figure 3b. Simple soil water model with uptake modelled as c.PeT

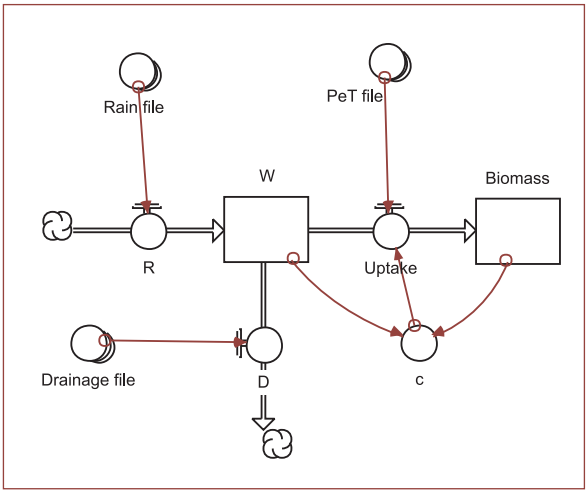


Figure 3c. Simple soil water model with uptake depending on both crop biomass and soil water

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* (semidecidous) 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 (2004).

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 1. The Craft of Research. Paul L. Woomer.

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

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 [accessed June 2009]

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. 2004. *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. and Ni'matul Khasanah. 2004. WaNuLCAS version 3.01: Background on a model of water, nutrient and light capture in agroforestry systems. International Centre for Research in Agroforestry (ICRAF), Bogor, Indonesia, 246 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* 1(5). http://www.iseesystems.com/community/connector/Zine/SeptOct03/jake.html [accessed June 2009]

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.

Vohnout, K. 2003 *Mathematical Modelling For System Analysis In Agricultural Research*. Elsevier, New York. 452pp

Internet resources

Ecological models http://ecobas.org/www-server/

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

FALLOW model at http://www.worldagroforestry.org/Sea/Products/AFModels/fallow/Fallowa.htm

FLORES model at http://www.cifor.cgiar.org/acm/methods/models.html An example of model building in participatory research

PARCHED-THIRST at http://www.ncl.ac.uk/environment/people/publication/13158/

WaNuLCAS model at http://www.worldagroforestry.org/SEA/Products/AFModels/wanulcas/

STELLA software; ISEE Systems at http://www.iseesystems.com

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

Vensim PLE. Vantana Systems Inc. http://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 http://www.modelkinetix.com/modelmaker/