Preparation for MSc Thesis Research at ITC and in Joint Education Programmes

D G Rossiter
International Institute for Geo-information Science & Earth Observation (ITC)
Enschede (NL)

19-March-2007

Contents

Purpose

For many ITC students the MSc research project is their first exposure to systematic scientific research. They need instruction in conceptual skills such as the scientific method, ethics and professionalism, and formulating research problems. They also need technical skills such as structgured technical writing, searching and interpreting scientific literature, proper use of citations, and abstracting. Finally, they need to know what is meant by a "good quality" ITC MSc project and thesis.

In recent years ITC and partners have developed several Joint Education Programmes (JEP) where ITC and the partner are jointly responsible for the taught and research components of the MSc. Each programme has a different structure, but they all require an MSc thesis of the standards, and following the concepts, presented here.

These notes were developed for a three-week module on preparation for MSc research at ITC. This module is intended to prepare students to write their research proposals. The notes contain more material than can be taught and practised in three weeks; thus they are also intended for reference during the thesis proposal and writing phases.

Finally, many parts of these notes are independent of the specific context of ITC MSc research. These should be useful to a wide variety of scientific workers.

Acknowledgements

The sections on statistical inference (§3) and scientific method (§1) have benefited from extended conversations with Hugh Gaugh Jr. and Jan de Leeuw; I have also profited from discussing scientific method and social ethics (§4.7) with Mike McCall, and social science research with Erik de Man. The sections on the research proposal and thesis structure (§2) have been informed by the ideas of Prof. Alfred de Gier and Jan de Leeuw. The subsection on argumentation (§1.6) is based on the original work of Ayako Kagawa. Thanks to Marcela Galindo (§2.2), André Nagelhout (§2.2), and Tomislav Hengl (§8.1) for their good-natured toleration of my adaptations of their MSc theses as examples.

1 The scientific method

Key chapter points

1. The **scientific method** is a manner of thinking and working towards more complete knowledge of the world.
2. To be scientific, a statement must, in principle, be **falsifiable**.
3. Sciences may be classified as **experimental**, **observational**, or **historical** (§1.1).
4. An important type of scientific reasoning is **deductive-inductive** (§1.2).
5. A scientific statement may be a **fact**, **hypothesis**, **theory**, or **law**, each with a level of **certainty** (§1.3).
6. To **argue** a point is to maintain its truth by reasoned debate, leading to a decision; this includes but is not limited to strict logic (§1.6).

The scientific method is not a belief system or religious dogma, but rather a manner of thinking and working towards more complete knowledge of the world. It has been **proven** to be **extremely successful** in:

- **explaining** the world as we observe it;
- **predicting** the previously-unobserved (the future, unvisited locations);
- **engineering**, i.e. building things that work.

A particularly important aspect of the scientific method is that it has a built-in mechanism to **check and revise itself**. That is, any statement in science is subject to revision or even *falsification* using the same methodology that was used to establish it in the first place. Thus it is *self-consistent* and does not allow for any *super-natural* reasoning.

Science is of course a human activity, and the pursuit of 'truth' is subject to all the human virtues and vices; fascinating discussions of how science really works may be found in books by, among others, Gower [6], Bauer [1], and Derry [3]. Gauch Jr. [5] gives an more philosophical view of the scientific method, while Okasha [10] is an especially accessible introduction to the philosophy of science underlying the scientific method.

1.1 Types of sciences

1. **Experimental**: controlled conditions under which measurements are made (e.g. laboratory experiments in physics or chemistry);

variable level of control of the context, but always quantifiable (e.g. temperature in a growth chamber can be controlled with a known precision)

2. **Observational**: uncontrolled or semi–controlled conditions
   - e.g. we can't order up an earthquake or extreme rainfall event
   - e.g. we can't manufacture survey respondents with certain characteristics[1].
   - Requires a sound sampling design.
3. **Historical**: we have evidence from the past, which can never be re-created experimentally (e.g. geology, archaeology)
   - can relate to current processes, assuming that the laws of physics etc. have not changed in the meantime
   - perhaps can reproduce some of the supposed processes in the lab.
   - relies heavily on inference and weight of evidence

Science vs. engineering:

- **Scientific research** is a method to discover facts about nature and to put these in a theoretical context ('why' the observed facts are so);
- **Engineering** is the design and manufacture of objects (which may be virtual, e.g. a computer program).

They both use *logical thinking*, and during the course of an engineering project many small experiments may be carried out to improve the design. The fundamental difference is that science *investigates* the world as it is and tries to explain it, whereas engineering *changes* the world by human activity.

1.2 The deductive-inductive scientific method

The best-known scientific method is known as the "deductive-inductive" approach. It has the following structure:

1. **Observe**;
2. **Invent a theory** to explain the observations $\Longrightarrow$ *induction*;
3. Use the theory to **make predictions** $\Longrightarrow$ *deduction*;
4. **Design experiments** to test these predictions;

[1] at least not with current technology ...

5. **Modify the theory** in the light of results $\Longrightarrow$ *induction*;
6. **Repeat** from step 3 **until** you can't think of any new predictions that might falsify or modify the theory.

As given above, this applies to the experimental sciences. For observational or historical sciences, Step 4 is modified:

4. **Make more observations** to test these predictions.

Step 4 is the crucial stage of **experimental design**: make new observations where they are **most likely to contradict what is expected** or where **an unexpected result would make maximum damage** to the theory. That is, the maximum information from a new experiment or observation comes either when the outcome is least predictable, or when it so predictable that an unusual result would be devastating.

Since we don't start from the beginning, the "Observe" and "Theory" steps are based on others' previous work and our general knowledge. This is nicely-shown in a famous diagram by Box *et al.* [2] (Fig. 1):

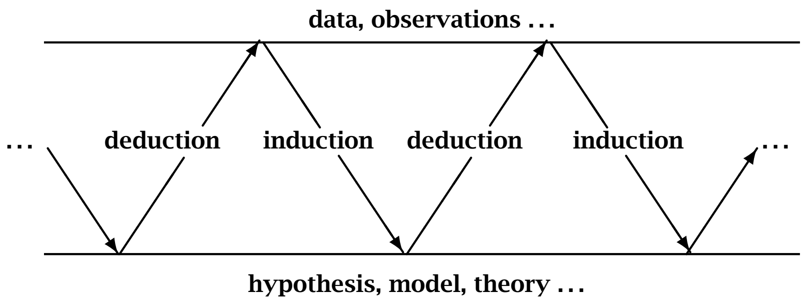

Figure 1: The deductive-inductive iterative approach to scientific knowledge (after Box *et al.* [2])

- The *deductive* step goes from existing theory or hypothesis to design a new experiment or set of observations, with expected results:

  "If my theory is true, and if I do this experiment (or make these observations), I should obtain these results."

- This is then compared to the actual results, leading to an *inductive* step where the existing theory is modified to account for the new results:

  "My experiment did not give all the expected results. (My observations are not all as I expected.) However, if I modify my theory this way, then the experiment (observations), as well as my previous knowledge, would fit this new theory".

This continues until the researcher is satisfied (and can satisfy others) that the theory is complete within its assumptions.

Induction and deduction are more fully explored below (§1.5)

A natural resources example

A soil mapping project[2] is a good example of a deductive-inductive approach with many iterations.

1. The initial *inductive* step: From background knowledge of soil forming process, the geological and tectonic environment, and surveys of presumably similar areas, the soil geographer forms a hypothesis of how soils have formed in this landscape, reasoning as follows:.

   > "There is evidence of recent (Pleistocene) glaciation in this region. The bedrock is known to be hard, massive limestone. The surface topography is an oriented field of rounded hills oriented in the presumed direction of the glaciation. My hypothesis is that the glacier flowed over existing small hills and formed them into a streamlined shape. As it did this, it removed the pre-glacial soil and scraped the bedrock almost bare. Since de-glaciation (about 12,000 years ago), soils have formed in the humid cool continental climate with typical northern hardwood vegetation."

   Furthermore, there is evidence from a somewhat similar region:

   > "In New England the same landform is observed on hard granite bedrock; these soils are indeed thin organic layers directly on the scraped rock."

2. The initial *deductive* step: From the hypothesis, the mapper can predict the soil types expected in each location; these are typically characterised by their landscape position. *Observations* are planned where they will best test this theory, for example, where the surveyor considers the most typical of each landscape position. In this case, the surveyor may reason as follows:

   > "If my theory is right, the shallowest soils should be on the steep side hillslope where the ice stream had the most

[2] based on the well-known drumlin field of the Lake Ontario plain between Syracuse and Rochester, NY (USA) [8]

> pressure; there should be hard limestone close to the surface and even out-cropping where erosion has been most active. There should be a dark topsoil with substantial organic matter, neutral to slightly alkaline, in equilibrium with the underlying limestone; there should be fragments of weathering limestone in the topsoil and increasing into the subsoil; total soil thickness over the hard limestone bedrock should not exceed 50 cm."

3. The soil is examined in the field and turns out to be a very thick, slightly acid, clay with no bedrock within tens of meters of the surface and no rock fragments. Obviously, something is wrong with the hypothesis!

4. The second *inductive* step: The theory must be re-formulated (a new *inductive* step). Some possibilities:

   (a) There never was a glacier here;

   (b) The glacial period was much longer ago, so soil formation had longer in which to operate;

   (c) There were no pre-existing hard rock hills for the glacier to mold.

   The first two have much regional evidence against them, so we hesitate to propose them, when there is a simpler alternative. Then we have to account for the oriented streamlined shape of the hills and their composition; i.e. we have to invent a new theory. One possibility is that the glacier encountered a flat lake plain with clayey soils, and molded these into a regular pattern of streamlined hills.

5. The second *deductive* step: The surveyor may now reason as follows:

   > "If the hills were formed by the glacier molding local clayey material, this same soil material should be found in the valleys between the small hills. Furthermore, the clay must have been formed in the inter-glacial period, so it should consist of medium- to low–activity clay minerals."

   Again, *observations* are planned where they will best test this theory, in this case in the valleys directly between two small hills. Furthermore, a laboratory determination must be made of the type of clay mineral in both landscape positions.

6. These observations show indeed the same type of clay in the hills and small valleys; furthermore the composition of the clay minerals is as predicted; the theory so far is *not contradicted* (is *supported*); note it is never fully *confirmed*.

This is only the beginning of the story; the soil geographer must build up a coherent theory of all the soils in the region and their inter-relation, in order to make a complete map. In addition, the reasoning (both induction and deduction) is much more complicated in reality.

1.3 Levels of certainty

We use the words "fact", "hypothesis", "theory" and "law" in common speech with a variety of meanings which often overlap. When discussing scientific certainty we must be more precise:

- **Fact**: something directly observable and measurable (but always with some uncertainty; no instrument is perfect);
- **Hypothesis**: a tentative theory, not yet tested; what we believe to be the true explanation or true state of nature, based on previous work or first principles; "[An] idea or a suggestion that is based on known facts and is used as a basis for reasoning or further investiation [7]; note the emphasis on "starting point", so that if a hypothesis is then supported by more evidence it becomes a ...
- **Theory**: a conceptual framework:
  - which explains existing facts;
  - allows predictions;
  - and is in principle *falsifiable* (some experiment or observation could contradict it or force its modification).

  "[A] reasoned supposition put forward to explain facts or events" [7]; note the emphasis on "reasoned", meaning that a theory must be supported by *evidence* and *logical argument* from this.
- **Law**: a theory with overwhelming evidence including the conditions under which it is true.

  For example, Newton's "laws" of motion are valid in cases where relativistic effects are not important (velocities low compared to the speed of light).

The boundaries in the sequence hypothesis ⇒ theory ⇒ law are of course fuzzy. A *law* can be defined as a *theory whose falsification, within its context, is almost inconceivable.*

The word "theory" seems to give the most trouble, since it only has to be a "reasoned supposition". So some theories are tentative, based on scanty evidence and easily-falsifiable, while others have much evidence behind them and are approaching "laws".

By the way...

> The following statement must be placed on the cover of secondary school science textbooks in some states of the USA:
>
> > "Evolution by natural selection is a theory, not a fact."
>
> This suggests some questions:
>
> 1. According to the definitions given above, is this a technically-correct statement?
> 2. What impression do you think it is intended to give to young students?
> 3. Given the evidence since Darwin and Wallace, where does the "theory" of evolution by (variation and) natural selection fit in the sequence *fact*, *law*, *theory*, *hypothesis*?
> 4. What would be a fair statement of the evidence so far?

1.4 Is a hypothesis necessary for science?

A **hypothesis** as defined above is a reasonable first explanation of the true state of nature based on previous work or first principles; the research must be designed to test or challenge this hypothesis. The research will either:

1. confirm;
2. contradict; or
3. cause a modification of ...

... the hypothesis.

Here's a simple example:

Research question Do students preferentially associate with others of their own nationality in academic activities at ITC?

Hypothesis No, ITC students mix freely.

Experiment Observe groups formed by free association (not by instructors) and compare their national composition to one that would be expected by random association.[3]

Revised hypothesis (Update the original from the experiment)

After several iterations and confirming experiments, this can be the starting point for a *theory* of behaviour.

[3] This would have to be expanded into a detailed experimental design.

Question: would you design the experiment differently if you had the hypothesis "Students from country (or region) X tend to stick together group exercises, but the others associate freely"? This illustrates the importance of the hypothesis for experimental design.

Some philosophies of science advocate *hypothesis-free* research, since just by stating a hypothesis we are constructing a context for the research and limiting its outcomes. This is often advocated in social sciences where researchers immerse themselves in communities with "no preconceptions" and "allow the theory to follow the observations".

This appears impossible in principle. No person can escape their life experiences, which form an implicit hypothesis (even theory) of how things (including societies) work. It is better to make these hypotheses explicit and then design the research to test them.

One could use the same argument for a natural resources survey. If the soils of a region have never been studied, how can the surveyor have a hypothesis of what soils are there, and how they are distributed on the landscape? Should observations be made without any theory? That is, should the sampling design be based on total ignorance?

This "total ignorance" attitude is:

- *inefficient*: because sampling can not be directed to extract the maximum information (i.e. to confirm or disprove the hypothesis); and
- *wasteful*: because it ignores previous work on soils and soil geography in other regions; the surveyor can reason from first principles of soil behaviour and from analogous regions elsewhere in the world, so is not truly in a state of ignorance.

So in fact a soil survey in an unmampped area must begin with a set of hypotheses based on previous knowledge (in this case, theories of soil formation and reasoning from similar areas). Then the survey can be designed to confirm or, more likely, modify that hypothesis.

1.5 Logical thinking

In science we use a combination of strict **deductive logic** and probabilistic **inductive reasoning**. How we actually think 'logically' in science is a fascinating topic [12]; here we give only a simplified view.

Induction and deduction:

- **Induction**: *generalise* from observations to theories
  - Logical process of inference

- this is how we make theories and laws

E.g. We have taken a transect of soil samples every 100 m across a dry lake bed, and these show higher salinity towards the centre of the lake bed. We generalise to all possible soil samples in the study area. Further, we develop an equation to predict salinity based on the distance from the edge of the lake.

- **Deduction**: *specialise* from a general law to a specific case
  - provides ideas for experiments or observations
  - "If this theory is true, then the following should occur or be observed"

Following the previous example, the equation predicts the salinity we should observe at any point in the study area. We may take a further step and apply this equation to all dry lake beds in the region (... to all dry lake beds in the world?)

Assumptions:

- Taken as true in the context of this research;
- Can not be tested within the time, budget or experimental design;
- If they are not true, the research is not valid;
- Often difficult to express, "taken for granted" at many levels;
- Established laws are often taken as assumptions, without explicit mention (e.g. we don't repeat the laws of universal gravitation each time we model landslide hazard);
- The more problematical should be made explicit;
- Could an *assumption* be a good *research question*? I.e. maybe the "assumption" should be tested!

Assumption vs. Hypothesis

Note the main difference between an *assumption* and a *hypothesis*: the latter is tested as part of the research, the former not.

Proof:

- In science very little is actually 'proven' in the strict sense of the word;
- Nature is very complex and subtle; simple answers are almost never satisfactory;
- Instead of 'proof' in the strict sense, *accumulate evidence*; additional evidence should support a good theory;

- However, it may cause the theory to be *modified*: either simplified or made more complex;
- 'house of cards': sometimes the theory falls under its own weight (too complex); leading to a **paradigm shift** (completely new way of conceptualising a set of observations), as famously explained by Kuhn [9];
- In practice, we are looking for proof within some **context**;
- Experiments are often designed to find the **limits of applicability** of a theory.

Statistical inference:

- This is discussed in more detail in §3;
- A formal way to accumulate evidence in support of a **model** which is a mathematical expression of the corresponding **hypothesis**;
- Can not by itself prove anything, must also have some **meta-statistical** argument about **causes** and, if possible, **mechanisms**.

Ockham's Razor:

- "*Pluralitas non est ponenda sine neccesitate*" ("Entities should not be multiplied unnecessarily");
- In practice: if two theories both explain the observed facts, then use the simplest;
- More evidence may require a more complex theory.

By the way…

> This is where conspiracy theorists and scientists have an un-bridgeable conceptual (and communication) gap: the conspiracy theorist is only happier as the theory gets more complex (e.g. if many hundreds of people would have had to be in on the assassination of JF Kennedy) whereas the scientist prefers the simpler explanation *unless* there is sound evidence for more complexity.

Parsimony:

- This is a technical term used in statistics to express the idea that the simplest relation that explains the data is the best [4].
- Motivation as for Ockham's Razor
- "Fit the relation, not the noise": maximum information

- Don't fit just one dataset
- Various measures of parsimony or information, adjustments to naïve measures of model success like regression R^2 (e.g. Akaike Information Criterion)

1.6 Argumentation

Argumentation may be defined [11] as "methodical reasoning; debate". So, to *argue* a point is to maintain its truth by reasoned debate.

> Note the transitive form of the verb: the intransitive form "to argue" means "to reason contentiously"; here we must have a point to argue (the object) and there is no sense of contention.

Argumentation it is not about "winning" the argument, rather about reasoning towards the best approximation to the truth. The usual aim to is take some sort of action based on the results of the argument, So, argumentation is a constructive debate to reach a solution [12].

How to argue a point: Argumentation is often based on logic, deductive and inductive. But, it can employ less rigorous methods, it does not have to be strictly logical, rather it can be based on *weight of evidence* and human intuitions of *likelihood*. The aim is to build sufficient evidence for the claim in order to make a decision.

Structure: An argument has a stereotypical structure:

1. a *claim* to be established;
2. *evidence* to *support* the claim;
3. a *warrant* or justification: the "since ..." which provides the link from data to claim;
4. a *backing* that provides the context (not to be argued); this is often difficult to state precisely, and includes the entire context of the argument.

For example:

1. Claim: "Private automobiles in Mexico City should only be allowed to drive on odd or even dates, according to the last digit of their license plate";
2. Evidence: "Air pollution from traffic is causing serious health problems";
3. Warrant: "Half of the cars means half of the pollution";
4. Backing: "People's driving habits will not otherwise change".

Now that this argument has been made explicit, we can look for flaws in the argument:

1. Is this reasoning (logic) correct as such?
2. Is the evidence correct? Is it complete?
3. Is the warrant a sufficient justification?
4. Is the backing true, and does it contain all the relevant information?

The critical examination of the argument to find weak points then leads to counter-arguments, which should be formulated as successive approximations to a final correct statement.

Argumentation styles:

1. From *definitions*, "define the problem away"; not too useful but may set up a more focused argument;

   > For example, the claim "Vegetarians are healthier than carnivores" can be defined away by narrowly defining what is meant by "healthier" to ensure that the available evidence supports the claim.

2. From *cause and effect*, but these may be difficult to separate;
3. From *contributions and impacts*, a weaker form of cause and effect;
4. By *analogy* or *comparison* with similar cases; must establish similar context (geographic, social, environmental ...) for the analogy to be valid; the argument must clearly state what is different in this case, and how it affects the argument;

Here is an example of argument by analogy:

1. Claim: "Community forestry (CF) should be introduced in [name your country]";
2. Evidence: Success of CF in Nepal;
3. Warrant: "What works there should work here";
4. Backing: "there are no relevant differences in society or environment between [here] and Nepal".

Putting it this way, it is clear that the backing is obviously false. However, this provides a means to sharpen the argument, by identifying the relevant differences and modifying the argument to account for them. Some of the differences in this case might be:

- Social structure
- Administrative structure (government as a whole, forest sector)

- Infrastructure
- Economic, educational level, other social indicators
- Religion, beliefs
- ...

What does each of these differences imply about the claim?

- The differences can be identified but then “argued away”, arguing that they don’t have any relevance to CF;
- The claim can be modified; rather than adopt CF in its Nepalese form, modify it for local conditions. Claim: “CF as practised in Nepal, but with [list the modifications here], should be introduced ...”

References

[1] Bauer, H. H. 1992. *Scientific literacy and the myth of the scientific method*. Urbana, IL: University of Illinois Press

[2] Box, G. E. P.; Hunter, W. G.; & Hunter, J. S. 1978. *Statistics for experimenters: an introduction to design, data analysis, and model building*. New York: Wiley

[3] Derry, G. N. 1999. *What science is and how it works*. Princeton, NJ: Princeton University Press

[4] Gauch Jr., H. G. 1993. *Prediction, parsimony & noise*. *American Scientist* **81**:468–478

[5] Gauch Jr., H. G. 2002. *Scientific method in practice*. Cambridge: Cambridge University Press

[6] Gower, B. 1996. *Scientific method : an historical and philosophical introduction*. London ; New York: Routledge

[7] Hornby, A. S. (ed.). 1995. *Oxford advanced learner’s dictionary*. Oxford: Oxford University Press. ITC 038.20.951 reference

[8] Hutton, F. Z. 1971. *Soil survey of Cayuga County, New York*. Washington, DC: US Government Printing Office

[9] Kuhn, T. S. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 2nd edition

[10] Okasha, S. 2002. *Philosophy of science: a very short introduction*. Oxford: Oxford University Press

[11] Sykes, J. (ed.). 1983. *The concise Oxford dictionary of current English*. Oxford, England: Clarendon Press, 7th edition

[12] Toulmin, S. E. 1958. *The uses of argument.* Cambridge: University Press

2 The MSc thesis project

Key chapter points

1. **Research** is discovering something new about the natural world, the built world, or society, including research methods.
2. **Types of research** include (1) designed experiments; (2) systematic observations; (3) review and synthesis; (4) system design; (5) social sciences (§2.1).
3. The **research proposal** establishes the relevance, novelty, methodological soundness, and feasibility of the project; it **convinces** the reviewer that the research should be undertaken (§2.5).
4. The research proposal often has a conventional structure: **Problem** ⇒ **Objectives** ⇒ **Questions** ⇒ **Hypotheses** ⇒ **Methods** (§2.2).
5. The **research problem** is a general statement of what is not known and should be discovered by research (§2.2.1); the **research objectives** are statements of what is expected as the results of the research (§2.2.2); the **research questions** are specific questions that the research can answer (§2.2.3); the **hypotheses** are expected answers to each question (§2.2.4).
6. **Research methods** are chosen in order to answer the research questions (§2.2.5).
7. A **design thesis** must present a new design that is demonstrably "better" in some sense than existing systems (§2.3).
8. A **social thesis** emphasizes concepts and definitions and argues from diverse evidence (§2.4).
9. Supervisors may give advice and comments, but the MSc thesis is the **student's responsibility** (§2.6).
10. Effective **time management** is necessary to produce a good-quality thesis in the alloted time (§2.7).

Research for an MSc thesis is one small part of the scientific enterprise. It must be grounded in the scientific method (§1) and follow a sound research plan. In this chapter we discuss the steps of the research process, emphasising how to prepare a good **research proposal**. With suitable modifications, these steps can be followed for writing more extensive proposals.

2.1 Types of research projects

The term "research" is from the French *réchercer*, "to find out". This general term can be used for:

1. **Designed experiments**, e.g. laboratory or field research where the researcher imposes the treatments in a (semi-)controlled situation;
2. **Systematic observations**, e.g. resource survey or community meetings, where the researcher makes measurements or observations according to a plan but without complete control of the process;
3. **Synthesis**, where the researcher imposes a new conceptual framework on previous data and establishes that this is a better or more unifying explanation;
4. **System design**, where the researcher designs a system and shows that it is "better" in some sense than previous designs; this includes design of algorithms and methods.

Natural vs. social sciences

There is in general a distinction between the **natural** and the **social** sciences:

Natural: The principal object of study is "nature", i.e. physical reality; there is a clear separation between observer and observed; argumentation is as logical and objective as possible;

Social: The principal object of study are humans and human society; so we can can not impose treatments at will; we are studying ourselves, so it is very difficult to avoid subjectivity; argumentation grades into humanities; see also §2.4.

In all of these, the main object of study can be the **thing in itself** (e.g. natural world, the built world, society), or the **methodology**[4] used in the study. For example:

Thing in itself: Changes in land use in a study area; commerce patterns in a district;

Methodology: How to assess land-use changes with multiple satellite sensors of different resolution; how to visualise spatio-temporal commerce patterns.

Often an ITC thesis includes both aspects; we are interested in the thing in itself (ITC development relevance) but also the methodology (ITC technology focus).

[4] "study of methods"

2.2 The “research” thesis

This is applicable to both designed experiments and systematic observations. It has a conventional scientific structure following the deductive-inductive approach (§1.2).

The research proposal has a conventional structure:

Problem ⇒ Objectives ⇒ Questions ⇒ Hypotheses ⇒ Methods

Following these concepts in order is a systematic way to approach research. It must first fit a known *problem* (so that it is important), then it must have a defined *objective* (so that it is clear what it should accomplish), which is then specified as a list of *questions* that the research should answer. For each question, the researcher must have a *hypothesis*, i.e. what answer is expected.

We will illustrate these concepts with examples modified from two ITC MSc theses, one in Naivasha, Kenya [6] and one in Lake Cuitzeo, Mexico [3]. The Naivasha example deals with the applicability of Small-format Aerial Photography (SFAP) to monitor wind erosion; we will also examine a different thesis topic (not actually carried out) that deals with the same area and general problem, but more conceptually.

2.2.1 Research problem

- A general statement of *why* the research should be done;
- Something that is not well-understood or solved and can be addressed by research;
- Not a social ‘problem’ (poverty, environmental destruction, war, ...), but social problems can motivate research (*relevance*).

The novelty of the research must be supported by a **literature review**. If someone else has already solved the problem, why re-do the work? This is explored further in §5.

The fundamental questions:

1. **Why should anyone care about the outcome of this research?**
2. **Who would use the results of this research?** and **for what?**
3. **Why should anyone sponsor this research?**

A reasonable answer to the first question might be “because it’s intrinsically interesting to know ...”. This is not a strictly utilitarian approach; however for most ITC research there should be a more concrete reason to undertake it.

(And of course, one important outcome of MSc research is that you will receive an important professional qualification (the MSc degree) and further career . . . so *you* care about the outcome of the research . . . but that's not sufficient for the sponsor or supervisor!)

Example (Naivasha SFAP): This is an example of a technology-oriented thesis, where new methods must be developed:

> "Wind erosion is causing widespread destruction of crop land and pastures in the rift valley of Kenya." (a *social* problem, context)

> "We do not know the priority areas for intervention." (a *management* problem)

> "It is impractical to monitor wind erosion over large areas by ground survey or conventional aerial photography." (a *technical* problem).

This sequence of problems leads naturally to an *objective*, namely to find a cost-effective way to monitor wind erosion over large areas and from these surveys to determine priority areas for intervention.

Note that there are many problems implicit in this example, and they could lead to useful research problems:

- How to monitor wind erosion over large areas in a cost-effective manner? (the problem actually chosen for this research);
- What are the priority areas for intervention?
- What land use practices are most associated with wind erosion? (Note: this is not yet a statement of *casuation*);
- What are the physical and social causes of wind erosion in this area?;
- How can these causes be addressed to minimize erosion?

These questions are inter-related: we must be able to monitor before we can determine the priority areas; and the monitoring is the basis for associating land-use practices with erosion. This association is then used as evidence when arguing about the physical and social causes; and once we know these causes we can design interventions.

The question for the MSc student is: **What can I realistically accomplish in the thesis research?** and what part do I leave for others? This is partly determined by the status of the problem. For example, if no good monitoring method is available, we should work on this, because the other problems depend on it.

Example (Naivasha causes): This is an example of a cause-oriented thesis, where the emphasis is on determining *why* something is occurring.

> “Wind erosion is causing widespread destruction of crop land and pastures in the rift valley of Kenya.” (a *social* problem, context)

> “Previous studies have established the extent of the problem and its historical development.” (not a problem, but rather an *opportunity*, a basis for deeper investigation)

> “We do not know the proximate or ultimate causes.” (a *knowledge* problem)

So the emphasis here is on determining *why* some known phenomenon is occurring.

Example (Lake Cuitzeo): This is an example of a survey-oriented thesis, where the principal problem that there is no map of something of interest.

> “The water of Lake Cuitzeo is used for multiple purposes, including irrigation and human consumption.” (context)

> “Almost nothing is known about its quality, but large areas are suspected to be sub-standard for both purposes.” (a *knowledge* problem)

> “There is no map of the different water quality parameters in the lake.” (also a *knowledge* problem)

> “It is not known whether there is any trend in the quality.” (also a *knowledge* problem)

> “Nothing is known about the causes of poor water quality, although it is suspected that high-input irrigated farming is a major contributor.”

In the limited time available for an MSc thesis, only some of this lack of knowledge can be addressed. In particular the time series necessary to determine a trend can not be collected in one field visit. The final problem depends on the knowledge that is lacking as expressed by the second and third problems.

2.2.2 Research objectives

These are statements of what is expected as the *output* of the research. Each of the objectives must be at least partially met at the end of the project.

There is usually a single *general* objective which is not operational, which is then broken down into a list of *specific* objectives which can be addressed by operational research methods.

Example of a general objective (Naivasha SFAP):

- “To determine the applicability of Small-format Aerial Photography (SFAP) to wind erosion mapping and monitoring in the rift valley of Kenya, and the main factors which affect its success.”

Example of a general objective (Naivasha causes):

- “To determine the causes of wind erosion in the rift valley of Kenya.”

Example of a general objective (Lake Cuitzeo):

- “To map the water quality of Lake Cuitzeo on one sampling date and suggest possible causes for any spatial variation in water quality.”

The *specific objectives* should be built up from simple (easy to formulate and investigate) to complex. If there is an inventory to be done, the objective is simply to do it; this may be followed by objectives that require more inference.

The thesis should at least partially meet all the objectives.

Example of specific objectives (Naivasha SFAP):

- “To determine which wind erosion features, and of what dimensions, can be visually interpreted on SFAP”
- “To determine the accuracy with which SFAP can be georeferenced with single-receiver GPS and mosaicked into a seamless image”
- “To determine the costs of a SFAP mission in local conditions”

Example of specific objectives (Naivasha causes)

In this case, the word “causes” is very broad, and it is customary to distinguish between *proximate* (immediate) and *ultimate* causes, also between *factors* and *processes*.

- “To determine factors related to wind erosion in the study area”
    - “To determine which land use practices are most associated with wind erosion”
    - “To determine which soil properties are most associated with wind erosion”
    - “To determine which soil properties are most associated with wind erosion”
- “To relate these factors with presumed processes”
- “To identify and quantify the proximate and ultimate causes of wind erosion in the study area”

Note that questions about *factors* may be answered by investigating the *association* (roughly speaking, "correlation") between them and the erosion; this then is information to be analyzed in terms of the *processes* by which wind erosion occurs, to finally discuss *causes*.

Example of a specific objective (Lake Cuitzeo):

- "To determine the water quality status of the central and Eastern parts of the lake"
- "To map the spatial distribution of the water quality components measured at one sampling time."
- "To map the spatial distribution of aquatic vegetation density."
- "To determine if there is a relationship between the reflectance values of optical multi-spectral sensors and measured water quality parameters including vegetation density".
- "To determine whether land use affects water quality, and if so, which constituents are affected by which land uses."

It's clearly easier to simply sample and map the water quality (first two objectives) than to determine why the water is of higher quality in some areas than in others (last objective). The third and fourth objectives (vegetation density and whether multi-spectral sensors can detect this) are technology objectives in support of the other research objectives. They could be the entire thesis if a difficult enough problem; this would be a different research focus.

2.2.3 Research questions

These specify what the research will actually address. *Each research question must be answered* by the thesis, therefore it must be a specific question to which an answer can be given. Questions follow objectives and may be simple re-statements in *operational form*, i.e. where an experiment or sample can answer it.

Questions are of two main types:

observational 'What' or 'where' questions;

analytical 'Why' questions.

The research project typically has a set of observational questions whose answers help in turn answer a set of analytical questions.

"Question" words

Here are some words that can be used to introduce research questions; first for those that do not require much analysis:

- "Where?" (mapping), e.g. "Where (in the study area) is the most severe accelerated erosion"

- “Is there” or “Does” (presence, existence), e.g. “Is there a water quality gradient with depth?”; this could be re-formulated “Does water quality vary with depth?”
- “Can?” (technique), in the sense of “Is it possible?”, e.g. “Can a light aircraft with GPS carry out a photo mission to specified accuracy standards?”; “Is it possible to see blow-outs on an air photo?”
- “What?” (results of a technique), e.g. “What is the accuracy of georeferencing?”
- “What?” (is encountered in the field), e.g. “What are the most common species of trees planted in domestic gardens?”
- “How?” (observational), e.g. “How has water quality changed since the establishment of the irrigation project?”; this could be re-formulated “What, if any, are the change in water quality . . . ”.

Another type of question requires deeper analysis:

- “What is?” (effects), e.g. “What is the effect of increased grazing on vegetation density?”
- “What is?” (relation), e.g. “What is the relation between increased grazing and vegetation density?”; this must be answered with a *statistical model*.
- “Why?” (causes), e.g. “Why does increased grazing affect vegetation density?”; this must be answered with some proposed *mechanism*.
- “How?” (function), e.g. “How does increasing pesticide use in surrounding farmland affect reproductive success of migratory bird species in the lake?”

Example of research questions (Naivasha SFAP):

> “What are the photointerpretation elements for different wind erosion features?” (e.g. in this case the blowouts may be darker because of the different ash in subsoil; elongated form in wind direction etc.)

> “Can blow-outs and dunes caused by wind erosion be seen on SFAP, and if so, of what dimensions?”

> “What is the smallest wind erosion feature than can be recognised, measuring both vertically and horizontally?”

> “Can sufficient group control points be established to convert the set of SFAP photos to orthophoto mosaic?”

- “What is the accuracy of such a conversion, using a single GPS receiver for ground control?”
- “What is the cost of a SFAP mission and how does this compare with conventional survey?”
- “What is the time required to organise a SFAP mission and produce an wind erosion assessment, and how does this compare with conventional survey?”

Note that although the general objective speaks of “monitoring”, there is no research queswtion directly related to this, because the research is only done in one time period. However, several questions relate to monitoring: what can be detected, and how much a mission costs. So in the conclusion the author can use the answers to these questions to discuss the applicability of the method for monitoring.

Example of research questions (Naivasha causes):

- “What are the land use practices in the study area?”
- “Which of these are most associated with wind erosion features?”
- “What is the quantitative relation between the intensity of specific land uses and wind erosion?”
- “What is the physical process which relates the intensity of a specific land use to wind erosion?”
- “What are the synergistic or antagonistic effects of specific land uses and other caustive factors?”
- “What is the principal cause of wind erosion in the study area?”

The above list only mentions land use intensity; other causitive factors should be added. Note that the questions go from easiest to answer to hardest. The last question can not really be answered as such; instead we can *argue* from the results of the previous questions to a more-or-less convincing story about causes.

Example of research questions (Lake Cuitzeo):

- “What is the water quality (turbidity, salinity) and depth at representative sample points in the Central and Eastern parts of Lake Cuitzeo?” (sampling)
- “What is the spatial structure of the lake depth as modelled by (i) geographic trend surface, (ii) distance from shore, and (iii) ordinary variograms?
- “What is the spatial structure of the water quality parameters as modelled by (i) geographic trend surface; (ii) distance from shore;

(iii) depth; (iv) ordinary variograms; (v) residual variograms from the trend and feature space models?" (modelling)

> "How much of the spatial structure can be explained by these models and how much remains unexplained?" (success of modelling)

> "What is the spatial distribution of water quality parameters and depth?" (mapping, using the models)

> "What is the relationship between reflectance values of optical multi-spectral sensors and water quality parameters including vegetation density?" (modelling, depends on the previous map)

> "What is the spatial distribution of water quality parameters including vegetation density as mapped from optical multi-spectral sensors?" (mapping, using the models)

> "What land uses are associated with areas of poorer water quality?"

Note how the water quality parameters are now specified. The analytic methods (trend surfaces, variograms) are also specified. Some questions depend on the results of others. For example, if there is no relation between aquatic vegetation and MSS, it is impossible to make a map.

This is a long list of questions and may be too much for a single study. Not all questions may be answered to the same depth.

2.2.4 Hypotheses

Hypothesis: "[An] idea or suggestion that is based on known facts and is used as a basis for reasoning or further investiation" [4]

In the context of research, these are the researcher's ideas on what the research will show, before it is carried out. They are statements that can be *proved* or *dis-proved* by the research. They are *based on previous work*, usually discovered in the *literature review*. They should match the research questions one-to-one.

The hypothesis must be *specific*, not a general statement. For example, given the research question "What is the effect of grazing intensity on vegetation density?" we can formulate the corresponding hypotheses:

- **Wrong**: "Grazing affects vegetation density"
- **Right**: "Above a threshold (to be determined), vegetation density is reduced linearly (coefficient to be determined) with grazing intensity, measured as animal-months.

The first hypothesis is too general, "affects" could be anything.

Example of hypotheses (Naivasha SFAP): The following statements refer to SFAP at a nominal photo scale of 1:5 000:

- “Blow-outs and dunes caused by wind erosion can consistently be seen on SFAP”
- “Both blow-outs and dunes with a vertical relief difference of as little as 1 m, and an minimum horizontal dimension of 5 m can be seen.”
- “It is always possible to find sufficient points for direct linear transformation within a single SFAP.”
- “SFAP can be converted to an orthophoto mosaic with a horizontal accuracy of 5 m using GPS ground control.”
- “The cost of a SFAP mission is an order of magnitude less than a conventional air photo mission.”
- “The time required to organise a SFAP mission and produce an wind erosion assessment is less than two weeks.”

Example of hypotheses (Naivasha causes):

- “The principal land uses are small-scale subsistence farming, paddock grazing of cattle, and extensive grazing.”
- “Wind erosion is found only in paddock grazing.”
- “No erosion is observed until grazing intensity reaches a threshold, after which the extent increases exponentially with grazing intensity until the whole area is destroyed.”
- “Overgrazing leads to removal of the surface cover (grasses), exposing the soil to the full kinetic energy of the wind.”
- “Fine-grained volcanic ash soils are more susceptible to wind erosion, when exposed by overgrazing, than coarse-textured ash and lacustrine soils.”

These hypotheses came from previous land use and soil studies in the study area, wind erosion studies in similar areas, and general physical principles. They look like conclusions but they are not! They are hypotheses to be verified, modified, or refuted.

Note especially the third hypothesis, giving the *form* of the presumed quantitative relation.

Example of hypotheses (Lake Cuitzeo):

In this case there is very little known about the study area, so the hypotheses are not very specific. An examination of an optical image gives some clues.

- “The central part of the lake is shallower and more turbid than the eastern part of the lake; salinity is absent to moderate.”

- “There is no geographic trend to depth; depth increases quadratically (bowl-like) with distance from shore; there is strong spatial dependence at ranges to 1 km.”
- “There is an east-to-west geographic trend in salinity; turbidity increases quadratically (bowl-like) with distance from shore; . . . ”
- “Models explain about 80% of the spatial variability.” (this based on studies in “similar” areas)
- (No hypothesis, output is the map)
- “Turbidity is linearly related to blue reflectance/”
- (No hypothesis, output is the map)
- “Areas of the lake receiving discharge water from high-input irrigated agriculture have the poorest water quality”.

Hypotheses are discussed more conceptually in §1.4.

2.2.5 Research methods

These are chosen in order to answer the research questions. This is why specific questions are so important.

For example, to answer the question “Can blow-outs and dunes caused by wind erosion be seen on SFAP, and if so, of what dimensions?”, we must:

1. Make a legend of wind erosion features and their characteristics to be measured in the field;
2. Identify test features in the field and geo-reference them;
3. Produce the SFAP;
4. Geo-reference the SFAP;
5. Interpret the SFAP at the locations of test features according to the legend;
6. Compare the interpreted features with the known features;
7. Quantify the degree of agreement.

All of these require definite methods. In this case we also have to protect against photo-interpreter bias: knowing the features in the field will the interpreter imagine them on the image? Perhaps the photo-interpretation should be before the field visit? Or should a block be photo-interpreted, not just specific features? It requires careful thought to make the methods able to answer the questions.

For each research method selected, the Methods section of the thesis should state:

1. Either:
   (a) the name of the method that was chosen, with a reference to the literature that describes it; *or*
   (b) a detailed description of the method, if it is being developing as part of this project;
2. Why this method was chosen:
   (a) Why is it applicable in this study?
   (b) Why is it preferred to other methods that could have been applied?
      - For example: cheaper, faster, more precise, adapted to the specific environment ...
3. What are the assumptions for applying this method, and how are they met in this study?
   - For example, a 1-dimensional water flow model (vertical flux only) assumes that there are no lateral fluxes (in the other two dimensions); this assumption is met in horizontally-homogeneous soils on level landscapes, so if such a model is applied the modeller must prove that these conditions are met.

2.3 The "design" thesis

Another type of research is a **design**, for example of a computer program, a user interface, a database structure, or an algorithm. Here the key question is why a design should be considered MSc research and not just a project. This is essentially the difference between engineering "research" and "development".

A "research"-level design must have:

- A high level of *creativity*; in particular it must create something really new, or at least a new synthesis;
- It must result in a design that is demonstrably *better* in some sense than the alternatives;
- The thesis must both *define* and *demonstrate* this superiority.

The *hypothesis* of the "research" thesis is then replaced with evaluation criteria: in what sense is the new design better than previous designs? This has several aspects:

1. What defines "better"?

2. How can this superiority be established?

In a "design" thesis, superiority is often established by a demonstration that certain *design criteria* have been met, which were not met in other products.

Similarly, in the *Results* section of the thesis, the discussion of how well the results support the research hypothesis is replaced by an argument that the design is "better".

Example

An example of a design thesis project is a proposal for a new structure of a soil geographic database. Here "better" could be defined as "allows the representation of real-world objects that can not be represented in any existing design" or "supports a class of queries that can not be carried out in any existing design". The thesis would have to:

1. Establish that there is a demand for a design;
2. Review existing designs and identify their shortcomings;
3. Show the proposed design;
4. Show how it is used on some sample data, i.e. a *proof-of-concept*, the design really can represent what was promised;
5. Show that it can represent concepts that are impossible with existing designs.
6. Show that this improved design is useful for answering a richer class of questions; for example, the database user can easily extract parameters for a defined class of models.

The "demand" for a design replaces the "research problem" of the research thesis. The question is still why anyone should bother to undertake this work.

Another example of a "design" thesis project is the design and implementation an improved user interface for statistical modelling. Here, "improved" would probably be tested by a series of designed experiments with target users, with measurable outcomes, e.g. how quickly or correctly users could accomplish a particular task.

Example We illustrate these concepts with a PhD thesis [9] which included the design and implementation of the *ALES* microcomputer program to assist in land evaluation [10, 11].

This thesis has sections on:

1. System demand;
2. System objectives;

3. System requirements;
4. System design;
5. System implementation;
6. System application to the problem field.

These are argued as follows:

Demand — Here the **demand for a system** is established. In this example we find sentences such as "There is today a high demand worldwide for information on the suitability of land for a wide range of land uses" (*social* problem) and "There are no comprehensive computer programs that allow the land evaluator to organise knowledge from diverse sources . . . " (*demand* in the narrow sense; identifies potential users).

Objectives — Here the **general and specific objectives of the system** (i.e. why it is being built) are presented. In this example the stated objective is ". . . to allow land evaluators to collate, systematise, and interpret this diverse information using the basic principles of the FAO's *Framework for Land Evaluation* [2], and to present the interpreted information in a form that is directly useful to land use planners"; this is then expanded to discuss the target group and the type of models that should be automated.

Requirements — Here the specifics of **what the system must be able to do** are listed. In this example, the system should allow expert judgement on the interactions between land characteristics to be captured in the system.

Design — Here the **new or unusual aspects of the system design** are explained. In this example, it is proposed to represent the interactions by decision trees; the question is then how to represent these in the system.

Implementation — This section contains an **explanation of how the system is implemented**. By itself it is only a project, not research; however as part of the wider discussion, and if justified, it becomes part of research. In this example, the selected computer language and database system are explained, and the program control flow is presented.

Application — This section **demonstrates that the system can meet the stated requirements**. In this example, several models were built in the system, some as duplication of existing manual methods [5, 12, 13] to show that the system could replace these, and some that were beyond the capabilities of existing methods. The proof of the system was that these models could be built, they could be applied to data, and results (in this case land evaluation tables) could be produced.

2.4 The "social" thesis

Another class of thesis project is a social analysis, i.e. the study of humans and human societies. These are complicated and even contradictory objects of study, and it is notoriously difficult to come to firm conclusions. Also, ethical and practical considerations make designed experiments either impossible or inadvisable. Still, it seems unavoidable that this species must sometimes be studied.

Here the "hypothesis" takes the same form as a research thesis, but the research *method* is different; in particular the **evidence can be subjective and anecdotal**, rather than the objective result of a measurement. The *Results* section of the thesis then takes the form of a reasoned *argument* from *evidence* as interpreted by the researcher.

A "social" thesis usually needs a section on **Definitions** or **Concepts**, where terms such as "participatory", "sustainable", "equitable" etc. are well-defined, so that they can be consistently identified in the research. A good example is the paper by Roling [8] on concepts of sustainability.

A typical ITC example is the hypothesis that "participatory" land use planning is "more successful than" top-down or technocratic approaches. Such a thesis also must clearly address the concept of "better" (as in the "design" thesis): what defines "better", and how can this be established?

The "social" thesis project may include some structured interviews or meetings, but these are much less controlled because of the unfortunate tendencies of human beings (both researchers and subjects) to distortion, fabrication, imagination, wishful thinking, etc. Social scientists have developed a range of techniques for increased objectivity, which should be used if possible (e.g. questions that ask for what should be the same information in different ways).

2.5 The thesis proposal

After you have mastered research skills, you will have several weeks to prepare a detailed research proposal, with the guidance of a supervisor. This proposal is used to decide if you are admitted to the research phase.

The **purpose** of a research proposal is to **convince** the research sponsor that you know the previous work on a subject and that you have an idea on how to go beyond it. A reviewer should be able to read a proposal and be taken along a path from a **research problem** (what is not known) -ÂŰ complete with a sound literature review that proves there is really a prob- lem that has not been solved ÂŰ- to **research objectives** and then a sound **research methodology**, also backed up by literature.

- The written research proposal is graded by scientific staff working in the relevant educational Programme, including the ITC Chair or Associate Professor responsible for quality control of the specialisation.
- The student presents the proposal to Programme staff and the Programme Director, and answers (pointed) questions.
- ! → Fundamental question: **Does this candidate have the ability to conduct MSc research and write a thesis about it?**, assuming that the candidate will receive a normal amount of supervision?
- ! → Fundamental question: **Is the proposed research feasible** within the time allocated, and given the resources (secondary data, field support, ...) available?
- ! → Results of the assessment are submitted by the Programme Director to the Programme Board, which makes the decision **whether the student can continue with the MSc thesis phase**, or if their study must be terminated.

Students who are not admitted to the MSc thesis phase leave ITC with a certificate of attendance and their course record.

2.6 Whose thesis is it, anyway?

Ultimately, **the contents of the thesis are the responsibility of the student** (candidate), not ITC in general, nor the Programme or supervisor in particular. The student designs the thesis project, collects the data (if applicable), writes the thesis, and defends it.

Each supervisor has a different style of working with students, and indeed each student and project is different. The following are **guidelines for supervision**:

- The student can expect **an average of two hours per week staff time** in the thesis writing phase. This includes face-to-face meetings, but also the time that the supervisor needs to read drafts, check calculations, check literature, etc.
- During the proposal writing phase the supervision time will be more, about **four hours**.
- Approximately every two weeks, the supervisor will give the student and Programme Director a brief **written report of the student's progress** with respect to the time plan, and the main points

that need attention. This is usually just a short summary; the student gets detailed feedback on drafts of thesis chapters and during meetings.

A checklist may be used during these meetings, to make sure all important points have been covered.

- **The supervisor does not do routine work for the student**. For example, if a satellite image needs to be georeferenced, and the student has forgotten how, the supervisor may point to the relevant section of a program documentation or lecture notes, but then the student must review the method and do the work.
- The MSc period is **not for individualised teaching**. If you have to learn things that were not taught during the course, you will have to learn them on your own, with advice from the supervisors of course, but they do not have time to give you individualised lectures or tutorials.

  Often, new computer programs that you want to learn come with tutorials and set-up guides; new statistical techniques are explained in textbooks of various levels of difficulty and often in the documentation of computer programs.
- The primary supervisor may be **absent** for several weeks at a time, either due to other work (e.g. consulting) or personal reasons (e.g. vacation). You will have a designated **second supervisor** to work with during those times. It is expected that the primary supervisor will brief the secondary supervisor about your status before s/he leaves, and vice versa on her or his return.

You can not expect the supervisor to tell you what to do, or what to think. They can give ideas, keep you clear of known dead ends or poor methods, suggest references, etc. But **you** are responsible for the work. During the thesis defence, **never** say something like "I did it this way because my supervisor told me to". A correct response is: "I did it this way because my supervisor suggested it; I then compared it with other methods and decided this is indeed best, because . . . ".

2.7 Time management

ITC MSc students (and other researchers) often complain that "time is too short". Yet, some manage to produce outstanding work. How can you cope with the perceived lack of time? Here are a few suggestions.

1. **The work must fit the time available**; design it accordingly. Time can not be expanded but work can be reduced.

2. **Your MSc thesis is not your life's work!**. It is a well-defined original piece of work in a well-defined scientific context answering a well-defined research question; however it must be feasible in the given time and with the given data-gathering possibilities. You will have plenty of time in your career to improve on and extend the work you begin in your thesis. So, **limit the scope** of your work accordingly.
3. The **quantity** of work is not as important as its **quality**. It is rare that an MSc thesis gives the definitive answer to a research question, because the time for data collection is short. However, it can be placed within the larger research context and well-argued.
4. **Set priorities**; concentrate on the most important points. For example, if you are comparing several methods of image classification, in your research proposal you should already have established which are the most relevant (i.e. which ones you really want results from) and which are more speculative (i.e. ones where results would be nice but not vital).
5. **Work smart**: Before undertaking tedious calculations or samplings, be sure you are calculating or sampling the right thing.
6. Make sure the simple things are done before moving to the complicated ones. There should be no problem in quickly writing up the Methods, for example.
7. Estimate how much time should be spent on each section of the thesis; **work from an outline**. Don't get stuck in any section; if it is proving too time-consuming, discuss with your supervisor ways to limit that part of the work.
8. **Plan ahead**: Make a work plan (in consultation with your supervisor) and stick to it, as much as possible. You have to limit the work you do in each phase.
9. **Be realistic** in your time planning. A human being is not a machine and needs food, sleep, social time and relaxation. Also, you should plan for the unexpected: setbacks both personal (e.g. sick time) and professional (e.g. slower-than-anticipated progress).
10. Do creative work at the times you work best; for example, some people write well in the morning, others while burning the midnight oil. Save routine tasks for other times.
11. Take time to relax and re-focus your energy.
12. Don't waste time writing things that are not central to *your* thesis.
13. Keep a **log book** of your work; this will allow you to show your

supervisors what you have done, how you did it, and where you had difficulties.

14. **Leave time to check and revise your work**. In particular, you should not be still producing results in the last weeks before the thesis is submitted; rather you should be putting the finishing touches on the conclusions and making sure the format is correct.

15. Make sure to **back up your computer files** at regular intervals. Material stored on the ITC servers (`M:` drive) are backed up every night by the IT department. Burn a CD once a week. Especially, make several CD copies of your primary data as soon as possible. Keep copies of each thesis draft; you may want to go back to (parts of) a previous version.

References

[1] Booth, W. C.; Colomb, G. G.; & Williams, J. M. 1995. *The craft of research*. Chicago: University of Chicago Press. ITC 001.818

[2] FAO. 1976. *A framework for land evaluation*. Soils Bulletin 32. Rome, Italy: FAO

[3] Galindo De Obario, M. 2005. *Water quality and its spatial variability in Lake Cuitzeo, Mexico*. MSc, International Institute for Geo-information Science & Earth Observation (ITC), Enschede (NL)

[4] Hornby, A. S. (ed.). 1995. *Oxford advanced learner's dictionary*. Oxford: Oxford University Press. ITC 038.20.951 reference

[5] Hudson, N. 1981. *Soil conservation*. Ithaca, NY: Cornell University Press, 2nd edition

[6] Nagelhout, A. 2001. *Performance analysis of small format aerial photography SFAP in assessing current status and trends in wind erosion : a case study in the Longonot – Kijabe hill area, Naivasha district, Kenya*. MSc thesis, International Institute for Aerospace Survey and Earth Sciences, Enschede (NL)

[7] Parsons, T. & Knight, P. G. 1995. *How to do your dissertation in geography and related disciplines*. London: Chapman and Hall. ITC 001.818

[8] Roling, N. 1997. *The soft side of land: socio-economic sustainability of land use systems*. *ITC Journal* pp. 248–262

[9] Rossiter, D. G. 1988. *The Automated Land Evaluation System: a microcomputer program to assist in land evaluation*. PhD, Cornell University

[10] Rossiter, D. G. 1990. *ALES: A framework for land evaluation using a microcomputer*. *Soil Use & Management* **6**(1):7–20

[11] Rossiter, D. G. & Van Wambeke, A. R. 1997. *Automated Land Evaluation System: ALES Version 4.65 User's Manual*. SCAS Teaching Series No. T93-2, Revision 6. Ithaca, NY: Cornell University, Department of Soil, Crop & Atmospheric Sciences, February 1997 edition

[12] Sánchez, P.; Couto, W.; & Buol, S. 1982. *The fertility capability soil classification system: interpretation, applicability and modification*. *Geoderma* **27**(4):283–309

[13] Sys, C.; Van Ranst, E.; Debaveye, J.; & Beernaert, F. 1993. *Land evaluation, Part 3 : Crop requirements*. Agricultural Publications 7. Brussels: General Administration for Development Cooperation

3 Statistical inference for research

Key chapter points

1. **Quantitative** statements are generally more useful than qualitative ones.
2. Statistical **inference** is used to quantify the **certainty** of quantitative statements.
3. A clear distinction must be made between **populations** and **samples**; the population that a sample represents must be unambiguously specified (§3.1).
4. There are two interpretations of probability: **frequentist** and **Bayesian** (§3.2).
5. **Bayesian** probability is the degree of rational belief that a statement is true; Bayesian inference works by updating **prior** to **posterior** probabilities, based on new observations. (§3.3).
6. **Frequentist** probability is the proportion of time an event would occur, should the experiment that gives rise to it be repeated a large number of times. Observations represent a sample from a population that has some fixed but unknown parameters (§3.4).
7. Frequentist **hypothesis testing** calculates the probability that **rejecting** a given **null hypothesis** is an incorrect decision. This involves the concepts of **significance levels**, **Type I and Type II errors**, and **confidence intervals** (§3.5).
8. Inferences are based on **statistical models**: their **functional form** and **parameters** (§3.7). The aim is to model the **structure**, and not the **noise**.
9. A clear distinction is made between model **calibration** ("*post*diction") and model **validation** ("*pre*diction") (§3.7.1).
10. **Correlation** and **regression** are often used uncritically and inappropriately; distinctions must be made between **fixed** and **random** predictors, and between **descriptive** and **predictive** models. (§3.8).
11. **Correlation** does not necessarily imply **causation**; this link requires **meta-statistical** reasoning.
12. When **law-like relations** are to be modelled, **structural analysis** should be used instead of regression (§3.8.1).
13. Models should be **parsimonious**; this avoids fitting noise rather than structure (§3.8.3).

Most research questions should be posed so that the answer is *quantitative*; this leads to deeper understanding and better information on which to base decisions. Kelvin made the definitive statement about the value of numerical measurement:

> "In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be." [10, 1:73]

A similar sentiment has been the inspiration for the development of *inferential statistics*, which seeks to quantify the plausibility of statements about the world. These inferences are the main result of scientific research. Davis [5, p. 11] puts it nicely:

> "Statistics ... may best be considered as the determination of the *probable* from the *possible*."

Here are some examples of statements we might like to make in the conclusions of a research project:

- "The projective transformation can successfully georeference a small-format air photo (SFAP) from ten ground control points measured with a single-receiver GPS".
- "In villages where a participatory GIS was developed there was less conflict between government plans and local goals."
- "Shifting cultivation systems have expanded in the past ten years, mainly at the expense of primary forest."

These statements need to be quantified; in particular we would like to give a precise meaning to words like "successfully", "less", "expanded", "mainly".

Some statements already are a quantitative statement, based on some observations, so seem adequate as they stand:

- "Primary forest covers 62% of the study area."
- "On 10–September-2000 Lake Naivasha contained $8.36 \cdot 10^9 m^3$ of water."
- "Twice as many boys as girls attend secondary school in District X."

Yet unless these are made by exhaustive sampling (visiting every pixel, pumping the contents of Lake Naivasha though a water meter, counting every school child), they are *uncertain*, so we'd like to give some *range* in which we are fairly sure the true value lies.

We then use the *inferential paradigm*:

- We have a *sample* which represents some *population*;
- We want to make a quantitative statement about the population;
- This requires us to *infer* from sample to population.

3.1 Basic concepts

Population

Populations and samples Any inferential statement refers to a *population*, which is the set of objects about which we want to make this statement. Most of these objects have not been observed, yet we would like to make some statement about them. It can be surprisingly difficult to precisely specify the population.

In the small-format air photo (SFAP) example, the population might be:

- All small-format air photos (SFAP) that were taken in this project
- All SFAP that could have been taken under 'similar' situations (how 'similar'? only in the study area? in 'similar' areas?);

Sample

Sampling design

The *sample* is that portion of the population that we have observed. The relation between population and sample is the *sampling design*. This is specified by the experimenter. The same must somehow *represent* the population for inferences to be correct. We must carefully consider how the sample was planned and executed. If some locations in geographic or feature space were purposely not sampled, it is difficult to argue that they be included in the population about which statements can be made. Typical reasons for not sampling include:

- Inaccessibility
- Lack of permission
- Uninteresting for purposes of the study (e.g. rock piles in a study of soil moisture)

In the first two cases, it *might* be possible to argue that the unsampled areas are similar to ones that were sampled, but this would have to be a convincing argument that the reasons for lack of access has no relation to the phenomenon being sampled. This is unlikely if, for example, a border village is not included in the sample because it is "too unstable".

Statements of fact vs. inference The word "statistics" by itself simply refers to mathematic summaries of a set of data. These are just statements of fact:

- "The median sigma of georeferencing of 14 photos was 5.16 m".
- "Participants in the workshop had from two to ten years of formal education."
- "Twelve of the 40 crop fields surveyed in 2004, with an area of 6.3 ha out of the 18 ha total crop land surveyed, were covered by primary forest in 1990".

It is another step entirely to draw *inferences* from such statements. This must be from a sample (what has been observed) to a population (what could have been observed), being careful to specify the population:

- "The median sigma of georeferencing with the projective transform is no greater than ..."
- "Small farmers in the district have from ... to ... years of formal education."
- "X% of the crop fields active in 2004 and Y% of their area were covered by primary forest in 1990".

To make such statements, we use statistical inference.

For example, here is a set of the sigma values from the projective transform applied to 14 small-format air photos from a particular study [14]:

```
4.36 3.63 6.01 3.78 7.58 8.36 5.18
4.77 4.80 7.18 5.79 5.14 5.42 3.81
```

It is just a statement of fact to say that the minimum of this set is `3.63`, the maximum is `8.36`, the median (half greater, half less) is `5.16`, and so forth. It is even a statement of fact to report the sample mean and standard deviation, computed according to the standard formulas, as `5.145` and `1.453`. But what we'd really like to say is how successful this procedure *would* be (or would have been) when applied in "similar" circumstances. This has several related meanings:

1. If, now, photos have already been taken and GPS points collected in the study area but have not yet been processed;
2. If, hypothetically, a different set of photos had been taken and GPS points collected in the same area during the same mission;
3. If, in the future, more photos are taken and GPS points collected in the same area and under the same conditions;
4. As these three, but in other areas.

We can not avoid this sort of “meta-statistics”, and it leads us to consider the *plausibility* (not provability!) of each. In all cases we are arguing that the data in hand are a *representative sample* of the larger *population*.

The first statement seems the most secure: the population is then all photos and GPS points that were obtained. That is, the success with 14 photos can be used to predict the success with the rest of them. The second is similar but deals with a hypothetical population: all the photos that could have been taken; those that were are a sample of what was possible. The third is of interest if we want to repeat the study in the same area and the fourth if we want to extend it.

3.2 Frequentist and Bayesian interpretations

Meta-inference, that is, what do we really mean with an inferential statement, is still a contentious topic. There are two principal interpretations [8]:

- *Frequentist*, also called *classical* or *British-American*; and
- *Bayesian*.

The two approaches begin from quite different ideas about what is meant by “probability” and then carry these differences over to methods of inference.

Historically, the frequentist approach was developed under the leadership of R A Fisher, a statistician working at the Rothamstead Experimental Station in England, and was propagated in his highly-influential works *Statistical methods for research workers* (first appearing in 1925) and *The design of experiments* (1935) and by his disciples such as Neyman and Pearson. He worked in the 1930’s at the Iowa State University (USA) and there influenced well-known workers such as Snedecor, Cochran, and Yates. Because of the close historical connection with field and laboratory research, and the well-developed theory of inference promulgated in many texts [e.g. 3] and computer programmes, the frequentist approach is the most common in practical work today.

The Bayesian approach is named for the English nonconformist minister and mathematician Thomas Bayes (1701–1761) but he did not develop it; he is however responsible for the first statement of *Bayes’ Rule* of probability (§3.3), published posthumously in 1763. Inspired by Bayes’ ideas about the meaning of inference, a group of statisticians, including Laplace, Jeffreys, de Finetti, Wald, Savage and Lindley, developed another view of statistical inference, known (somewhat misleadingly) as *Bayesian* inference [11]. This approach can reproduce the frequentist interpretation, but can also be extended to a much richer set of inferences where frequentist methods fail.

3.3 Bayesian concepts

Bayesian interpretation of probability The Bayesian viewpoint begins from a subjective definition of probability: it is the *degree of rational belief* that I have that something is true. The restriction to "rational" beliefs means that certain rules of consistency must be followed; I can't simply state a belief with no evidence. In this viewpoint, all probability is conditional on evidence, and can be updated in view of new evidence.

The Bayesian goes further and asserts a frankly *subjective* view of probability: any parameter that we are trying to estimate is *not* fixed, i.e. some hypothetical "true" value, but instead is something we want to develop a personal *probability distribution* for. Naturally, we want a distribution that is consistent with all the evidence we can find, but we give up the attempt to narrow down the estimate to some hypothetical but ultimately unknowable true value. Also, there are limits on our personal distribution: it must in some sense agree with distributions estimated by others with similar subjective beliefs.

Types of probability

- *Prior* probability: before observations are made, with previous knowledge;
- *Posterior* probability: after observations are made, using this new information;
- *Unconditional* probability: not taking into account other events, other than general knowledge and agreed-on facts;
- *Conditional* probability: in light of other information, specifically some other event(s) that may affect it.

The distinction between conditional and unconditional probability depends on one's standpoint with respect to the possible conditioning event.

Simple form of Bayes' rule In its simplest form, *Bayes' Rule* is used to update a *prior* probability $P(A)$, based on new information that an event B with prior probability $P(B)$ has occurred, and knowing that the conditional probability $P(B|A)$ of B given A, to a *posterior* conditional probability $P(A|B)$ [3, 1.3.5]:

$$P(A|B) = P(A) \cdot \frac{P(B|A)}{P(B)} \tag{1}$$

The last factor is the proportion by which the prior is updated, sometimes called the *likelihood function*.

Equation 1 is derived by reformulating the definition of intersection probability from conditional probability:

$$\begin{aligned} P(A \cap B) &= P(A|B) \cdot P(B) \\ &= P(B|A) \cdot P(A) \end{aligned} \tag{2}$$

Equating the two right-hand sides and rearranging gives the rule.

This rule can be used for diagnosis. For example, suppose we have a fever (event B) and therefore suspect that we may have malaria (event A); we would like to calculate the probability that we in fact have malaria, so that we can take the appropriate medication. To compute this, we need to know:

1. The conditional probability of a person with malaria having a fever, $P(B|A)$, which we estimate as, say, 0.9 (some people who are infected with malaria don't have a fever);

2. The unconditional probability $P(A)$ of having malaria, i.e. the proportion of the population that has it, say 0.2; this is the *prior* probability of having malaria before looking at our symptoms;

3. The unconditional probability of having a fever from whatever cause, say $P(B) = 0.25$.

These probabilities can be estimated from a large random sample of people, independent of their health, where they are tested for malaria to give $P(A)$ and observed for fever to give $P(B)$, and together to give $P(B|A)$, the presence of fever in those that tested positive for malaria. Note that this prior would be quite different in different locations. Then the *posterior* probability that, given that an individual has a fever, that they have malaria is $P(A|B) = 0.2 * (0.9/0.25) = 0.72$. The probability of malaria has been greatly increased from the prior (0.2) because the presence of fever is so closely liked to the disease. The likelihood function was thus $0.9/0.25 = 3.6$; the odds increased by 3.6 times in the presence of the information about the symptom.

If fever were more prevalent overall in the population, or if a smaller proportion of malaria sufferers showed a fever, the updated probability would be different. For example if $P(B|A) = 0.5$ (fever less symptomatic), then $P(A|B) = 0.2 * (0.5/0.25) = 0.4$; if $P(B) = 0.5$ (fever is more common overall), then $P(A|B) = 0.2 * (0.9/0.5) = 0.45$; in both cases it is less likely that our symptom (fever) indicates the disease (malaria). If malaria were less prevalent overall in the population, the posterior probability will be reduced proportionally; this is because fever from other causes is now more likely.

General form of Bayes' rule Bayes' rule has a general form which applies when a sample space A of outcomes can be divided into a set of mutually-exclusive outcomes $A_1, A_2, \ldots$. Then the conditional probability of any of these outcomes A_i, given that event B has occurred, is [3, 1.3.6]:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)} \tag{3}$$

An example here is the land cover class at a particular location. This is one of the possibilities given by a legend. The *prior* probability $P(A_i)$ of a location belonging to class i is estimated from prior knowledge of the area to be mapped, perhaps a previous map or even expert opinion. The conditional probability $P(B|A_i)$ of some event (such as an aspect of a spectral signature) in for all possible land must also be given either from theory or statistical estimation. Then we can compute the *posterior probability* that a given location is in fact in the given class. This is precisely what "Bayesian" image classification algorithms do.

To take a simple example, consider a legend with three classes: open water (A_1), grassland (A_2), and forest (A_3) with prior probabilities $P(A_1) = 0.1$, $P(A_2) = 0.4$, $P(A_3) = 0.5$; these must of course sum to 1. That is, we expect the final map to have 10% open water. The event which is used to update this prior could be an NDVI < 0.2, for which we could estimate $P(B|A_1) = 0.95, P(B|A_2) = 0.02, P(B|A_3) = 0.05$; that is, in known pixels of water (from a training set), 95% of them had NDVI < 0.2; for grassland and forest there were only 2% and 5%, respectively with such low NDVI. Now, if we observe a pixel with NDVI < 0.2, we compute its posterior probability of in fact being open water as:

$$P(A_1|B) = \frac{0.95 \cdot 0.1}{0.95 \cdot 0.1 + 0.02 \cdot 0.4 + 0.05 \cdot 0.5} = 0.7422$$

The information that this pixel has a low NDVI has increased the probability that it represents open water from 0.1 (in the absence of spectral information) to 0.7422; the likelihood function was thus 7.422. The probability of being grassland or forest are similarly calculated as 0.0625 and 0.1953, respectively; note that the three probabilities add to 1 as expected. You might be surprised by the fairly high probability that the pixel is forest (nearly 20%); but recall that we expect half the map to be forest, and an appreciable proportion (5%) of pixels from these areas have low NDVI.

These formulas are not at all controversial in case the prior absolute and conditional probabilities are known. However, even if they're not, we may have some idea about them from previous experience, and that should give us better results than simply accepting the *non-informative* priors, i.e. that all outcomes $P(A_j)$ are equally-likely. In the above example, if we didn't know anything about the overall proportion of land covers in the area, we'd take $P(A_1) = P(A_2) = P(A_3) = 0.\bar{3}$, and compute the posterior probability of water, given low NDVI, of:

$$P(A_1|B) = \frac{0.95 \cdot 0.\bar{3}}{0.95 \cdot \bar{3} + 0.02 \cdot \bar{3} + 0.05 \cdot \bar{3}} = 0.9314$$

with the probability of being grassland or forest being 0.0196 and 0.0490, respectively. These probabilities are much lower than computed above (because the prior probability of these classes is lower); thus we see the major influence of prior probability. This has led to criticism of this approach as being *subjective*. But in image classification we often do have estimates of the proportion of various land uses or covers, either from previous studies or just reconnaissance; all classes are not *a priori* equally likely at each pixel in the classified image.

Bayesians argue that we are rarely in a state of ignorance about the object of study, and it makes sense to take account of what we already know. The medical diagnosis example supports this: doctors would be foolish not to take into account the difference between *a priori* rare and common diseases, even if they can not put a precise number on the relative occurrence. It's much more likely that someone with a fever in Yaoundé has malaria than someone with a fever in Enschede, and the doctors in those two places should not reason otherwise.

Practical problems with the Bayesian approach For a single condition there is no problem. But of course diseases have many symptoms, and land covers give rise to many spectral conditions, and these are often not completely independent. So Bayes' rule can't simply be applied sequentially, symptom-by-symptom, it has a much more complicated form when there are conditional probabilities between conditions.

3.4 Frequentist concepts

Frequentist interpretation of probability To a frequentist, the probability of an event is the proportion of time it would occur, should the experiment that gives rise to the event be repeated a large number of times. This is intuitively-appealing in the case of throwing dice, for example; we can imagine throwing dice in the same way many times. It is less appealing if we think of an agricultural yield trial; in this case we're imagining that the trial could have been done in many similar locations, in many similar years. Yet since we can't repeat the same conditions, the interpretation of 'frequency' becomes difficult; there is always a hypothetical aspect to the argument.

In this view, the observed data from an experiment represent a sample from a population that has some fixed but unknowable parameters. For example, we have evaluated the transformation sigma of a set of air photos with a set of GPS measurements; if we decide these represent all possible photos and GPS readings that could have been taken on the day, this is the population about which we'd like to make some statement, for example, how successful is the projective transformation in the study area.

Frequentist and Bayesian approaches agree exactly in some situations:

1. Uninformative prior probabilities of various outcomes; or
2. Exactly-known (objective) prior probabilities of various outcomes.

3.5 Frequentist hypothesis testing

A common use of frequentist inference is to decide whether a hypothesis is probably true or false. More strictly, the frequentist can give the probability that rejecting a given hypothesis is an incorrect decision. This has a clear interpretation for the decision-maker: it's the chance of making a wrong decision. It also has an interpretation for the scientist: the chance of making an incorrect statement about nature.

The null and alternate hypotheses Frequentist reasoning distinguishes the null and alternate hypotheses:

- The *null* hypothesis H_0: Not rejected until proved otherwise ("innocent until proven guilty"); if the evidence is not strongly against this, we can't reject it.
- The *alternate* hypothesis H_1: Something we'd like to prove, but we want to be fairly sure

A classic example of a null hypothesis is that a new crop variety does *not* have a higher yield than the currently-grown variety. The alternative in this case is that it does; note that this is a *one-tailed* alternate hypothesis because we don't care whether or not the new variety is worse.

On the other hand, we might have an *informative* null hypothesis; this is where some ideas from the Bayesian viewpoint are incorporated. For example, many studies may have shown that wood from hardwood species are denser than softwoods, so if we are repeating the study in a new area, we'd be quite surprised if the softwoods turned out to be denser. The null hypothesis then would be that the hardwoods are denser, unless proven otherwise; we might even use a specific numerical difference as the null hypothesis.

Significance levels and types of error In frequentist tests we need to quantify the risk of making an incorrect inference. These are of two types:

- α is the risk of a *false positive*: rejecting the null hypothesis when it is in fact true; this is called *Type I* error;
  - "The probability of convicting an innocent person" (null hypothesis: innocent until proven guilty)
- β is the risk of a *false negative*: not rejecting the null hypothesis when it is in fact false), this is called *Type II* error. The quantity $(1 - \beta)$ is called the *power* of a test.
  - "The probability of freeing a guilty person"

The following matrix shows how these kinds of error arise from the decision which we take and the truth of the matter (which of course we don't know):

	Null hypothesis H_0 is really . . .	
Action taken	True	False
Reject	Type I error committed	success
Don't reject	success	Type II error committed

Note that in strict frequentist thinking we can never "accept" it; all we can say is that we don't have sufficient evidence to reject it. We can never say that it's probably true, only that it's probably not false.[5]

Deciding on a significance level In frequentist inference, α is set by analyst, whereas β depends on the form of the test. These must be balanced depending on the *consequences* of making each kind of error. For example, if the null hypothesis is that a new crop variety is no better than the current one:

- The cost of introducing a new crop variety if it's not really better, and the lost income in case the new crop is in fact worse (Type I error), vs.
- The lost income by not using the truly better variety (Type II error)

This reasoning is mirrored in concepts of law. The British-American legal system is heavily weighted towards low Type I errors (to keep innocent

[5] This sort of convoluted reasoning is frequently cited by Bayesians as evidence that the frequentist approach is misguided.

people out of prison, even if some criminals are walking free), whereas the Napoleonic system accepts more Type I error in order to lower Type II error (to keep criminals off the street, even if some innocent people are sitting in prison).[6]

Significance levels

Often α is set by convention, or several are reported with conventional levels:

- "Marginally Significant" : $\alpha = 0.1$
- "Significant" : $\alpha = 0.05$
- "Highly Significant" : $\alpha = 0.01$
- "Very Highly Significant" : $\alpha = 0.001$

This can roughly be equated to "sure", "very sure", "extremely sure" that a Type I error is not being committed. Which level we choose to accept is subjective – and here we see that Bayesian 'subjectivity' is not absent from frequentist inference.

3.6 Examples of frequentist inference

In the frequentist paradigm, there is one true value of a population parameter, and we try to estimate it from the sample. We compute the "best guess" estimate by some procedure which we justify from the *assumed characteristics of the underlying population.*

The most common inferences are *point estimates*, to infer the true value of a single parameter, such as a population mean or a correlation between two variables. Since we only are estimating from a sample, we can't pin such an estimate down exactly, so we also compute a *confidence intervals*, which is a range having a *known probability* of containing the true value, again under our *assumptions.*

A simple example of point estimation is of the population *mean* or centre of gravity. *If* we can assume that the n observations we make are from a single population, with (unknown) *identically–* and *independently–*distributed (abbreviation "IID") errors of observation, *then* the most likely ("expected") value of the true mean is given by the well-known formula:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4}$$

The notation $\hat{\mu}$ means that we are *estimating* the true population mean μ; whereas $\bar{x}$ is simply shorthand for the right-hand side. So $\bar{x}$ is not an inference, but $\hat{\mu}$ is.

[6] Or maybe the British and Napoleonic systems have opposite null hypotheses about human nature.

The interval which has probability $(1 - \alpha)$ of containing the true value is:

$$(\bar{x} - t_{\alpha/2,n-1} \cdot s_{\bar{x}}) \;\leq\; \mu \;\leq (\bar{x} + t_{\alpha/2,n-1} \cdot s_{\bar{x}}) \tag{5}$$

where $t_{\alpha/2,n-1}$ is Student's t with $n-1$ degrees of freedom at confidence level $\alpha/2$ and $s_{\bar{x}}$ is the standard error of the mean:

$$s_{\bar{x}} = \frac{1}{\sqrt{n}} \cdot \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{1/2} \tag{6}$$

Note that the confidence level α, say 0.05, is halved, say to 0.025 for each side of the interval, because this is a two-sided interval. The t-distribution must be used because we are estimating both the mean and variance from the same sample; for reasonably-large sample sizes the normal distribution itself can be used.

The null hypothesis here is that the true mean μ is the value estimated from the sample $\hat{\mu}$. The alternate hypothesis is that the true mean is not this value; outside the confidence interval we can be fairly confident in rejecting the null hypothesis.

Using the geometric correction example above, recall we had 14 values of transformation sigma:

```
4.36 3.63 6.01 3.78 7.58 8.36 5.18
4.77 4.80 7.18 5.79 5.14 5.42 3.81
```

from which we compute the sample mean `5.145` and sample standard error of the mean is `0.403`. These are not yet inferences about the population, only statements about the sample. Then we find the required value of t ($\alpha = 0.025$, 13 degrees of freedom) is 2.160.[7] Then the confidence interval that we assert covers the true mean with only 5% chance that we are wrong is:

$$\begin{aligned}(5.145 - 2.160 \cdot 0.403) \;&\leq\; \mu \;\leq\; (5.145 - 2.160 \cdot 0.403)\\ 4.274 \;&\leq\; \mu \;\leq\; 6.015\end{aligned}$$

Now we make the *inferential statement* "With only 5% chance of being wrong, I assert that the mean transformation error is at most 6.015 m".

The variability of small samples Figure 2 shows an example of inference: four samples of size 30 were drawn from a known normal distribution[8] and then we attempted to infer the true mean and standard deviation, which in this case was known. The four random samples gave estimates from 177 to 184.3 for the true mean (180) and 16.5 to 20.0 for the true standard deviation (20). This is typical of inferences from small

[7] R code: `qt(.975,13)`

[8] R code: `rnorm(30, 180, 20)`

samples. In this simulation we can draw as many samples as we wish, but in a field experiment where we are again assuming a true mean and standard deviation, and even assuming the distribution of the variable, we can not easily repeat the experiment, and certainly not in the exact same conditions.

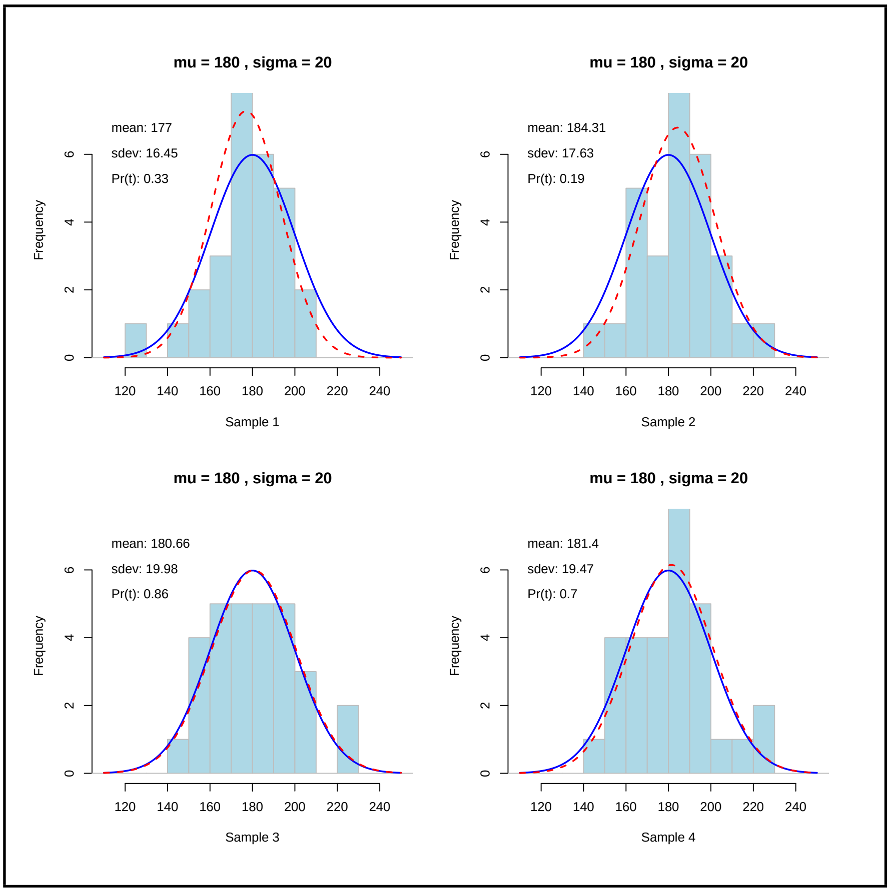

Figure 2: Four estimates of the parameters of a known population

What do we really mean by a 'confidence interval'? In the frequentist view, the confidence interval for a parameter is said to *cover* the its true value with some probability p. This means that if we would (or could) repeat the procedure many times, in that proportion of the cases the realised confidence interval would contain the true value of the parameter. For example, if $\alpha = 0.05$, in 95% of the hypothetical repetitions of the sampling the computed interval would in fact contain the true (but unknown) value. Looking at this from the other side, the one realised confidence interval we have from our one sample has probability $(1 - p)$ that it does *not* contain the true value; that is the risk of a Type I error.

Map accuracy assessment

An appropriate statement for map accuracy assessment based on a binomial test of ground truth vs. mapped class might be:

> "This land cover map was made primarily by manual and automatic interpretation of a satellite image, so that most locations have not been visited. However, a representative sample of locations was visited to assess the thematic accuracy of this map, which is reported as the proportion of 15x15 m ground locations areas where the land cover as reported on the map agrees with the actual dominant land cover. Under the assumption that the ground truth locations were representative of all the possible samples that could have been chosen, we used the binomial distribution to calculate a 90% 'confidence interval', which gives a minimum and maximum accuracy. These intervals as reported here have a nine-in-ten chance of containing the true accuracy. If we had been able to take a large number of similar samples, 90% of the confidence intervals calculated from these would have contained the true accuracy. We have no way of knowing whether the one sample we did take is one of those 90% or one of the 10% where the computed confidence interval, reported here, does not contain the true accuracy. So, there is a one-in-ten chance that the true accuracy is outside the interval we report here."

Specifically for this example, given an unbiased sample of size n, with n_t successes, the proportional accuracy and its standard deviation are estimated as parameters of the binomial distribution by [16]:

$$p = n_t / n \tag{7}$$

$$s = \left[\frac{p \cdot (1 - p)}{n} \right]^{1/2} \tag{8}$$

If the sample size is large enough[9], the confidence interval of the estimate may be approximated as:

$$p \pm \left[s \cdot Z_{1-\alpha} + \frac{1}{2n} \right] \tag{9}$$

where $Z_{1-\alpha}$ is the two-tailed normal score for the probability of non-coverage α; this can be obtained from tables or computed in software. The factor $1/(2n)$ is a small-sample correction The lower and upper limits as computed by Equation 9 are truncated at 0 and 1, respectively, if necessary.

To be specific, suppose we have $n = 163$ total ground truth samples, of which $n_t = 86$ are correctly classified. Then $p = 0.5276$ and $s = 0.0391$. To limit the probability of non-coverage to 5%, the corresponding area under the normal curve is $\Pr = 0.95$, which is obtained for the two-tailed test with $Z = 1.96$[10], so that the 95% confidence interval for p is

[9] For small samples, especially if p is near 0 or 1, the confidence interval must be determined directly from the binomial distribution as explained by Rossiter [16].

[10] R code: `qnorm(.975)`

$[0.4479 \ldots 0.6073]$. This is interpreted to mean that if we had repeated the same sampling scheme a large number of times, in 95% of these samples the observed accuracy would be somewhere between 44.8% and 60.7%. We are taking a 5% risk that the true proportion is $< 44.8\%$ or $> 60.7\%$.

(Note that we can narrow the confidence interval at the expense of a higher risk of non-coverage. For example, increasing this risk to 10%, we obtain $Z = 1.64$[11] and an interval for p of $[0.4602 \ldots 0.5950]$, i.e. about 2.5% narrower. Increasing the risk to 20%, i.e. a one in five chance of the true value being outside our calculated interval, we obtain $Z = 1.28$ and an interval for p of $[0.4744 \ldots 0.5808]$, now 5.3% narrower.)

3.7 Building a statistical model

Every inference we make is based on an underlying statistical model. For example, an inference about a population mean depends on the assumed distribution of the variable (normal, log-normal, Poisson, Weibull ...). There are three steps:

1. Selecting a *functional form*, i.e. the model to be fitted;
2. Determining the *parameters* of the model; this is called *calibration*;
3. Determining how well the model describes reality; this is called *validation*.

The following conceptual equations show the inferences we are making:

- Observations = f (Structure, Noise)
- Observations = f (model, unexplained variation)
- Observations are a subset of Reality, so ...
- Reality = f (Structure, Noise)
- Reality = f (deterministic processes, random variation)

The aim is to **match our model with the real deterministic process** and **match our estimate of the noise with the actual random variation.** It is equally an error to model the noise (*overfit* the model) as to not model the process (*underfit* the model).

Evidence that a model is suitable For most datasets a numerical solution can be computed for many models. The question naturally arises as to whether it should be. In other words, is a model *meaningful* or *applicable*?

[11] R code: `qnorm(.95)`

There are two levels of evidence:

1. **external** to the model:
   (a) what is known or suspected about the process that gave rise to the data; this is the connection to the reality that the model is trying to explain or summarise;
   (b) how well the model fits further data from the same population: success of *validation* against an independent dataset
2. **internal**: from the model itself:
   (a) how well the model fits the data (success of *calibration*);
   (b) how well the fitted model meets the *assumptions* of that functional form (e.g. examination of regression diagnostics).

For example, the set of errors associated with georeferencing a satellite image from control points identified on a topographic map would seem to conform to the model of many small, independent errors[12] that we know (from theory) give rise to a normal (Gaussian) distribution. So it makes sense to estimate the standard deviation (so-called "sigma") of that distribution, to evaluate the average size of these errors and therefore the quality of the transformation.

However, even in this example we may find evidence that the errors are not independent:

- the distribution of individual errors across the image does not seem to be random → georeference sections of the image separately?
- the distribution of individual errors does not seem to be fitted by a normal distribution → use a different transformation? exclude some points? (but on what basis?)

This last point highlights the *assumption* underlying the Gaussian model: errors are all the result of small, random processes. If we make a gross error (e.g. mis-identify a road intersection on the image with one several km away) this is a different kind of error, which violates the model, and that is why we are justified in eliminating it, once it is identified.

3.7.1 Model calibration vs. model validation

The process of fitting a model to observed data is *calibration*, that is, the model parameters are adjusted ('calibrated') to best fit the available experiments. In the case of regression, this is part of developing the equation, given a functional form. This yields a goodness-of-fit measure

[12] map compilation and printing, image distortion, map registration to digitiser, . . .

such as R^2 (the *coefficient of determination*), which expresses how well we were able to match the model to the data. This is the complement of the *residual sum of squares* (RSS) as a proportion of the *total sum of squares* (TSS):

$$
\begin{aligned}
R^2 &= 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \\
RSS &= \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 \\
TSS &= \sum_{i=1}^{n} (z_i - \bar{z})^2
\end{aligned}
$$

where $\hat{z}_i$ is the predicted (modelled) value and $\bar{z}$ is the sample mean. If there are no residuals, RSS = 0 so that $R^2 = 1$; if the model explains nothing, RSS = TSS so that $R^2 = 0$. However, this only measures how well the model fits the data set, i.e. how well it is *calibrated*.

Once a functional form is selected, we estimate its parameters by formulas that were developed *assuming* the functional form is correct, e.g. maximum-likelihood estimators. For example, having decided on a simple linear regression, we must estimate the slope and intercept of the best-fit line; the maximum-likelihood method if all errors are independent and identically-distributed is least-squares. This is *model calibration*.

Postdiction vs. prediction

Another name for calibration is *postdiction* (as opposed to prediction), from the Latin 'post' (after) and 'dicere' (to say). This allows us to use the past (already observed) to make probabilistic statements about the how well the observations are explained by the calibrated model. If the observations were representative of a population, we would expect to obtain the same parameters, within experimental and observational error, in similar repeated studies. However, there is no way to be sure that, because we can't in general re-do the study. We can compare the predicted and actual values of our one sample, to see how well they match; this is the *goodness-of-fit* with respect to the sample. This tells us how well the model can match the sample, but it says little about how well it would match other similar samples. An example is the reported coefficient of determination (R^2) from a regression; this is a measure of the success of *calibration* (postdiction).

If we have a second independent sample, we can compare its values with what the model predicts. Note that the model calibration procedure did not use these observations, so this is an independent test, which can fairly be termed *validation*.

There are several measures of validity:

- *Root mean squared error* (RMSE) of the *residuals*: the *actual* vs. *estimate* (from the model) in the *validation* dataset; lower is better:

$$\mathrm{RMSE} = \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right]^{1/2}$$

- *Bias* or mean error (ME) of estimated vs. actual mean of the *validation* dataset; should be zero (0):

$$\mathrm{ME} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

- *Gain* of the least-square fit of estimated vs. actual data; this should be 1, otherwise the estimate does not increase at the same rate as the actual data.

These can be visualised by plotting fitted vs. actual values on the same scale, with a 1:1 line (Figure 3). The residuals are the vertical distances from this line:

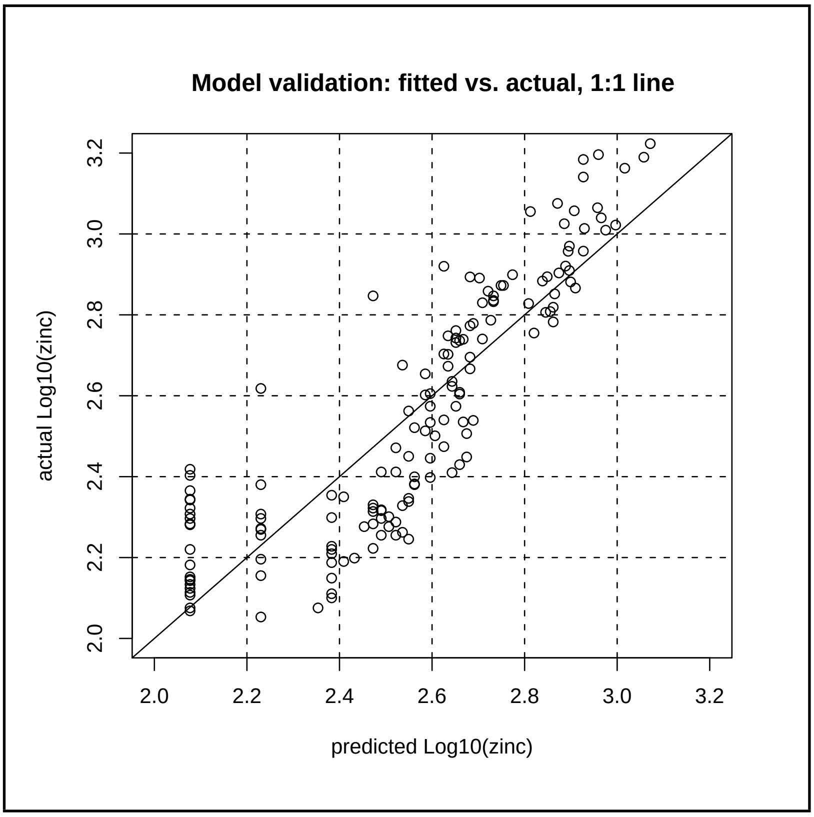

Figure 3: Validation with an independent dataset

The null hypothesis of no bias (intercept 0) or gain (slope 1) (i.e. the model is valid) can be tested by fitting a regression model to the actual vs. fitted values and simultaneously testing the two regression parameters of this model with an F-test [13]. A simpler approach is to consider the tests of each parameter separately in the regression analysis of variance table.

Unimodal vs. multimodal populations We can always compute a sample mean; this is just a summary statistic. But we typically do so to *infer* the mean of the population of which the sample is representative. But, how do we know our sample comes from a population with only one central tendency?

For example, is it helpful to speak of "the mean" of the 400-observation sample whose histogram is shown in Figure 4?

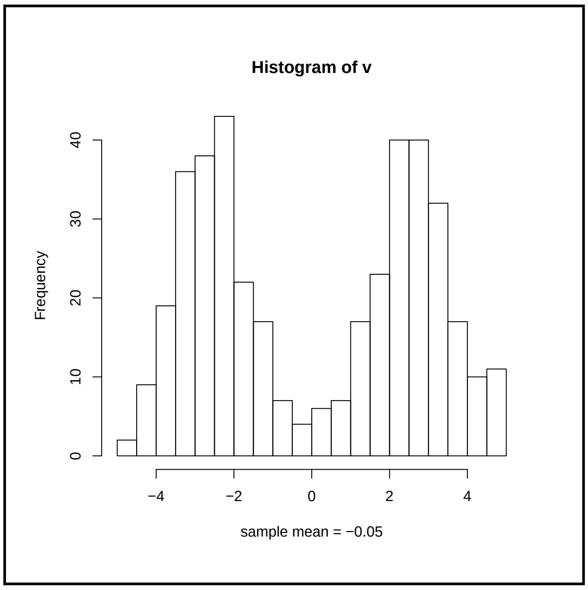

Figure 4: Histogram of a bimodal sample

It seems likely that there are two distinct populations in this1 sample[13],so that we'd like to estimate are the means of the two populations, not the (meaningless) overall mean.

This becomes even clearer if we imagine calculating a confidence interval for the population mean.

3.8 Conceptual issues in correlation and regression

Correlation and various kinds of regression are often misused. There are several articles that explain the situation, with examples from earth science applications [12, 19]. A particularly understandable introduction

[13] In fact there *are* two populations; this sample was created with the R code `v<-c(rnorm(200,-2.5,1),rnorm(200,2.5,1))`

to the proper use of regression is by Webster [20], whose notation we will use.

Correlation vs. causation A fundamental distinction must be made between two of concepts:

1. The *relation* between two or more variables, often described mathematically as the *correlation* ('co-relation');
2. The *causation* of one variable by another, often described by *regression* techniques.

This second is a much stronger relation than the first. The issue of causation also includes some *conceptual model* of how the two phenomena are related. Statistics can never prove causation; it can only provide evidence for the strength of a causative relation supported by other evidence. Thus we must always make a *meta-statistical* argument about causation.

Description vs. prediction Regression analysis can be used for two main purposes:

1. To *describe* a relation between two or more variables, whether the relation is supposed to be causative or not;
2. To *predict* the value of a variable (the *predictand*, sometimes called the *dependent* variable or *response*), based on one or more other variables (the *predictors*, sometimes called *independent* variables.

These can lead to different inferential procedures.

A statistical model that does not assume causation can still be useful for prediction. For example, the prevelance of two plant species may be correlated, so that we can develop a model to predict the presence of one from the presence of the other, without having to make any argument that the presence of one in any way "causes" the presence of the other. (In fact, we are more likely to argue that these have a common cause.) So we can have a regression equation that we use for prediction, not at all based on any notion of causation.

Types of models A *simple* correlation or regression relates two variables only; a *multiple* correlation or regression relates several variables at the same time. Modelling and interpretations are much trickier in the multivariate case, because of the inter-relations between the variables.

A *linear* relation models one variable as a linear function of one or several other variables. That is, a proportional change in the predictor results in a proportional change in the predictand or the modelled variable. Any other relation is *non-linear*.

Non-linear relations may be *linearisable* by means of a transformation of one or more variables, but in many interesting cases this is not possible; these are *intrinsically non-linear*.

Fixed vs. random variables A distinction is made between predictors which are known without error, whether fixed by the experimenter or measured, and those that are not. Webster [20] calls the first type a "Gauss linear model", because only the predictand has error, and the predictor a *mathematical* variable, as opposed to a *random* variable which is measured with error. The regression goes in one direction only, from the mathematical predictor to the random response, and is modelled by a **linear model with error**:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

There is no error associated with the predictors x_i, only with the predictand y_i. Thus the predictors are assumed to be known without error, or at least the error is quite small in comparison to the error in the model. An example of this type is a designed agricultural experiment where the quantity of fertiliser added (the predictor) is specified by the design and the crop yield is measured (the predictand); there is random error ε_i in this response.

An example of the second type is where the crop yield is the predictand, but the predictor is the measured nutrient content of the soil. Here we are modelling the relation as a **bivariate normal distribution** of two random variables, X and Y with (unknown) population means μ_X and μ_Y, (unknown) population variances σ_X and σ_Y, and an (unknown) correlation ρ_{XY} which is computed as the standardised (unknown) covariance $\mathrm{Cov}(X, Y)$:

$$\begin{aligned} X &\sim \mathcal{N}(\mu_X, \sigma_X) \\ Y &\sim \mathcal{N}(\mu_Y, \sigma_Y) \\ \rho_{XY} &= \mathrm{Cov}(X, Y)/\sigma_X\sigma_Y \end{aligned}$$

In practice, the distinction between the two models is not always clear. The predictor, even if specified by the experimenter, can also have some measurement error. In the fertiliser experiment, even though we specify the amount per plot, there is error in measuring, transporting, and spreading it. In that sense it can be considered a random variable. But, since we have some control over it, the experimental error can be limited by careful procedures. We can not limit the error in the response by the same techniques.

3.8.1 Structural Analysis

The regression of two variables on each other depends on which variables is considered the predictor and which the predictand. If we are predicting, this makes sense: we get the best possible prediction. But sometimes we are interested not in prediction, but in *understanding* a relation between two variables. This so-called *structural analysis* is explained in detail by Sprent [17] and more briefly by Webster [20] and Davis ([4, pp. 214–220] and [5, pp. 218–219]).

In structural analysis we are trying to establish the best estimate for a *structural* or *law-like* relation, i.e. where we hypothesise that $y = \alpha + \beta x$, where both x and y are mathematical variables. This is appropriate when there is no need to predict, but rather to understand. This depends on the prior assumption of a true linear relation, of which we have a noisy sample.

$$\begin{aligned} X &= x + \xi \\ Y &= y + \eta \end{aligned}$$

That is, we want to observe X and Y, but instead we observe x with random error ξ and y with random error η. These errors have (unknown) variances σ^2_ξ and σ^2_η, respectively; the ratio of these is crucial to the analysis, and is symbolised as λ:

$$\lambda = \sigma^2_\eta / \sigma^2_\xi \tag{10}$$

Then the maximum-likelihood estimator of the slope, taking Y as the predictand for convention, is:

$$\hat{\beta}_{Y.X} = \frac{1}{2s_{XY}} \left\{ (s^2_Y - \lambda s^2_X) + \sqrt{(s^2_Y - \lambda s^2_X)^2 + 4\lambda s^2_{XY}} \right\} \tag{11}$$

Equation 11 is only valid if we can assume that the *errors in the two variables are uncorrelated*. The problem is that we don't have any way of knowing the true error variance ratio λ, just as we have no way of knowing the true population variances, covariance, or parameters of the structural relation. We estimate the *population* variances σ^2_X, σ^2_Y and covariance σ_{XY} from the sample variances s^2_x, s^2_y and covariance s_{xy}, but there is nothing we've measured from which we can estimate the *error* variances or their ratio. However, there are several plausible methods to estimate the ratio:

- If we can assume that the two error variances are equal, $\lambda = 1$. This may be a reasonable assumption if the variables measure the same property, use the same method for sampling and analysis, and there is an *a priori* reason to expect them to have similar variability (heterogeneity among samples).

- The two *error* variances may be estimated by the ratio of the *sample* variances: $\lambda \approx s_y^2/s_z^2$. That is, we assume that the ratio of variability in the measured variable is also the ratio of variability in their errors. But, these are two completely different concepts! One is a sample variance and the other the variance of the error in some random process.
- The variance ratio may be known from previous studies.

Figure 5 shows the large difference that may result from viewing one variable as a function of the other or vice versa, compared to the structural relation between two correlated variables.

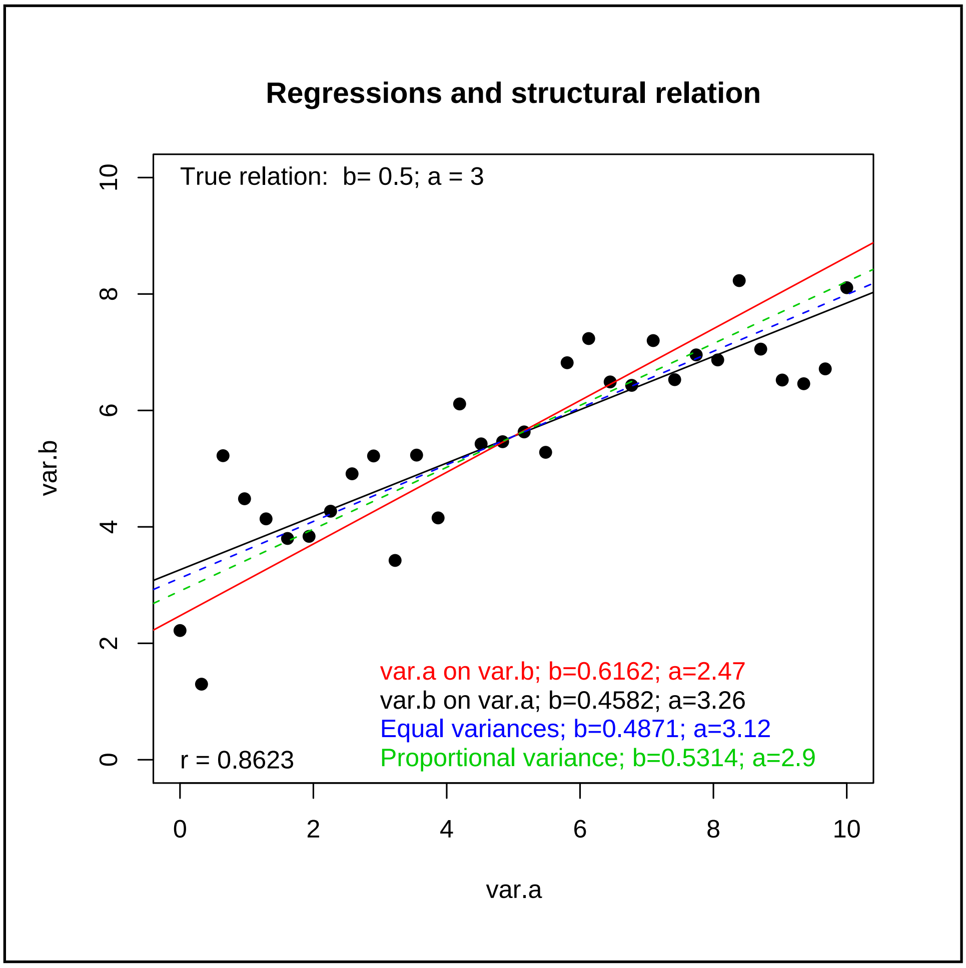

Figure 5: Regression in two directions compared to structural relations

3.8.2 Selecting the correct regression model

A classic example is provided by Anscombe [1], who developed four different bivariate datasets, all with the exact same correlation coefficient $r = 0.81$ and linear regression equation $y = 3 + 0.5x$ (Figure 6).

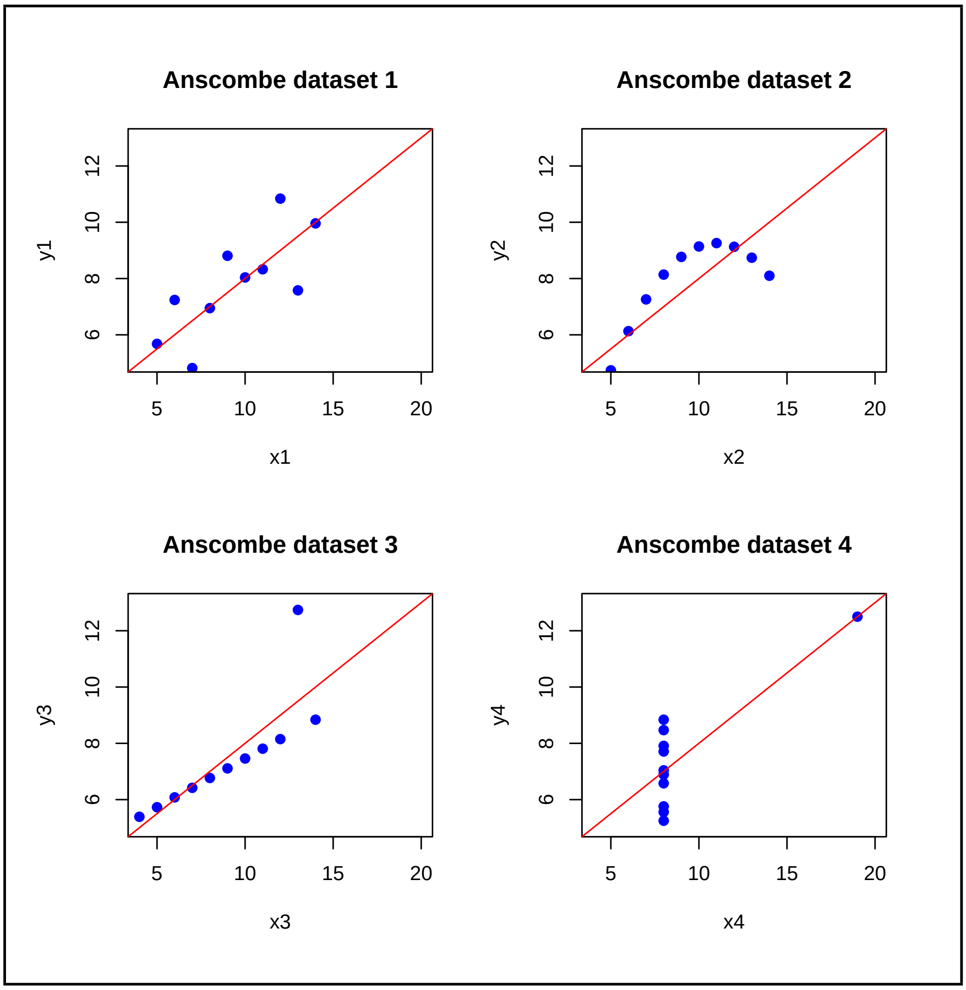

Figure 6: Anscombe's bivariate regression examples

The question is whether the linear regression model, i.e. that the value of y depends linearly on x, is applicable. Second, whether the least-squares estimate of the regression coefficients gives a correct summary of the relation.

1. Yes and yes, the data seem well-fitted by a line, and errors are equally-distributed around it;
2. No and no, the data seem to fit another functional form perfectly;
3. Yes and no, all the data except one perfectly fit a substantially-different line, $y = 4 + 0.346x$[14]
4. Yes and yes; except we are quite uncomfortable with the best estimate, because we suspect that if we took more observations at $x = 19$ we would see a similar spread to the observations at $x = 8$, and we have no way of knowing where the single observation is in

[14] *Robust* regression methods [2, 6, 18] can successfully fit this relation.

within this distribution.

How do we know that the chosen model is appropriate?

1. From *a priori* knowledge of the process;
2. From *internal evidence* when we try to fit the model.

In the second case there are many so-called *regression diagnostics* with which we can evaluate how well the model satisfies its assumptions. A common set of diagnostics examines the *residuals*, that is, the discrepancy of each fitted point from its observation. If any are unusually large, it may be that the observation is from a different population, or that there was some error in making or recording the observation. If large residuals are associated with large values, this is evidence of *heteroscedasticity* (i.e. variance is not constant across the range of the predictor). Texts on regression [e.g. 6] explain these in detail.

Figure 7 shows an example of a regression diagnostic for the Anscombe data. The 'diagnostic' here is that the residuals should show no relation to the fitted value; we can see that is the case in regression 1 (the 'normal' case) but violated badly in all the others. This gives evidence that the selected model was not correct.

3.8.3 Parsimony

This is a technical term used in statistics to express the idea that the simplest relation that explains the data is the best. Gauch Jr. [7] gives an accessible introduction to this concept. It is especially applicable in *multiple regression* models, where the model can be made increasingly complex, apparently explaining more and more of the dataset (as measured by the unadjusted R^2).

However, after a certain point the more complex model is explaining the *noise* (experimental error), not the *relation*. Even with only one predictor, it is always possible to fit n data points exactly by using a polynomial of degree $n - 1$. This effect is shown in Figure 8. The points should all lie on the dashed line (the true relation), but because of experimental error they deviate from it with error mean 0 and standard deviation 3; each experiment will have a different error. The best fits to two different sets of points for increasing polynomial degree are shown. Note that the underlying relation is the same. Also note that the lower-order (linear) fits are similar for both noisy datasets, but the higher-order fits differ greatly, as each fits its own noise, rather than the structural relation.

One measure, which applies to the standard linear model, is the *adjusted* R^2. This decreases the apparent R^2, computed from the ANOVA table,

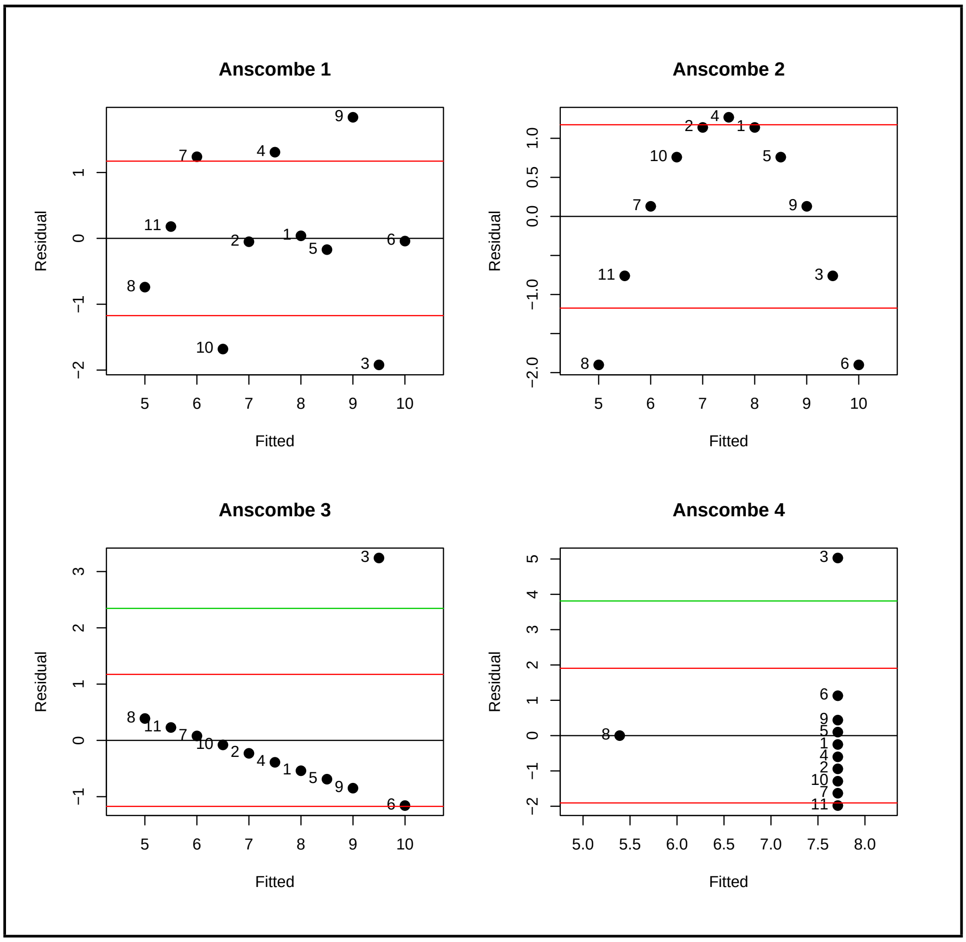

Figure 7: Anscombe's bivariate regression examples; residuals vs. fits

to account for the number of predictive factors:

$$R^2{}_{\text{adj}} \equiv 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

That is, the proportion of variance *not* explained by the model $(1 - R^2)$ is *increased* with the number of predictors k. As n, the number of observations, increases, the correction decreases. A more general measure, which can be applied to almost any model type, is Akaike's *An Information Criterion*, abbreviated AIC. The lower value is better.

References

[1] Anscombe, F. 1990. *Residuals.* In Eatwell, J.; Milgate, M.; & Newman, P. (eds.), *The new Palgreave: time series and statistics*, pp. 244–250. New York: W.W. Norton & Company

[2] Birkes, D. & Dodge, Y. 1993. *Alternative methods of regression.* Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc.

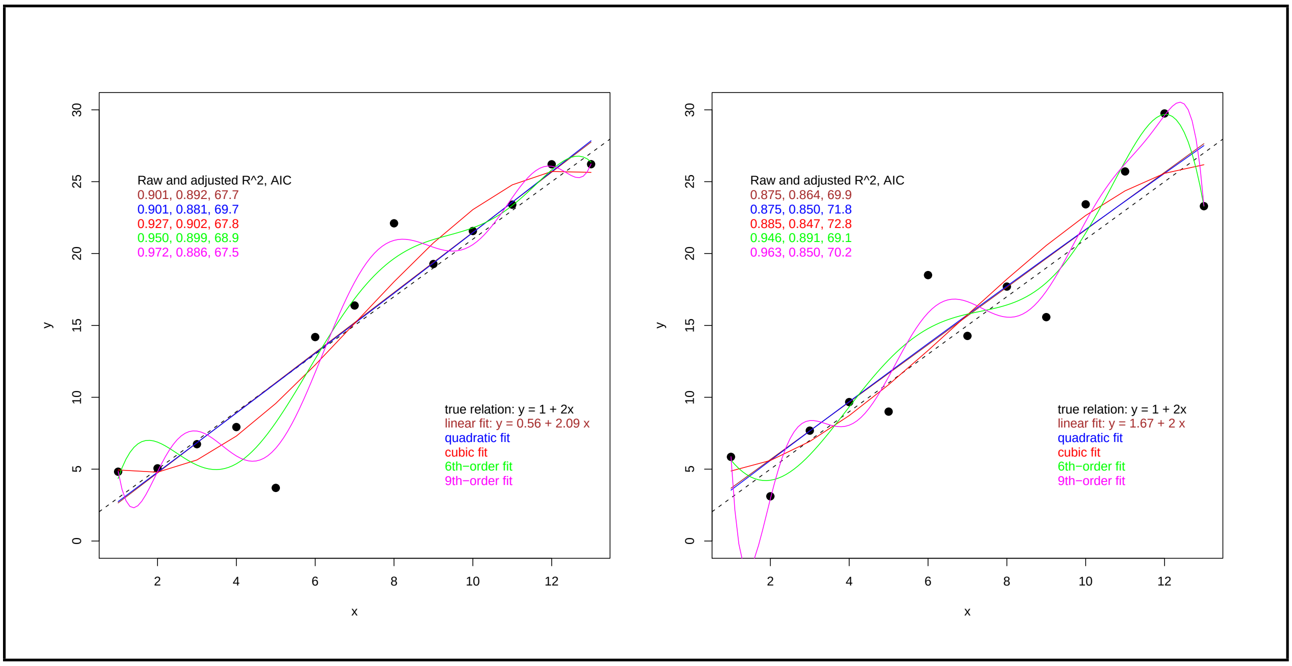

Figure 8: Fitting noise with higher-order polynomials

[3] Casella, G. & Berger, R. L. 2002. *Statistical inference*. Pacific Grove, CA: Duxbury, 2nd edition

[4] Davis, J. C. 1986. *Statistics and data analysis in geology*. New York: Wiley

[5] Davis, J. C. 2002. *Statistics and data analysis in geology*. New York: John Wiley & Sons, 3rd edition

[6] Fox, J. 1997. *Applied regression, linear models, and related methods*. Newbury Park: Sage

[7] Gauch Jr., H. G. 1993. *Prediction, parsimony & noise*. *American Scientist* **81**:468–478

[8] Gauch Jr., H. G. 2002. *Scientific method in practice*. Cambridge: Cambridge University Press

[9] Ihaka, R. & Gentleman, R. 1996. *R: A language for data analysis and graphics*. *Journal of Computational and Graphical Statistics* **5**(3):299–314

[10] Kelvin. 1883. *Popular Lectures and Addresses*

[11] Lee, P. M. 2004. *Bayesian statistics: an introduction*. London: Arnold, 3rd edition

[12] Mark, D. M. & Church, M. 1977. *On the misuse of regression in earth science*. *Mathematical Geology* **9**(1):63–77

[13] Mayer, D.; Stuart, M.; & Swain, A. 1994. *Regression of real-world data on model output: an appropriate overall test of validity*. *Agricultural Systems* **45**:93–104

[14] Nagelhout, A. 2001. *Performance analysis of small format aerial photography SFAP in assessing current status and trends in wind erosion : a case study in the Longonot – Kijabe hill area, Naivasha district, Kenya*. MSc thesis, International Institute for Aerospace Survey and Earth Sciences, Enschede (NL)

[15] Rossiter, D. G. 2003. *Introduction to the R Project for Statistical Computing for use at ITC*. Enschede (NL): International Institute for Geo-information Science & Earth Observation (ITC). URL: http://www.itc.nl/personal/rossiter/teach/R/RIntro_ITC.pdf

[16] Rossiter, D. G. 2004. *Statistical methods for accuracy assesment of classified thematic maps*. Technical note, International Institute for Geo-information Science and Earth Observation (ITC). URL: http://www.itc.nl/personal/rossiter/pubs/list.html#R_ac

[17] Sprent, P. 1969. *Models in regression and related topics*. Methuen's monographs on applied probability and statistics. London,: Methuen

[18] Venables, W. N. & Ripley, B. D. 2002. *Modern applied statistics with S*. New York: Springer-Verlag, fourth edition

[19] Webster, R. 1989. *Is regression what you really want?* *Soil Use & Management* **5**(2):47–53

[20] Webster, R. 1997. *Regression and functional relations*. *European Journal of Soil Science* **48**(3):557–566

4 Ethics & professionalism in science

Key chapter points

1. Scientific **ethics** are **rules of conduct** for carrying out scientific work (§4.1).
2. **Fraud** is any action which wilfully mis-represents the truth (§4.2); it has thee forms: **fabrication** (§4.2.1), **falsification** (§4.2.2), and **plagiarism** (§4.2.3).
3. **Plagiariasm** is knowingly representing the work of others as one's own; this includes **text**, whether directly copied or paraphrased, **data** and **ideas**.
4. A **simple rule to avoid written plagiarism**: Everything you write outside of quotation marks must be the result of your own creative effort.
5. **Intellectual property** is any product of a creative effort; it may be protected by copyright, which must allow **fair use**, e.g. for comparison; other uses usually require **license agreements** (§4.3).
6. **Professionalism** refers to scientists' behaviour towards the society in which they live (§4.4).
7. Research is embedded into the wider **social context**; the scientist must make ethical decisions about choices of topics and their effect on society (§4.6).
8. Relations between researchers and their **human subjects** and **local popluations** are subject to difficult ethical decisions (§4.7).

4.1 Ethics

In the context of scientific work, *ethics* refer to the *rules of conduct*: what is permitted. These rules have evolved along with science, both from more general codes of ethics such as religious value systems but also to aid scientific progress. The idea is that ethical behaviour isn't just "right" in some abstract sense, but also that it ensures good science. It also ensures that scientists are properly rewarded for their work.

In the wider sense, ethics also includes the relation between researchers and society as well as the relation between researchers and research subjects or colleagues; these are explored in §4.6 and §4.7, respectively.

Scientific ethics in the narrow sense (internal to the scientific community) is organised around two main principles:

- **Honesty and truth**: Science attempts to explain the natural world; technology attempts to manipulate the built world, including virtual 'buildings' such as computer systems. If we believe that there is an objective truth, we must be honest in reporting our observations of it.
- **Credit for work performed**: This is the currency of the scientific world; personal advancement of the scientist or engineer depends on receiving credit (and taking blame!) for what has been actually done by that person.

4.2 Fraud

Fraud is, broadly speaking, any action which wilfully mis-represents the truth. This can be the truth as to who did something (i.e. not correctly crediting someone with their idea or data) or the truth as to what was actually seen (i.e. data falsification or manipulation). You can commit fraud by *omission* (not saying something that should be said) as well as by *commission* (saying something false).

Intent to deceive

The key issue in fraud is the *intent to deceive* , in other words the *will-ful misrepresentation* of the facts (e.g. what was done, what was seen, who did what). When we read a piece of research, we may not accept the interpretations and conclusions of the author, but we expect that any statements of fact are indeed true, so that we can form our own conclusions or repeat the work.

The scientific enterprise (including ITC) is, justifiably, responds harshly when cases of fraud are detected. This can be years after the fact (see for example Broad & Wade [4] for the case of British psychologist Cyril Burt who falsified and invented data for a series of very influential studies on identical twins) but is more likely to be sooner. Scientists are naturally suspicious and inquisitive, and will probe behind what is written to find out what is true. ITC supervisors, Professors, and external examiners are very good at identifying suspect parts of the thesis and asking about them.

Examples are:

- results that seem too good to be true;
- data that shows very regular patterns, consistent with the hypothesis;
- data points from a small dataset that lie very close to a good model fit (e.g. regressions, variograms);
- beautiful writing in the middle of an otherwise sloppy text; and

- data that should have taken a long time to collect but which are supposed to have been obtained in a short fieldwork.

Fraud may be classified in three divisions, roughly in order of seriousness:

1. **Fabrication**: making up data, lying about procedures;
2. **Falsification**: manipulating data to obtain a desired outcome;
3. **Plagiarism**: taking credit for someone else's work.

Fabricating or falsifying data is the cardinal sin of science, since only with true data can we make progress towards the truth. Plagiarism is a crime against another author, stealing his or her credit.

4.2.1 Data fabrication

Fabricating data is inventing data or lying about the procedures by which it was obtained. This is the **cardinal sin** in science, because it can never be un-done. A simple example is filling in survey sheets without actually making field visits, based on what you expect to find. A bit less obvious but still fabrication is over-interpreting a survey response ("He said he wasn't sure about when his family came here, but to me 1965 seems about right, so I'll enter that") or field observation ("I don't see any gravel in the subsoil here but there really should be, so I'll enter them on the form").

Without accurate primary data, the entire research is invalid. You can always interpret or manipulate the data (with appropriate justification, of course), but that is a separate step from the primary data collection.

You should always **keep your primary field records and logs**. They are the ultimate proof of what you actually did.

4.2.2 Data falsification

Falsifying data is manipulating actual data to obtain a desired outcome. It comes in several forms: omitting 'inconvenient' observations as well as changing data values to more 'reasonable' ones.

- Discarding data during **sampling** is possible but (1) when explicitly acknowledged and (2) based on clear criteria.

  Example: a planned soil fertility sample was found to be located in the middle of an irrigation ditch; this can be discarded because it's not representative of the population being sampled (i.e. agricultural soils). This must be on the basis of criteria defined **prior** to beginning the sampling.

Example: a respondent in a household survey seems clearly to be mentally ill and delusionary. Record his or her answers, but add a note about their mental state as you interpret it, and then state that this response was discarded for the reason that, in your opinion, the respondent was not reliable. Another researcher can still make use of your primary observation if they disagree with your assessment of the respondent's state.

- Discarding data during **analysis** is possible but (1) when explicitly acknowledged and (2) based on clear criteria.

  A typical problem concerns so-called *outliers*, that is, data points that don't fit the pattern. In any case, they must be reported. But you don't have to include them in the analysis (e.g. to compute a correlation coefficient) if you can argue convincingly that they are **not part of the population being analysed.** Some possibilities:

  - Poor technique (but how do you that know only *this* sample was affected?)
  - Poor record-keeping (reflects poorly on your technique, but at least you are admitting it);
  - From a markedly-different site that is *not* included in the population you are studying.

  An obvious recording error (e.g. missing decimal point) may be corrected with no further observation, but this change should be shown in the original field book with a note.

  Leaving out an 'inconvenient' observation with no comment and no justification is fraud.

It's interesting that many of the advances in science come from researchers who rigorously pursued their data, or who noted anomalies in other researchers' data and tried to explain them. A classic example is the discovery of the microwave background radiation from the Big Bang by Wilson & Penzias.

4.2.3 Plagiarism

Plagiarism is defined and explained by many authors [e.g. 3, §11.5], more or less as follows:

Knowingly representing the work of others as one's own [6, p. 3]

This can occur many in several ways, for example:

1. Copying someone else's work;

2. Paraphrasing someone else's work, i.e. saying the same thing with slightly different words and phrasing;
3. Reporting someone else's work (e.g. fieldwork) as if it were your own;
4. Getting someone else to do your work for you ('ghostwriting');
5. Using a particularly apt term or phrase which you didn't invent.

Simple copying is easy to define, but some cases are not so straightforward. Here we go into detail on what is permitted and what is not, and the reasons for this. We start from some basic principles of honest writing:

Three golden rules

1. **Everything you write outside of quotation marks must be the result of your own creative effort.** Otherwise, you are taking credit for something you did not write.
2. **Every *idea* that is not your own must be credited.** Otherwise you are taking credit for the other person's idea.
3. **Every *fact* that you did not yourself establish must be credited.** Otherwise you are claiming direct knowledge that you do not have. This includes field or lab. work actually done by others which you are reporting.

Plagiarism by direct quoting without attribution is a common problem with ITC students for several reasons:

- The student feels that author is an all-knowing authority, and their text should not be altered;
- The student feels that the author has explained matters perfectly, and their text can not be improved upon;
- The student is not a confident writer (perhaps because they are not used to writing in English) and prefers to use a ready-made text;
- It is very tempting to cut-and-paste from easily-available electronic documents (web pages, full-text journals . . .).

Cut-and-paste plagiarism

The first reason is always false. The second may be true for the original author's purpose, but not for yours. The third may well be true, but paraphrasing is still plagiarism. If you really must, quoting is at least honest. The fourth (direct cut-and-paste) is **really stupid**; the same tools that the plagiarist uses to find the text will be applied by the examiner to find the text again and establish that it was plagiarised.

To be completely clear on this, here is an example of plagiarism by copying. First, from the original article by Bergsma [2]:

> Soil conservation is defined as the use of land, within the limits of economic practicability, according to its capabilities and the need to keep it permanently productive.

Second, from an MSc thesis, *not* written by Bergsma:

> Soil conservation is defined as the use of land, within the limits of economic practicability, according to its capabilities and the need to keep it permanently productive.

This is certainly plagiarism: straight copying. What if we add the citation?

> Soil conservation is defined as the use of land, within the limits of economic practicability, according to its capabilities and the need to keep it permanently productive [2].

This is not so bad, but it is still plagiarism. The author has credited Bergsma with the *idea* of this definition of soil conservation, but still implies that the actual *words* used are the author's interpretation, which they are not.

The correct way to use this exact definition and credit the author is:

> Bergsma [2] defines soil conservation as "the use of land, within the limits of economic practicability, according to its capabilities and the need to keep it permanently productive".

Or, if you don't want to use the author name as the subject of the sentence:

> Soil conservation is defined as "the use of land, within the limits of economic practicability, according to its capabilities and the need to keep it permanently productive" [2].

Note the use of quotation marks to set off the *exact words* of the original source.

Unless you intend to discuss the exact definition or wording, it is better to synthesize with other sources or adapt to your own argument. An example in this case might be:

> The concept of soil conservation was originally aimed at the physical protection of the soil from erosion at any cost and for indefinite time [10], but the emphasis is now on measures that are economically practicable and in line with the land's capabilities to provide productive and ecological services [2].

Here we use two sources to support an argument, and brings out the essence of what is meant by 'soil conservation', without plagiarizing either source. Bergsma is correctly credited with the emphasis on economics.

Unncessary plagiarism

Much of what is plagiarised is not really necessary for the thesis. Students sometimes plagiarise the bulk of their introductions and much of their literature reviews. Why do you need to define a GIS, anyway? Only if you will proceed from that definition to something specific is it necessary. And in any case you should phrase the definition in your own way, or quote (not plagiarise) an established definition.

Synthesis

In introductory material such as a literature review or the problem statement, it is common to make statements that are obviously not your own original ideas. If you have made a synthesis, that is, taken various ideas and facts and put them together to make your own argument or explanation, you have to give credit but you should not quote.

Quoting

It is almost always better to put things in your own words and argument rather than to quote. However, quoting is justified in these specific instances:

- Definitions that you will discuss;

> 'A common-language definition of land is "the solid part of the earth's surface" [15]. However, when we use the term 'land' in when defining 'land evaluation', we have in mind a more specific meaning, following the FAO [7], . . . '

> 'Bergsma [2] defines soil conservation as "the use of land, within the limits of economic practicability, according to its capabilities and the need to keep it permanently productive". Thus the emphasis is on economic sustainability.'

- Direct statements that you will discuss;

> 'Buol *et al.* [5] feel that there is widespread awareness of the existence of soils, calling them "objects of common experience and observation". We will argue that they are in fact not so widely perceived . . . '

- Especially clever or unique sayings, aphorisms, literary references that are particularly appropriate to what you want to say.

> 'As Yogi Berra[15] famously said, "You can observe a lot just by watching".'

[15] an American baseball player and folk hero well-known for his aphorisms

Paraphrasing

A particular difficulty comes with *paraphrasing*:

- You say the same thing as as a single author;
- You say it in the same order or with the same argumentation;
- You use quite similar words or synonyms.

This is also plagiarism, although certainly less egregious than out-and-out copying.

Example of plagiarism by paraphrasing

Consider this passage:

> "People seem to have a natural tendency and urge to sort out and classify the natural objects of their environment. Soils are no exception, being objects of common experience and observation – undergirding agricultural production and supporting buildings and highways"
> – Buol *et al.* [5, p. 180]

Here is a paraphrase that would certainly be considered plagiarism, *even if the citation is given*:

> 'As humans, we appear to have a built-in need to organise the things we find around us in the natural world. This is also true for soils, which everyone has seen, since soils are so necessary for agriculture and civil engineering [5, p. 180]'

Why is this still plagiarism? Because, although I have changed the words, the argument and sequence are the same. I have simply used synonyms and close paraphrases of the original:

Buol	**Paraphrase**
People	humans
natural tendency and urge	built-in need
to sort out and classify	to organise
natural objects	things . . . in the natural world
of their environment	we find around us
Soils are no exception	This is also true for soils
being objects of common experience and observation	which everyone has seen
undergirding	are so necessary for
agricultural production	agriculture
and supporting buildings and highways	civil engineering

Here is an acceptable compromise, where I give credit to the original authors and then extend their ideas with my own. I've also loosened up the style and argument. This is not plagiarism.

> 'Buol *et al.* [5] point out that, as humans, we seem to have a built-in need to organise the complexity of the natural world. Cognitive scientists such as Pinker [11] have even suggested that the desire to reduce complexity and form categories is 'hard-wired' into our brains by evolution. This tendency to classify extends to soils, at least to those properties that are readily perceived by soil users such as agriculturalists and engineers.'

4.3 Intellectual property and fair use

The intellectual, intangible product of a creative effort, such as writing, music, or a computer program, is as much the property of the creator as is a tangible object such as a work of art or a machine. In some cases intellectual property is put into the public domain for free use, in other cases its use is restricted.

Misuse of intellectual property is easier than misuse of tangible property, but it is equally theft.

4.3.1 Copyright

Copyright (indicated by the © symbol) is the means by which an author asserts ownership of a work. Laws vary between countries, and there are international treaties. The basic idea is very simple: *the work belongs to the author*, who grants you certain rights. If you obtain the work legally, you can use it for your own purposes (e.g. read it for pleasure or instruction). Other uses are made explicit, for example:

> "All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording or by any information storage or retrieval system, without the written permission of the publisher, except for brief passages quoted by a reviewer." [14]

4.3.2 The concept of 'fair use'

In science or art we may want to compare our work with that of others, and to make our point we need to quote directly. This sort of use is recognised by copyright law as *fair use*: you obtained the work and you may use it for your professional purposes.

We may also want to make photocopies of printed matter. If we want the whole book, we are required to buy it. If we only want 'reasonable' excerpts, it is considered fair use to make a copy of these parts.

4.3.3 License agreements

Some materials are made available only under the terms of a *license agreement*. That is, the person who obtains it in a legal manner, whether by purchase or free, must first agree how it is to be used. This is very common with *computer programs* and *digital data*.

For example, ITC has a license with certain publishers to allow on-line access to the full text of journal articles (usually as Adobe PDF files). The license allows full use within ITC, but it is *forbidden to supply a third party* with the file; they would need to obtain it under their own license.

! → You, your employer, or your educational institution (ITC) is *liable for your actions*. Remedies available to the copyright owner include expensive lawsuits and even criminal charges.

4.3.4 Restrictions on Datasets

Some digital data is supplied completely without restriction on what you can do with it, in particular data produced by the United States government. Most, however, is only supplied along with an *end-user license agreement* (EULA), to which you must agree.

You can not use data in your thesis which is not legally yours to use. This can be either via your own license, via ITC, or via some organisation of which you are considered part for licensing purposes.

4.3.5 Copyleft and open-source software

Some material is explicitly protected against theft, but its full use (including resale) is allowed if certain conditions are met. The most (in–)famous of these is the GNU General Public License (GPL) for certain open-source software[16], which requires that any new software that uses any code protected by GPL also itself be licensed under the GPL.

4.4 Professionalism

The term **professionalism** refers to scientists' role in the larger society in which they work . Many professional groups have codes of behaviour. These include ethical standards, but also deal with how the professional should behave and act within the society at large. These are sometimes called "codes of ethics" or "standards of professional conduct". They may have legal standing in some countries.

For example, the Soil Science Society of America [13] includes the following:

[16] http://www.gnu.org/copyleft/gpl.html

“Members [of the Society] shall:

1. Uphold the highest standards of scientific investigation and professional comportment, and an uncompromising commitment to the advancement of knowledge;
2. Honor the rights and accomplishments of others and properly credit the work and ideas of others;
3. Strive to avoid conflicts of interest;
4. Demonstrate social responsibility in scientific and professional practice, by considering whom their scientific and professional activities benefit, and whom they neglect;
5. Provide honest and impartial advice on subjects about which they are informed and qualified;
6. As mentors of the next generation of scientific and professional leaders, strive to instill these ethical standards in students at all educational levels.”

Point (2) was already covered under Ethics, but the others are socially-defined values. Notice in particular the ethical standards for consulting covered by points (3) and (5), There is also attention to the explicit social role of the professional; science is not value-neutral! In particular, point (4) means that the social implications of research must be considered (e.g., a technology that favours capital-intensive farming will have implications for the survival of family farming). Point (6) has to do with inter-generational transmission of values.

These aspects of ethics go far beyond simple considerations of honesty. They may well be defined differently in different societies. Here social and religious values are indeed relevant.

4.5 Codes of conduct

There may be conflicts between certain “universal” scientific values and the socio-cultural context; these difficult issues are often addressed by national scientific societies. For example, in the Netherlands the association of universities has published a “Code of Conduct for Scientific Practice” [16], which all researchers in the Netherlands must follow. This has sections on:

1. **Scrupulousness**: Scientific activities are performed diligently, with care, resisting pressure to cut corners in order to “achieve”;
2. **Reliability**: The scientist makes every effort for their work to be accurate and thorough, thus reliable;

3. **Verifiability**: Any publication based on research must clearly state the basis for the data and conclusions, including the data source and analysis methods; all of this so that the reader can in principle independently verify the work;
4. **Impartiality**: In scientific activities, the scientist must have no other interest than science, and be prepared to prove this. This is most relevant when the scientist works for industry or has commercial interests;
5. **Independence**: Scients operate in a context of academic freedom and independence from interference. If this is not possible for commercial, political or institutional reasons, this must be clearly stated and justified.

4.6 The social responsibility of the scientist

A wider ethical question than professionalism or a code of conduct is the role of the scientist in the wider social context – i.e. acting as a responsible member of society. This depends on the scientist's personal values and society's expectations.

An example is the **selection of a research topic**:

- Would the results of the research be **useful** to society?
- Is the topic related to a **social problem** of importance?
- Would the results of the research be **socially valuable**, or at least not damaging?
- Are various sectors of society **marginalized** or even directly **harmed** by the research?

Many research topics pose ethical problems, for example:

- Any remote-sensing project by its nature (view from above) invades the privacy of individual land owners; it also violates the sovereignity of the country imaged;
- Any natural resources survey or land suitability evaluation project implies that knowledge of these will be given to people outside the affected area, who may make investment or migration decisions that may not benefit the local population.
- A design thesis that builds on a specific computer program is implicitly endorsing that program and, if it is a commercial program, promoting the financial interests of the company that produced it (ESRI, Microsoft, ENVI . . .). Conversely, use of an open-source program reduces commercial opportunities. Which side are you on?

4.7 Social interactions of the scientist

Most scientists interact with colleagues, both in and out of their own institution. In any research which includes fieldwork, scientists also interact with local populations in the study area. These relations are a matter of ethics as well as professionalism.

Colleagues

Much of the interaction with **colleagues** is governed by narrowly defined scientific ethics as outlined above (§4.1), particularly the rules for assigning credit for work performed. However, there are often cultural differences (both general and scientific) in working methods, expectated roles and responsibilities, priorities, attitudes towards authorities, and communication style which can hinder scientific progress. Economic differences between colleagues can exacerbate these cultural differences. Awareness, sensitivity, communication, flexibility and common sense go a long way towards achieving a good working relationship.

Local populations

Professional societies whose members do research with and in **local populations** have had extensive debate about the relation between researcher and subject; an example is by the American Anthropological Association [1]. This is also dealt with in texts on social science research methods [e.g. 8, 9] and in the context of recent research on participatory GIS [12].

Here are some examples of ethical questions raised by such research:

- How should they be approached? What information about the research purpose should be given?
- Will the results of the research be 'returned', and if so, in what form?
- What to do if the research is not in their benefit, or even to their detriment? Example: studying soil erosion vs. farming practices, this may lead to a ban on certain crops or management on certain lands (e.g. steep slopes), which is a short-term economic loss to the farmers?
- If surveys are to be performed, what information about them is given to the participants? Should they be paid or otherwise rewarded?
- What are ethical methods of asking questions or making observations? Can subjects be "tricked" with false promises or pretexts?
- How intimate should the researcher be with the population? Does the researcher sacrifice neutrality or objectivity by identifying too closely with the subjects or target group?

- How should researchers balance their own cultural values with those of their subjects?
- How to extract reliable information within cultural limitations? Example: It is considered improper in the local context for a male researcher to talk directly with a female subject; should the researcher trust a male relative's interepretation of what the female says?

Research is never neutral – someone (maybe the researcher?) benefits more than others; the results may be used for political ends, and so forth. Researchers are themselves always biased and have their own cultural references. These must be made explicit, at least to the researcher.

References

[1] American Anthropological Association. 1998. *Code of ethics*. Arlington, VA: American Anthropological Association. URL: http://www.aaanet.org/committees/ethics/ethcode.htm

[2] Bergsma, E. 1996. *Terminology for soil erosion and conservation*. Wageningen; Enschede: International Society of Soil Science; International Soil Reference and Information Centre (ISRIC); ITC. ITC 631.459 REFERENCE

[3] Booth, W. C.; Colomb, G. G.; & Williams, J. M. 1995. *The craft of research*. Chicago: University of Chicago Press. ITC 001.818

[4] Broad, W. & Wade, N. 1982. *Betrayers of the truth*. New York: Simon and Schuster

[5] Buol, S. W.; Hole, F. D.; & McCracken, R. J. 1989. *Soil genesis and classification*. Ames, IA: The Iowa State University Press, 3rd edition

[6] Cornell University. 2000. *The code of academic integrity and acknowledging the work of others*. Ithaca, NY. URL: http://web.cornell.edu/UniversityFaculty/docs/AI.Acknow.pdf

[7] FAO. 1976. *A framework for land evaluation*. Soils Bulletin 32. Rome, Italy: FAO

[8] Frankfort-Nachmias, C. & Nachmias, D. 1996. *Research Methods in the Social Sciences*. New York: St. Martins Press, 5th edition

[9] Frankfort-Nachmias, C. & Nachmias, D. 2000. *Research methods in the social sciences*. New York: Worth Publishers, 6th edition

[10] Hudson, N. 1981. *Soil conservation*. Ithaca, NY: Cornell University Press, 2nd edition

[11] Pinker, S. 1997. *How the mind works*. New York: W. W. Norton

[12] Rambaldi, G.; Chambers, R.; McCall, M.; & Fox, J. 2006. *Practical ethics for PGIS practitioners, facilitators, technology intermediaries and researchers*. *Participatory Learning and Action* **54**:106–113

[13] Soil Science Society of America. 1999. *Statement of ethics*. On-line HTML document; URL: http://www.soils.org/pdf/statement_of_ethics.pdf. Access date: 27-December-2004

[14] Strang, G. 1986. *Introduction to applied mathematics*. Wellesley, MA: Wellesley-Cambridge Press

[15] Thompson, D. (ed.). 1995. *The concise Oxford dictionary of current English*. Oxford: Oxford University Press, 7th edition

[16] VSNU. 2004. *The Netherlands Code of Conduct for Scientific Practice*. The Hague: VSNU. URL: http://www.vsnu.nl/

5 Literature review

Key chapter points

1. The main purpose of a literature review for MSc research is to establish its **originality** and to put the proposed research in **context** (§5.1).
2. The literature review also justifies **choice of research methods**.
3. Citations are used to present **definitions and concepts**; **opinions** of others; details of **methods**; and **facts** which you did not yourself establish; and **quotations**.
4. Sources have different degrees of **reliability**, peer-reviewed journal articles (research and review); conference papers; book chapters, textbooks, technical reports, and web pages (§5.3).
5. The **list of references** in a proposal, thesis or article must contain every reference in the text, and vice-versa (§5.8).
6. Every item in the list of references should be easy for a competent librarian to find (§5.8).

This chapter explains some aspects of the literature review and list of references included in the thesis, including:

- Purpose of a literature review
- Why citations are used in a thesis or scientific paper;
- Different types of references and their reliability;
- Citing electronic sources;
- Different levels of sources;
- Some common problems in citations;
- Bibliographic style and use of EndNote; and
- Starting points for effective literature search.

5.1 Purpose of a literature review

The main purpose within the context of MSc research is to establish its **originality**; that is, that the work proposed has not already been done. Almost always something related has been done; the review organises these, discusses them, and points out their limitations, some of which will be addressed in the research.

A second purpose is to place the proposed research in **context**, that is, to show its importance within a wider problem area. This must be established from the opinions of others, who define the context and identify important unsolved problems.

A third purpose is to compare **methodological approaches** to your problem. There are almost several ways to address a research problem, and here you compare these approaches and justify your own approach (which may combine aspects of the others).

5.2 Purpose of citations

Science is a collective enterprise, with a history and a future. No one person can do all the work nor think up all the good ideas. The greatest scientists of all time, such as Newton and Gauss, explicitly acknowledged their intellectual debt to their predecessors. Furthermore, it would be impossible for one person to do all the experiments, collect all the primary data, or build all the information systems that have already been done by your colleagues.

Fortunately, science requires that we write down what we find and what we think. In a review of the literature we follow this historical trail, thereby saving us from having to duplicate previous work, and giving us the best possible basis for our own plans. It also saves a lot of writing, since you can just cite conclusions, and leave the detailed explanation for the original source.

In a thesis or other scientific writing, literature citations serve several purposes:

Definitions & concepts

1. They **present definitions and concepts** that are not yours, and **give proper credit** for them.

   > 'Heuvelink [18] distinguishes two major conceptual models of soil spatial variability: the Discrete (DMSV) and Continuous (CMSV). The DMSV hypothesises that the variation in soil classes and properties across the landscape can be partitioned by sharp boundaries into homogeneous areas, whereas the CMSV . . . '

   > 'There are two major conceptual models of soil spatial variability: the Discrete (DMSV) and Continuous (CMSV) [18]. The DMSV hypothesises that the variation in soil classes and properties across the landscape can be partitioned by sharp boundaries into homogeneous areas, whereas the CMSV . . . '

Of these forms, the first is more explicit that Heuvelink actually invented these terms; the second form might just mean that Heuvelink's article is a good review of the concepts.

> 'Sahay & Woolshan [30], reporting on the implementation of GIS in a USAID-sponsored project in India, distinguished between what they termed "inhibiting" and "enabling" factors. '

Opinions

2. They **present opinions** that are not yours, **give proper credit** for them, and allow the reader to **verify your interpretations** of these works (i.e., the reader can go back to the original source and check if you correctly summarised it).

> 'According to McBratney *et al.* [26], pedometric techniques are the future of soil survey *(Is that really a fair summary of that article?)*'

Data & results

3. They **substantiate data and results** that are not from your own research, and allow the reader to find the original source if necessary.

> 'The Hungarian Environmental Monitoring System is a point–vector database congaining 1236 soil profile descriptions [8].'

> 'About 48% of Africans, mainly in the centre and south, profess some form of Christianity, while about 41%, mainly in the north and Sahel, are Muslims [4, article "Afrika"]'

Note that the specific article in this reference work is mentioned in the citation; this is not strictly necessary, but it helps the reader find the information to verify it or see its context. The fact in this example could, for example, have been in an article about world religions; instead it happened to be in the article about Africa.

Introduction

4. They **refer to previous work on your topic**, which you use in your introduction to motivate your study and place it in context:

> 'The first systematic study of soil map quality was by Webster & Beckett [36]. Somewhat later, a group at Cornell University worked for several years on aspects of soil survey adequacy, including accuracy assessment [12, 32]. At this same time, group at the Staring Centre in the Netherlands developed methods for quantifying map unit composition and thematic quality [17, 25].'

Methods

5. They refer to **standard methods**, so that you don't have to repeat

them in your text. This is common in your ‘Methods’ chapter.

> ‘The particle size distribution was determined by the pipette method with pre-treatment for organic matter but not for carbonates [28].’

Formulas

6. They provide **detailed justification** of mathematical or statistical methods, so you don’t have to derive or defend them:

   > ‘A formula for computing the variance of the *kappa* map accuracy statistic was derived by Bishop *et al.* [3, §11.4.2]: $\sigma^2[\hat{k}]$ is computed as . . . ’

   Note that I mention the section in this long book where this particular formula is derived. This is not strictly necessary, but may be a great aid to the reader in finding and verifying your interpretation. If the index of the book provides an easy way to find this (here, if there is an entry for ‘*kappa*, variance of’), it would not be necessary to mention the section here.

Results

7. They **refer to other studies related to your results**, with which you should compare, in your ‘Results’ chapter.

   > ‘This result appears to contradict that of Webster & Beckett [36], who found that only 10% of the area was unsuccessfully mapped.’

   > ‘This successful clustering of the profiles by principal components analysis matches the results of Gobin *et al.* [14], who found that the first two PCs explained 64.7% of the total variance in a set of 72 pedons in southeastern Nigeria’

Further reading

8. They give the reader material to go **deeper into a topic** than was necessary for your purposes. This is not needed for your work, but can be useful to some of your readers.

   > ‘A formal development of the theory of spatial operations on cell complexes is given by Kainz [20].’

5.2.1 Citations must have been read by the author

! → In general, **only cite material you have actually seen.**

Otherwise you can not be sure that it says what you are asserting that it does, or even that it really exists. *You are relying on someone else’s interpretation* of what it says, which may well be wrong. You can not defend any interpretation of the material, since you haven’t read it yourself .

! → **When you cite something, you are implicitly representing that you have read it.**

The main exception to this rule is if the existence of the cited work is itself relevant to your study; for example, if you are writing a historical survey and need to refer to all works on a subject, even if you haven't been able to find it yourself. Another exception is if you can find the work but can not read its language.

See §5.6 for a solution if you absolutely must cite something you haven't seen.

5.2.2 When **not** to use a reference

Not everything you say needs to be supported by a reference.

1. If it's your idea or result (then your report is the reference others will use);

   > 'On closer observation, it was obvious that the water samples all contained insect larvae ...'

2. If the fact is known to any person with a basic education; this holds especially for general statements that will be developed further by argument.

   > 'France and Germany have long vied for European supremacy ...'

3. If the fact can be found in a standard secondary-school or general reference;

   > 'Since the area A of a circle is πr^2, we can compute ...'

4. If the fact is more or less fixed and can be verified in many ways;

   > 'Cuba is a Caribbean nation ...'

5.3 Types of sources

Not all sources are equally valid! At one extreme, anyone can place any opinion on the Internet, with no control. At the other extreme is a peer–reviewed paper in a highly competitive international scientific journal.

This section lists the principal types of sources for published scientific information, with some comments on their reliability.

5.3.1 Journal Article

This is an original contribution that appears in a published scientific journal.

These contributions have been **peer–reviewed** to ensure quality control (§5.4). However, not all peer–review is equally effective. In general, the more influential the journal (i.e., the more its work is cited and considered of top quality), the more likely that peer review has been rigorous. You are more likely to find reliable information in Nature than in some regional journal of development studies.

However, you should not take this argument from authority too far. Nothing done by man is free from the possibility of error, illogical thinking, or outright fraud.

There are several types of articles which may appear in a journal. Here I give examples from peer–reviewed, internationally-circulated soil science journals.

Research Article Describes an original investigation, method, or procedure. Specific and limited. Examples are Dobos *et al.* [8] and King *et al.* [21]:

- ▷ Dobos, E.; Micheli, E.; Baumgardner, M. F.; Biehl, L.; & Helt, T. 2000. *Use of combined digital elevation model and satellite radiometric data for regional soil mapping*. *Geoderma* **97**(3-4):367–391
- ▷ King, D.; Bourennane, H.; Isambert, M.; & Macaire, J. 1999. *Relationship of the presence of a non-calcareous clay–loam horizon to DEM attributes in a gently sloping area*. *Geoderma* **89**(1-2):95–111

Review Article Summarises a set of research articles; surveying the state-of-art in a particular field. The title typically includes words like “review”, “summary”, or “overview”. Here the originality lies in the synthesis, not the investigation. Examples are by McBratney *et al.* [26] and Goovaerts [15]:

- ▷ McBratney, A. B.; Odeh, I. O. A.; Bishop, T. F. A.; Dunbar, M. S.; & Shatar, T. M. 2000. *An overview of pedometric techniques for use in soil survey*. *Geoderma* **97**(3-4):293–327
- ▷ Goovaerts, P. 1999. *Geostatistics in soil science: state-of-the-art and perspectives*. *Geoderma* **89**(1-2):1–45

Opinion A scientific editorial, either by the journal editor or an invited contributor. An example is by Basher [2]:

- ▷ Basher, L. R. 1997. *Is pedology dead and buried?* *Australian Journal of Soil Research* **35**:979–94

5.3.2 Conference Paper

This is an original contribution that was presented at a scientific meeting. In most cases these are **not** peer–reviewed. The conference organisers typically allow anyone who pays the registration fee to present whatever they want. In fact, the reason that scientists presents a paper at a conference is to inform their peers of their work, especially their new results and ideas which may not be fully "cooked" yet, and to get feedback. So it is correct to present work that could not pass the peer review process. Think of the conference as a bazaar.

Conferences may publish their papers in several forms, in increasing level of reliability:

Conference Proceedings The original submissions, with no quality control; typically distributed at the meeting itself. "Published" by the conference organisers. Almost impossible to obtain after the fact. Avoid using this as a source if at all possible.

Edited Proceedings A book from a publisher with some of the submissions, at least reviewed by a scientific editor to eliminate obviously wrong papers. An example is by de Gruijter & Marsman [17]:

- ▷ de Gruijter, J. J. & Marsman, B. A. 1984. *Transect sampling for reliable information on mapping units.* In Nielson, D. R. & Bouma, J. (eds.), *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA*, pp. 150–163. Las Vegas: PUDOC, Wageningen

Special issue of a journal Selected papers are sent for peer review; these should be considered journal papers for the purposes of citation and literature search, even if they were first presented at a conference. They have an editor and may be cited as a whole; in this sense they are like an edited book. An example of a special issue is by de Gruijter [16]; a contribution from this issue is by King *et al.* [21]. This work was originally presented at the Pedometrics '97 conference in Montpellier, but subsequently revised to a journal article. The entire issue may be cited, as may individual articles.

- ▷ de Gruijter, J. 1999. *Special issue: Pedometrics '97. Geoderma* **89**(1-2):1–400
- ▷ King, D.; Bourennane, H.; Isambert, M.; & Macaire, J. 1999. *Relationship of the presence of a non-calcareous clay–loam horizon to DEM attributes in a gently sloping area. Geoderma* **89**(1-2):95–111

5.3.3 Book chapter

This is an original contribution that is collected into an edited book on a specific topic.

These are typically invited by the book editor and may undergo some peer review; certainly they are edited. Often they are review articles or summaries. Quality control is not as rigid as for journal articles. An example of such a book chapter is that of Skidmore [31]. The chapter is by Skidmore, an authority on the subject he was asked to review, and the book is edited by a group of well-known scientists. Still, this was not peer–reviewed in the same way as a journal article.

▷ Skidmore, A. K. 1999. *Accuracy assessment of spatial information*. In Stein, A.; Meer, F. v. d.; & Gorte, B. G. F. (eds.), *Spatial statistics for remote sensing*, pp. 197–209. Dordrecht: Kluwer Academic.

The entire book can also be cited if you want to make a summary statement about its contents:

▷ Stein, A.; Meer, F. v. d.; & Gorte, B. G. F. (eds.). 1999. *Spatial statistics for remote sensing*. Dordrecht: Kluwer Academic

5.3.4 Textbook

This is a published book meant to introduce a subject for classroom teaching or self-study. It can treat a topic at any level (i.e. pre-requisites for understanding it), but given that level, it is intended as the first contact with the subject.

These are not peer–reviewed as such but are typically extensively edited and sent by the publisher to people who might use the text in teaching, to see if they find the book accurate and useful. Beware, not all publishers do this. The reputation of a publisher is important here. Among the good ones are Wiley, Springer, McGraw-Hill, Addison-Wesley, Oxford University Press. Others may be sloppier.

An example of an elementary text is by Lillesand & Kiefer [23], of an advanced text by Bishop *et al.* [3].

▷ Lillesand, T. M. & Kiefer, R. W. 1994. *Remote sensing and image interpretation*. New York: John Wiley & Sons, 3rd edition

▷ Bishop, Y.; Fienberg, S.; & Holland, P. 1975. *Discrete multivariate analysis: theory and practice*. Cambridge, MA: MIT Press

5.3.5 Technical Report

These are publications from an institution or project, and often contain primary data and maps which do not appear elsewhere. They are often difficult to obtain, but if they are the only source of information, they should be cited. They are not peer–reviewed; the quality control was only as good as the project.

Examples are:

- ▷ Center for Advanced Spatial Technologies (CAST). 1998. *AR-GAP final report: State-wide biodiversity mapping for Arkansas.* Report, Center For Advanced Spatial Technologies (CAST), Fayetteville, AR.
- ▷ Anonymous. 1985. *Soils and soil conditions, Kali Konto upper watershed, East Java.* Project Report ATA 206, Universitas Brawijaya (Malang), Agricultural University (Wageningen)

Note that the Kali Konto project was completed in 1984 (this data appears on the cover) but not published until August 1985 (this date appears in the publication information inside the report). Not only is no author given, but the project name isn't even given, so we have no alternative but to list the author as 'Anonymous'.

5.3.6 Electronic sources

These are materials that are available in digital (computer-readable) form[17]. We distinguish first between the on-line and off-line cases:

On-line These are only available via the web. These pose the difficulty that they are not permanent: tomorrow's version may be different from today's, it may move to another cyber-address, or it may even disappear.

Off-line These are in electronic form, but do not rely on the web for access. Typically they are CD-ROMs, often published by book publishers. They are cited like books:

- ▷ FAO. 1998. *Digital Soil Map of the World and derived soil properties (CD-ROM).* Land and Water Digital Media Series No. 1. Rome: FAO

Then we distinguish between the case where the electronic source is just another form of a published source, or where it is the primary information:

Alternative source In this case, the electronic source is simply a copy or a differently-formatted version of a printed source. The electronic alternative is easier to access, so this information can be mentioned in the corresponding entry in the reference list. The electronic source has the same level of peer-review as the printed version.

[17] 'Electronic' is quite a misnomer but we are stuck with it.

Here is an example (also used above) published technical report that has been formatted for the Web (both HTML and PDF).

> 'In America, the National Gap Analysis Program (GAP) was carried out during the 1990's, in order to find a common language for discussing issues such as land cover mapping, vertebrate habitat characterisation, and biodiversity conservation. Most studies were carried out at state level, for example in Arkansas [7].'

▷ Center for Advanced Spatial Technologies (CAST). 1998. *AR-GAP final report: State-wide biodiversity mapping for Arkansas.* Report, Center For Advanced Spatial Technologies (CAST), Fayetteville, AR. URL: `http://www.cast.uark.edu/gap/`

Primary source

Other electronic sources have no printed equivalent. In this case, **you must include the *access date***, that is, when you actually viewed the content. This ensures that the information was at the given address; if it later is missing, it may be possible to find the version from the date in a search site's archives[18].

! →

Unless you have specific information to the contrary, **primary web sources are not peer-reviewed.** On the web you can easily find outright forgeries and lies (e.g. the so-called "Protocols of the Learned Elders of Zion"), highly improbable statements with no credible evidence (e.g. creationist web pages that state that there is no evidence for evolution or an Earth older than 6011 years (as of 2007), or that psychic energy from crystals can cure cancer, let alone poor science, incorrect history, and uncritical analysis. The Web is a global Hyde Park Corner[19], a wonderfully democratic medium where anyone can make a fool of themselves to a world-wide audience. So in the same way you would be critical of someone making a speech on a street corner, be critical of what you read on the Web.

You **must also include the *date the source was last modified***; this is equivalent to the edition for a book.

Here are examples of references to a web page ...

> 'In many American states, soil conservation is aggressively promoted through attractive web sites [e.g. 34].'

[18] You might want to make your own local copy for backup.
[19] a place in London where speakers harangue passers-by with any and all opinions

▷ USDA Natural Resources Conservation Service. USDA Natural Resources Conservation Service. 2000. *Welcome to the USDA Natural Resources Conservation Service home page for North Carolina*. On-line document; URL: http://www.nc.nrcs.usda.gov/. Access date: 02-May-2001

. . . and to a PDF file:

> 'A good introduction to ethnopedology is by Ettema [9].'

▷ Ettema, C. H. 1994. *Indigenous soil classifications*. On-line PDF document; URL: http://www.itc.nl/personal/rossiter/Docs/Misc/IntroToEthnopedology.pdf. Access date: 16-May-2002

Finding a web page's title and author

For Web pages, you should use the page's *title* as given in the HTML `<title>` tag. This is shown as the window title in most browsers. The *author* is the organisation or individual who sponsors the page; layout alone is not considered scientific authorship. You may have to go up one or several levels in the hierarchy, possibly to the organisation's home page, to find out who actually sponsors the page.

5.4 Peer review for quality control

The peer review process attempts to ensure that what is published is reliable and important. The approval of the authors' peers means that someone who is not intimately familiar with the research being discussed (e.g. you as MSc candidates) can pretty much trust that the publication is methodologically-correct and honestly-performed. The conclusions are another matter; here you should still form your own opinion from the body of the paper.

Here's how it works:

1. The authors submit a draft of the article to a journal editor.
2. The editor checks that the subject matter is relevant for the journal, and that the paper meets the required format (length, figures, required sections etc.).
3. The editor sends the draft to several other scientists familiar with the subject matter. They read the draft and advise the editor on what to do with the paper:
   (a) **Accept in present form.**

   If the article is accepted, it is typeset, sent to the authors for proofreading (not changes), and published in the journal.

(b) **Accept with minor revisions** (specified by reviewers). Suggested revisions are typically of format or style and not of substance; the reviewer finds the work as such correct:

- Better editing, revise language slightly, improve the English;
- Re-consider a specific statement which the reviewer does not consider justified;
- Reformat tables or graphics;
- Explain something more clearly or illustrate with an example.

If the article is accepted subject to minor revisions, the authors make the revisions, and re-submit to the editor, who checks that the revisions match what the reviewers recommended. If so, the paper is published. If the authors disagrees with the suggestions, they can argue the point with the journal editor, or withdraw the article.

(c) **Reconsider if major changes are made.** Recommendation to revise and possibly re-submit can also be for several reasons:

- Incorrect analysis, not suitable to the data (therefore the conclusions are not justified);
- Unjustified conclusions, poor reasoning;
- Work does not properly consider related work, i.e. does not compare its results to others.
- Poor writing.

If major revision is recommended, the authors are given the paper back (i.e. the journal releases potential copyright) and they have a chance to amend the paper and re-submit it here or elsewhere.

(d) **Reject.** Outright rejection can be for several reasons:

- Incorrect data collection or processing methods (therefore the data are not reliable);
- Work repeats what has already been done, nothing new is added to the existing literature;
- Work is too narrow ("light") to justify publishing, but could be incorporated into a bigger study;
- Work is not relevant for the proposed journal;

If the article is rejected, it may not be re-submitted.

5.5 Choosing among sources

Not all sources are equally useful to you or your readers. The references should be **relevant**, **reliable**, and **accessible**,

1. Only cite material that directly bears on your work (i.e., is **relevant**). It should fit one of the categories listed above (§5.2). Superfluous references do not impress, they confuse.
2. The order of **reliability** is approximately:
   (a) Peer–reviewed articles in international journals;
   (b) Book chapters in edited collections; Textbooks by well–regarded authors (i.e. with a publication record in peer–reviewed journals); Edited conference proceedings from international congresses;
   (c) Technical reports and electronic documents with no printed equivalent from well–regarded institutions;
   (d) Peer–reviewed articles in national or regional journals; Textbooks by lesser authors;
   (e) Unedited conference proceedings; Edited conference proceedings from local congresses;
   (f) Technical reports and electronic documents with no printed equivalent from unknown institutions.

   In all cases, use your common sense and natural scepticism; beware the argument from authority, but recognise its utility.
3. Cite the most **accessible** source among several that give similar information:
   - easy to find in many libraries, or at least easily obtainable by inter-library loan;
   - written as clearly as possible;
   - in English;
   - the most recent synthesis, rather than an isolated report.

   For example, a thesis that has later been turned into a book or article(s) in major journals is much easier for a reader in any country to find in these sources. Another example is an early study in a technical report or a minor journal that then is included in a synthesis (review paper or textbook).

If a work has been presented both in a national language and then in English, use the English version, unless the original has more information, in which case cite both.

4. For **primary** information, cite the **primary** source rather than an interpretation or summary, even if it is in a less accessible or reliable source. For example, the original census rather than an article on the results of that census. Sometimes both should be cited, since the summary may be more accessible.

5. An ITC thesis should be cited if it contains the primary data and analysis, even if much of this was later included in an article; these theses are obtainable by anyone on request to the ITC library.

5.6 Citing material you haven't or can't read

If absolutely necessary, use the 'cited in' approach: use the original author, but the bibliographic reference is to the book you actually saw. There are two reasons for doing this:

- You can't obtain the original source; or
- You can't read the original language.

But, you need to cite the fact, opinion, or data from the source.

Example 1: You can't obtain the original source, but you have an abstract This may be the case for conference proceedings, and works in minor journals, out-of-print books and reports. The abstract may be in a special abstracting publication (e.g. CAB Abstracts) or in an electronic database (e.g. GEOBASE). So, you can see the main conclusions of the work, and you want to cite it, but you can't see the full paper. In this case, place the notation "(Abstract)", with the name of the abstracting service, at the end of the reference:

▷ Oliver, M. A.; Webster, R.; & Slocum, K. 2000. *Filtering SPOT imagery by kriging analysis. International Journal of Remote Sensing* **21**(4):735–752. (GEOBASE Abstract)

This also applies if you can only read the English-language abstract of a paper written in a language you do not understand. In this case, place the notation "(English Abstract)", for example (assuming you can't read French):)

▷ Gaultier, J. P.; Legros, J. P.; Bornand, M.; King, D.; Favrot, J. C.; & Hardy, R. 1993. *L'organisation et la gestion des données pédologiques spatialisées: Le projet DONESOL (English abstract). Revue de Géomatique* **3**(3):235–253

If possible, give the title in English with the name of the original language in parentheses. This assumes that someone has translated the title for you, or that it appears translated in the source. Journal and book names are not translated, because they are searchable by librarians.

▷ Gaultier, J. P.; Legros, J. P.; Bornand, M.; King, D.; Favrot, J. C.; & Hardy, R. 1993. *Organisation and management of soil data: the DONESOL project (in French) (English abstract)*. *Revue de Géomatique* **3**(3):235–253

Example 2: You can't obtain the original source or its abstract If the information in the original source still needs to be cited (typically so that the reader could find it), use the "Cited In" approach:

"Cited In"

> 'The Hungarian Environmental Monitoring System has been collecting detailed information since 1995 [35, cited in 8].'

▷ Várallyay, G.; Hartyáni, M.; Marth, P.; Molnár, E.; Podmaniczky, G.; I., S.; & Kele, G. 1995. *Talajvédelmi információs és monitoring rendszer. 1 kötet módszertan*. Technical report, Földmüvelésügyi Minisztérium, Budapest. (In Hungarian)

▷ Dobos, E.; Micheli, E.; Baumgardner, M. F.; Biehl, L.; & Helt, T. 2000. *Use of combined digital elevation model and satellite radiometric data for regional soil mapping*. *Geoderma* **97**(3-4):367–391

Here Várallyay *et al.* [35] has the information; I can't get this technical report, so I have to rely on the account in Dobos *et al.* [8], which is where I found out about this system. Also note I do not give the English title for Várallyay *et al.* [35], since the whole work is in Hungarian, and it has no English abstract (as far as I know).

> 'In the South, many plantations faced problems of steadily-decreasing yields and economic ruin. Observant agriculturalists realised that different soils could sustain different levels of production, and recommended systematic soil surveys [29, cited in 5, p. 199].'

▷ Ruffin, E. 1832. *An essay on calcareous manures*

▷ Buol, S. W.; Hole, F. D.; & McCracken, R. J. 1989. *Soil genesis and classification*. Ames, IA: The Iowa State University Press, 3rd edition

Here I should cite Ruffin [29], even though I can't find this old book, because I am referring to a specific historical situation. If I simply want to say that different soils sustain different levels of production, I should cite a modern textbook or reference work on soil fertility.

Example 3: You can't read the language

> 'Kubiëna [22, cited in 5] was the first taxonomist to make the fundamental distinction between *terrestrial* and *aquatic* soils.'

▷ Kubiëna, W. L. 1948. *Entwicklungslehre des Bodens*. Wien: Springer-Verlag

Here we cite Kubiëna [22] because we are going to use this important distinction, which he proposed. Even if I have seen the book, if I can't read German, I must rely on the account in Buol *et al.* [5]. But unless your aim is to give a historical bibliography, you could just rely on the secondary source to establish the fact:

> 'Kubiëna was the first taxonomist to make the fundamental distinction between *terrestrial* and *aquatic* soils [5].'

Example 4: You can read the language, but the reader may not be able to If I can read German, but perhaps my reader can not, I should cite the original (since I myself read it) and a more accessible source (for my reader):

> 'Kubiëna [22] was the first taxonomist to make the fundamental distinction between *terrestrial* and *aquatic* soils [see also 5, p. 199].'

5.7 Miscellaneous citation problems

5.7.1 Corporate vs. Individual Authorship

This is sometimes difficult to determine, especially for project reports, technical manuals, or reference works. The basic rule is to credit an individual author or editor (or, several individuals) if it is clear from the work that they are largely responsible for its contents, even if the work is sponsored or published by an organisation. Simply compiling a group of papers, or summarising a discussion or project, is not enough for authorship.

In the case of the *Oxford Advanced Learners Dictionary* [19], Hornby is considered to be the overall editor, who was responsible for ensuring a consistent style, therefore the work is cited under his individual name.

In the case of the Brockhaus encyclopedia [4], there is an editor listed inside, but he is only named as the 'editorial leader', which sounds more like a co-ordinator than an editor, therefore the work is cited under the organisation.

In some cases, no author is given, even if one author wrote the work, because it was done under contract. An example is FAO Forestry Paper 48 [10], written by Anthony Young, although he is nowhere mentioned in the report. Then we have no choice but to cite the work under the organisation.

A common case is a project report with tens or scores of contributors or participants in a workshop. If there is an editor listed, the work is cited under the individual name as editor. If there is just a long list of contributors, the work is cited under the organisation.

5.7.2 Multiple sources for the same fact

Sometimes you have looked at several sources, all of which support a synthesis that you want to make. If you don't refer to the individual contributions for other reasons, cite all of the sources in one list, using language that makes it clear that you are referring to all of them. In the following example, I am referring to all of the works together:

> 'Soil map quality has been studied by several groups over the last thirty years [12, 17, 25, 32, 36]. The general conclusion is that we are a long way from making routine assessments of quality; indeed there is no agreement on the concept of 'quality' when applied to soil maps.'

5.8 Bibliographic style and the list of references

The **list of references** is an appendix to your thesis where you list those items that you actually refer to in the text. This is different from a **bibliography**, which is a (usually categorised) list of all sources consulted (whether cited or not), or a comprehensive list of sources relevant to some field. In the ITC thesis, it is the list of references that is required, unless one of the purposes of the study is to compile a bibliography.

Some simple, common-sense, rules are all that you really need to cite correctly:

! → 1. **Every reference in the text must appear in the reference list**.

! → 2. **Every reference in the reference list must appear in the text**.

! → 3. Items in the list of references should be **easy for a competent librarian to find**,

   - This implies that there can be **no personal communications or unpublished materials**. If these most be used, they are cited in the main text.

· This requirement implies what must be included in the reference as a *minimum*; other information may be included as a convenience for the reader. For example:

Journal article	Journal name or standard abbreviation; journal volume number; first page number
Book	Author; title; publisher; city

· Other information is sometimes added for the reader's convenience, in particular, the title of a journal paper and its full page range.

Citation style

4. **Any consistent style is acceptable.** Many good sources are available; especially recommended is the Council of Biology Editors' Style Manual [6] available in the ITC library. This uses an *author-date* system. The *APA-Published* style in EndNote is a reasonable implementation of this. *Numbered* styles take up less space in the text and do not distract the reader; you can still use the author's name in the text as a proper name if appropriate. This is the style used in these notes.

EndNote [24] takes care of (1), (2), and (4); consistent entry of items into EndNote, with correct information in important fields, takes care of (3). A consistent entry in EndNote can be re-formatted in hundreds of styles.

You can fairly easily create or modify EndNote's styles. There is no ITC house style.

There are some style manuals available on-line, for example from the American Society of Agronomy *et al.* [1].

5.9 EndNote tips

1. Make sure you select the correct *reference type* for each new reference. Some choices are: journal article, book, book chapter, and electronic source. You will notice that the fields in the record are different for each type.
2. Don't trust any *import filter*. Review every downloaded record to make sure the reference type is correct and that information is in the correct fields.
3. **Review at the formatted references**, i.e. read your own list of references! If something seems out of place, it may be that the output style you are using doesn't use a certain field, or it may be that you have the wrong reference type for the reference, or it may be that the fields are not correctly filled in.

4. If you use the automatic link from EndNote to Word, never make a change in the formatted Word document; always change the EndNote reference or style. This ensures that all items of the same type will have the same format.

Corporate authors

5. Enter corporate authors with a single comma after the complete name; EndNote will not try to abbreviate it:

   - "National Resources Conservation Service,"

   If you forget the final ",", in some styles EndNote will present this as *Service, N.R.C.* which is certainly not what you want.

6. Check the author names in your EndNote entry; there should be one per line, with no punctuation at the ends of lines. Using the format "LastName, First I." for "western" authors is most reliable.

Non-European names

7. If the name does not follow the European convention of personal and family names, write the entire name *with a comma after it*; EndNote will not try to force it into a European format and will never try to abbreviate it. Examples are Chinese, Vietnamese, Indonesian and Ethiopian names:

   - "Gao Yan,"
   - "Nguyen Thi Thu Ha,"
   - "Muhibuddin Bin Usamah,"
   - "Atkilt Girma,"

   Sometimes an author from one of these groups will "westernize" their name, especially if they follow a career outside their country; it can be difficult to recognize this if you don't know the language. For example, the Chinese author listed above may follow the European convention and write the family name (*Gao*) second, as "Yan Gao"; you have to know that "Gao" is a family name and "Yan" a given name to recognize this shift. In this case the entry in EndNote would be "Last Name, First Name":

   - "Gao, Yan"

   which EndNote could abbreviate in some styles as "Gao, Y.".

8. Spanish surnames are written three ways depending on the degree of formality, and you may encounter any of them:

   - "José Antonio Navarrete Pacheco": formal, with both father's family name (*Navarrete*) and mother's family name (*Pacheco*); this will be used in formal documents and theses; enter in EndNote as "Navarrete Pacheco, José Antonio".

- "José Antonio Navarrete P.": less formal, with father's family name written out and mother's family name abbreviated; enter in EndNote as "Navarrete P., José Antonio".
- "José Antonio Navarrete": every-day usage, with only father's family name; this may also be used if the Spanish-surname author publishes in a "western" journal; enter in EndNote as "Navarrete, José Antonio", which EndNote could abbreviate in some styles as "Navarette, J. A.".

5.10 How to Search

Finding relevant material, and especially the most important for your purpose, is not easy. It requires patience, detective skills, some luck, and continued hard work. Fortunately, in the electronic age it is possible to make much more rapid progress than previously.

Starting points include:

- Reference lists in lecture notes;
- Reference lists in earlier theses;
- Reference lists in textbooks;
- Review papers; these have the advantage that the literature is placed in context, and you can already have an idea of which references are most important;
- Keyword searches in electronic resources, both on- and off-line;
- Recent issues of relevant journals. This can be intimidating because the articles tend to be specialised, but if the topic is interesting to you, you can often find more basic references in the article's Introduction.

Once you have found some relevant literature, you can often go further by looking for:

- Works by the same author(s); very often the author continues working on related problems.
- Papers in the same journal or conference proceedings; a given journal tends to group papers that cover related areas.

If the work is relatively old, make sure to look at newer sources to see if it has been superseded, challenged, or revised.

5.11 Effective use of on-line search

The ITC library has excellent on-line access to a variety of sources; these are collected on the "Digital Library" web page[20].

Several scientific publishers maintain large on-line databases of abstracts and full-text articles. ITC has access to Elsevier's Science Direct[21], Blackwell's Synergy[22], and SpringerLink[23], among others. There are also several independent databases; most notable is ISI Web of Science[24], from the company (Thomson) that maintains the science citation index.

An important feature of both Science Direct and Web of Science is the **forward citation search**. This allows you to find all the later works that cite a given paper. So, once you have found a key reference, you can find who has gone further on the same subject. In Science Direct this is marked "Cited By"; in Web of Science "Times Cited".

Google has Google Scholar[25]; like other parts of Google this relies on "crawling" around the web to look for articles, it is thus hit-or-miss and is not a database. As with the rest of the web, there is no quality control. However, you may be able to find full-text of articles, as well as "grey literature" reports.

References

[1] American Society of Agronomy; Crop Science Society of America; & Soil Science Society of America. 1998. *Publications Handbook and Style Manual*. Madison, WI: American Society of Agronomy. URL: http://www.asa-cssa-sssa.org/publications/style/

[2] Basher, L. R. 1997. *Is pedology dead and buried? Australian Journal of Soil Research* **35**:979–94

[3] Bishop, Y.; Fienberg, S.; & Holland, P. 1975. *Discrete multivariate analysis: theory and practice*. Cambridge, MA: MIT Press

[4] Brockhaus GmbH, F. A. 2000. *Der Brockhaus in einem Band*. Leipzig: F. A. Brockhaus GmbH, 9. edition

[5] Buol, S. W.; Hole, F. D.; & McCracken, R. J. 1989. *Soil genesis and classification*. Ames, IA: The Iowa State University Press, 3rd edition

[6] CBE Style Manual Committee. 1994. *Scientific style and format : the*

[20] http://www.itc.nl/library/digital_library.asp
[21] http://www.sciencedirect.com/
[22] http://www.blackwell-synergy.com/
[23] http://springerlink.metapress.com/
[24] http://www.sciencedirect.com/
[25] http://scholar.google.com/

CBE manual for authors, editors and publishers. Cambridge: Cambridge University Press, 6th edition

[7] Center for Advanced Spatial Technologies (CAST). 1998. *AR-GAP final report: State-wide biodiversity mapping for Arkansas*. Report, Center For Advanced Spatial Technologies (CAST), Fayetteville, AR. URL: http://www.cast.uark.edu/gap/

[8] Dobos, E.; Micheli, E.; Baumgardner, M. F.; Biehl, L.; & Helt, T. 2000. *Use of combined digital elevation model and satellite radiometric data for regional soil mapping*. *Geoderma* **97**(3-4):367–391

[9] Ettema, C. H. 1994. *Indigenous soil classifications*. On-line PDF document; URL: http://www.itc.nl/personal/rossiter/Docs/Misc/IntroToEthnopedology.pdf. Access date: 27-December-2004

[10] FAO. 1984. *Land evaluation for forestry*. Forestry Paper 48. Rome, Italy: Food and Agriculture Organization of the United Nations

[11] FAO. 1998. *Digital Soil Map of the World and derived soil properites (CD-ROM)*. Land and Water Digital Media Series No. 1. Rome: FAO

[12] Forbes, T. R.; Rossiter, D.; & Van Wambeke, A. 1982. *Guidelines for evaluating the adequacy of soil resource inventories*. SMSS Technical Monograph 4. Ithaca, NY: Cornell University Department of Agronomy, 1987 printing edition. ITC 631.47

[13] Gaultier, J. P.; Legros, J. P.; Bornand, M.; King, D.; Favrot, J. C.; & Hardy, R. 1993. *L'organisation et la gestion des données pédologiques spatialisées: Le projet DONESOL (English abstract)*. *Revue de Géomatique* **3**(3):235–253

[14] Gobin, A.; Campling, P.; Deckers, J.; & Feyen, J. 2000. *Quantifying soil morphology in tropical environments: Methods and application in soil classification*. *Soil Science Society of America Journal* **64**:1423–1433

[15] Goovaerts, P. 1999. *Geostatistics in soil science: state-of-the-art and perspectives*. *Geoderma* **89**(1-2):1–45

[16] de Gruijter, J. 1999. *Special issue: Pedometrics '97*. *Geoderma* **89**(1-2):1–400

[17] de Gruijter, J. J. & Marsman, B. A. 1984. *Transect sampling for reliable information on mapping units*. In Nielson, D. R. & Bouma, J. (eds.), *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA*, pp. 150–163. Las Vegas: PUDOC, Wageningen

[18] Heuvelink, G. B. M. 1998. *Error propagation in environmental mod-*

elling with GIS. Research Monographs in GIS. London: Taylor & Francis. ITC 910.1:681.3 : 502.5

[19] Hornby, A. S. (ed.). 1995. *Oxford advanced learner's dictionary*. Oxford: Oxford University Press. ITC 038.20.951 reference

[20] Kainz, W. 1995. *Logical consistency*. In Guptill, S. C. & Morrison, J. L. (eds.), *Elements of spatial data quality*, pp. 109–137. Oxford: Elsevier (on behalf of the International Cartographic Association). ITC 910.1:681.3

[21] King, D.; Bourennane, H.; Isambert, M.; & Macaire, J. 1999. *Relationship of the presence of a non-calcareous clay–loam horizon to DEM attributes in a gently sloping area*. *Geoderma* **89**(1-2):95–111

[22] Kubiëna, W. L. 1948. *Entwicklungslehre des Bodens*. Wien: Springer-Verlag

[23] Lillesand, T. M. & Kiefer, R. W. 1994. *Remote sensing and image interpretation*. New York: John Wiley & Sons, 3rd edition

[24] Maas-Prijs, P. 2002. *How to use EndNote 5.0*. Handout, ITC

[25] Marsman, B. A. & de Gruijter, J. J. 1986. *Quality of soil maps : a comparison of soil survey methods in a sandy area*. Soil survey papers 15. Wageningen: Soil Survey Institute. ITC 631.4 S

[26] McBratney, A. B.; Odeh, I. O. A.; Bishop, T. F. A.; Dunbar, M. S.; & Shatar, T. M. 2000. *An overview of pedometric techniques for use in soil survey*. *Geoderma* **97**(3-4):293–327

[27] Oliver, M. A.; Webster, R.; & Slocum, K. 2000. *Filtering SPOT imagery by kriging analysis*. *International Journal of Remote Sensing* **21**(4):735–752. (GEOBASE Abstract)

[28] van Reeuwijk, L. P. 1995. *Procedures for soil analysis*. ISRIC Technical Paper 9. Wageningen: ISRIC

[29] Ruffin, E. 1832. *An essay on calcareous manures*

[30] Sahay, S. & Woolshan, G. 1996. *Implementation of GIS in India: organizational issues & implications*. *International Journal of Remote Sensing* **10**(4):385–404

[31] Skidmore, A. K. 1999. *Accuracy assessment of spatial information*. In Stein, A.; Meer, F. v. d.; & Gorte, B. G. F. (eds.), *Spatial statistics for remote sensing*, pp. 197–209. Dordrecht: Kluwer Academic. ITC 519.2 : 528.88

[32] Soil Resources Inventory Group. 1981. *Soil resource inventories and development planning: Proceedings of workshops at Cornell Univer-*

sity 1977-1978. Soil Management Support Services (SMSS) Technical Monograph 1. Washington, DC: Soil Conservation Service, USDA

[33] Stein, A.; Meer, F. v. d.; & Gorte, B. G. F. (eds.). 1999. *Spatial statistics for remote sensing*. Dordrecht: Kluwer Academic

[34] USDA Natural Resources Conservation Service. 2000. *Welcome to the USDA Natural Resources Conservation Service homepage for North Carolina*. On-line document; URL: http://www.nc.nrcs.usda.gov/. Access date: 27-December-2004

[35] Várallyay, G.; Hartyáni, M.; Marth, P.; Molnár, E.; Podmaniczky, G.; I., S.; & Kele, G. 1995. *Talajvédelmi információs és monitoring rendszer. 1 kötet módszertan*. Technical report, Földmüvelésügyi Minisztérium, Budapest. (In Hungarian)

[36] Webster, R. & Beckett, P. H. T. 1968. *Quality & usefulness of soil maps*. *Nature* **219**:680–682

6 Technical writing

Key chapter points

1. The purpose of technical writing is to **communicate information** to the reader.
2. One way to structure technical writing is to **write from an outline** (§6.1).
3. The key skill in technical writing is **writing clearly** (§6.2).
4. Problems are often encountered with **verbs** (tense and voice) (§6.3) and **punctuation** (§6.4).
5. Special problems are encounted by **non-native speakers** of English (§6.6).

The purpose of technical writing is to **communicate information** to the reader in as **compact**, **clear**, and **efficient** way possible..

In the case of the MSc thesis, the purpose is to explain *what* you did, *why* you did it, *how* you did it, what you *discovered*, and what *conclusions* can be drawn, as succinctly and clearly as possible. **The MSc thesis is the 'story' of a research project** and must be written clearly, concisely and attractively, so other scientists (including your examiners) can determine what you have done and how well you have done it.

There are many aspects of technical writing; in this section we only deal with:

- Structured technical writing; and
- Style in technical writing.

These might be termed 'macro'– and 'micro'–English, respectively.

Please keep the following guiding principle in mind:

Communication comes from clarity:

- Clarity of **structure**;
- Clarity of **logic**;
- Clarity of **grammar**;
- Clarity of **vocabulary**.

A classic short guide to clear technical writing is by Katzoff [7]; Gopen & Swan [4] present a psychology-based approach to effective scientific writing, with several worked examples.

6.1 Structuring technical writing by outlining

It is intimidating to most researchers (and even some professional writers!) to begin with a blank piece of paper. There are several techniques to get started and to keep going. This is highly personal. Some people make great progress with a structured technique, whereas others find their creativity blocked. If you have a system that works, fine. But if not, the *outlining* technique presented here may be attractive.

One approach which is especially appealing to those with structured minds (such as many scientists and engineers) is *outlining*, that is, working from the overall structure of the document in a *hierarchical* manner to arrive at the specifics.

A major advantage of this method is that you are sure all the pieces will be there before you have to write. Also, you see their inter-relation, in particular, that the order of argumentation is clear.

MS-Word directly supports outlining, with its 'View | Outline' command, which uses heading styles to organise the outline. It allows you see the document structure at different levels and to easily re-organise the document.

The structure of the thesis already provides an outline, namely your chapter titles. For example:

1. Introduction
2. Materials & Methods
3. Results
4. Discussion
5. Conclusion

Then you break down each of these into sub-topics. For example, 'Materials & Methods' could be divided as follows:

1. Introduction
2. **Materials & Methods**
    - **2.1. Sampling design**
    - **2.2. Field methods**
    - **2.3. Data processing**
    - **2.4. Data analysis**
3. Results
4. Discussion
5. Conclusion

These sub-topics in turn may be broken down into sub-sub-topics, and so forth. For example, 'Field methods' could be broken down as follows:

1. Introduction
2. Materials & Methods

2.1. Sampling design
2.2. Field methods
2.2.1. Infiltration and saturated water content
2.2.2. Soil profile description
2.2.3. Bulk density
2.3. Data processing
2.4. Data analysis
3. Results
4. Discussion
5. Conclusion

This process can continue indefinitely; however, more than four levels is unusual.

Finally you reach the level of the **paragraph**, which is a set of sentences that work together to make one point. The limit between a paragraph and a section is not sharp. Sections go deeper into a topic, and require several related paragraphs.

The 'topic sentence' technique

Within each section, you may want to use the *topic sentence* method to define paragraphs. The idea is to write a single sentence that introduces the paragraph, and leave the details of that paragraph for later. Remember, many readers will skim your writing exactly this way: they will read the first sentence and then decide whether to read the fuller explanation.

6.1.1 An example of composition from an outline

Here is an example adapted from Rossiter [8], where a single section is expanded first in topics and then in topic sentences. The title of the paper is "Economic land evaluation: why and how." Within the paper one of the sections is "Factors to consider in the calculation of economic suitability":

Topic list

§: Factors to consider in the calculation of economic suitability

¶1 Don't account for land purchase or rental in any case.

¶2 Return to labour vs. return to land, when to use each.

¶3 Externalities and when to use them.

¶4 Shadow prices and when to use them.

These show the four issues that, according to the author, should be taken into account when computing economic suitability. The next step is to write a complete first (topic) sentence for each paragraph:

Topic sentences

§: Factors to consider in the calculation of economic suit-

ability

¶**1** Since land is the entity being compared in land evaluation, no costs associated with acquisition or rental of the land should be included in any of the economic measures. ...

¶**2** Economic suitability may be expressed in terms of the return to labour or the return to land. ...

¶**3** Externalities are off-farm effects that are not reflected in the production unit's budget. ...

¶**4** Shadow prices are those set by the economist to reflect the true value of an input or output to society. ...

The next step is to summarize what the paragraph should say; most authors skip this step and just start writing, but it can be useful to be clear to oneself what should be in the paragraph.

For example, ¶**3** will cover the following points:

Paragraph points

¶3
‡ define externalities; example
‡ handled differently in financial vs. economic analyis
‡ in financial analysis
‡ in economic analyis
‡ analyst must be clear on which was used

Each of these points will become one or two sentences in the final paragraph.

Finally, each paragraph must be expanded until it is complete. Each new sentence must add something to the paragraph. If it doesn't, it probably belongs elsewhere. Here is a complete paragraph from the above example, with each sentence shown separately for emphasis, and with an explanation of what it is meant to contribute to the argument:

A final paragraph

¶3
‡1 (*define externalities*): Externalities are off-farm effects that are not reflected in the production unit's budget.
‡2 (*make this statement concrete to the reader*): Examples include water pollution and sedimentation of reservoirs.
‡3 (*when to use them? two cases ...*): They are accounted for differently in financial and economic analyses.
‡4 (*define financial analysis*): A financial analysis is from the point of view of the individual land user, so externalities are ignored unless a monetary cost to the land user is imposed by society.
‡5 (*make this statement concrete to the reader ...*): An exam-

ple is a tax on sediment discharge.
→ (*this is to be contrasted with ...*)
‡6 (*define economic analysis*): An economic analysis is from the point of view of society, so externalities must be included.
‡7 (*how?*): The techniques of resource economics are used to assign them a negative economic value.
→ (*bring the two thoughts together*)
‡8 (*conclusion and opinion*): In both cases the analyst must state which type of analysis was used, and, in the case of an economic analysis, which externalities were included and how.

Note the **parallel construction** used in ‡4 and ‡6 to emphasize the contrast between the two kinds of analysis proposed in ‡3.

Here is the final paragraph, without comments, as the reader will see it in the paper:

> "Externalities are off-farm effects that are not reflected in the production unit's budget. Examples include water pollution and sedimentation of reservoirs. They are accounted for differently in financial and economic analyses. A financial analysis is from the point of view of the individual land user, so externalities are ignored unless a monetary cost to the land user is imposed by society; an example is a tax on sediment discharge. An economic analysis is from the point of view of society, so externalities must be included. The techniques of resource economics are used to assign them a negative economic value. In both cases the analyst must state which type of analysis was used, and, in the case of an economic analysis, which externalities were included and how.

6.2 Writing clearly

Now you've organised, outlined, organised your argument into paragraphs, and decided what to say. But, how do you actually write readable, coherent text?.

Don't try to be literary or clever! Clear, direct, unambiguous and forceful writing is called for in scientific communication. One reason that English is the international scientific language is because it is suitable for "plain speaking".

Booth *et al.* [2, §11.4.1] explain two styles of drafting, which appeal to different types of writers:

- Write as fast as possible, correct later;

- Write carefully, don't leave any problems.

You have to experiment to find the one which fits your style of work. An intermediate form is to write fast, but keep a separate page with notes to yourself of things to check on later.

Everyone has a time of day or days of the week when they are at their most creative. Schedule the creative phase of your writing for these times. Other times are suitable for detailed work such as checking spelling and details of the argument.

Re-read your text with a fresh eye

In a separate step, **re-read your text as if it were written by someone else**, and ask yourself:

- Is the **argument** clear? Does it say what you want it to? Remember, *you* know what you wanted to say, but can the *reader* find that in your text?
- Does it read **smoothly**? Would it benefit from more explanation or connectives?
- Is it as **short and direct** as possible? Are there redundant words or phrases? Have you said the same thing twice? Can the text be shortened?
- **Does it read like English?** Writers who are not accustomed to writing English often use constructions typical of their native language. If what you are reading doesn't sound like it was written by a native, most likely it is because of such problems.

Write for your intended reader

There are many styles of writing. A good criterion for selecting a style is to think about your intended reader. For a thesis, the most important readers are the members of the examining committee, especially the external examiner.

Shorter is (usually) better

Make your point in as few words as possible. A shorter text is easier on the reader. But . . .

Don't sacrifice length for clarity

Use all the words you need to make your point clearly. In particular, use *connectives* to link parts of your text and bring out their relation. For example, look at the use of the connective 'in particular' in the previous sentence.

Use a dictionary

You should always work with a standard dictionary [e.g. 9] to be clear on word meaning and usage.

Technical points of style are dealt with in standard English-language references, such as the Chicago manual [5]. Each field has its own difficulties regarding nomenclature, units of measure, etc.; these are covered in specialised style manuals. An example for soil science is American Society of Agronomy *et al.* [1, Ch. 2].

Use the spelling checker

Use your word processor's **spelling checker**, but remember that just because a word is correctly spelled, that does not mean it is the word you want!

> "Infiltration measurements were located on **bear** soils ..." (Are these soils where bear like to walk? Usually we prefer **bare** soils.)
>
> "All the evidence points to one **confusion** ..." (That's unfortunate; I was hoping that you could come to a **conclusion**. But maybe you really **do** mean 'confusion'!)
>
> "We had to decide **weather** or not to include **boarder** trees."

Pick the right word

Be especially alert for words that seem right, but aren't. Common errors are with the pairs (*imply, infer*), (*much, many*), and (*affect, effect*).

Avoid barbarisms

Avoid barbarisms, i.e. words that seem to be English but which are not, for example *to impact* (correct: *to affect, to have an impact*). This is a common error of sloppy native speakers, especially provincial newspapers in the USA.

In all this, if you are unsure, **use the dictionary!**.

Maybe use the grammar checker

MS-Word's **grammar checker** is surprisingly good, but again, make sure its advice is correct for your situation.

When you have finished a piece of text, put it aside for a day, and then **re-read it**. You will easily see the obvious mistakes in spelling and punctuation; when in doubt use your dictionary or style manual.

6.3 Verbs: tense and voice

English verb constructions are very rich and can cause problems both for native and non-native speakers. Here we present only some aspects of verb use that cause problems in a thesis project; consult a grammar for more information.

6.3.1 Tense

This is the time to which the verb refers. It should present few difficulties if you place yourself in the position of the reader and consider the time to which the statement refers *at the time it is written.*

Future This is used for events in the future when the document was written. It is often used in the "Methods" section of the research proposal:

> ‘Fifty plots **will be** sampled.’

It is also used for speculative statements in the “Conclusions” section of the thesis:

> ‘Sub-pixel classification **will become** a routine technique for forest inventory.’

Past This is used for events already in the past when the document was written. It is often used in the “Methods” section of the thesis:

> ‘Fifty plots **were** sampled.’

Note that only the tense had to change once the work was done.

It is also used for a result that is specific to our study:

> ‘Ground control with single-receiver GPS survey **was** able to adequately rectify 1:5 000 scale small-format aerial photography to map accuracy standards.’

This is simply a statement of fact about our study, and does not implies that we think the method is *in general* successful. If we believe that, we should use the present (see below).

Present This is used for statements that are always true, according to the author, for some continuing time period.

> ‘Sub-pixel classification **is** a new technique for forest inventory.’

This statement may be false some years from now, but it’s true at the time of writing and for some time thereafter.

It is also used for a result that is widely-applicable, not just to our study:

> ‘Ground control with single-receiver GPS survey **is** adequate to rectify 1:5 000 scale small-format aerial photography to map accuracy standards.’

This statement implies that we think the method is *in general* successful; our data show this and we are confident it is everywhere true.

Several other tenses are less common:

Past perfect This is used for events already in the past when another event in the past occurred.

> ‘Fifty plots **had been** sampled as part of a previous project.’

This implies a context such as "...when we decided how many more plots to sample."

Future perfect This is used for future events that will have been completed when another event in the future occurs.

> 'Fifty plots **will have been** sampled by this project.'

This implies a context such as "...before we arrive to sample our plots."

6.3.2 Voice

The two voices are **active** and **passive**. In the first case the subject of the sentence controls the verb:

> 'Pests damage crops.'

In the second case there is no explicit subject, only (perhaps) one implied in the complement.

> 'Crops are damaged by pests.'

Note that "by pests" is not necessary for a grammatical sentence. We could have said "Crops are damaged in the spring".

Passive voice The passive voice should be used when the object is more important than the subject, especially when the identity of the subject doesn't matter:

> 'Nelson Mandela is widely respected.'

This implies "by everyone"; the sentence could be re-written:

> 'Nelson Mandela is respected by everyone.'

It could also be re-written in the active voice:

> 'Everyone respects Nelson Mandela.'

However, this puts the main information at the end of the sentence, thereby making it weaker. But if this sentence were paired with a contrasting sentence with a different subject, the active voice would bring out the contrast:

> 'Everyone respects Nelson Mandela, but very few know his middle name[26].'

Avoiding "I" One use of the passive is to avoid calling attention to oneself; this is considered egotistical:

[26] Rolihlahla

'I selected five representative villages.'

as opposed to:

'Five representative villages were selected.'

The second form is *ambiguous* unless the entire context implies the actor. It could be made explicit:

'Five representative villages were selected by me.'

but that sounds awkward. It sounds better if referring to a third party:

'Five representative villages were selected by the project team.'

and these forms are grammatically equivalent.

When the active voice is required

This is nicely discussed by Webster [10]:

> "There are several actions that I very definitely want reported in the active voice: they are ones of assumption, decision and choice. When I read 'It is assumed that', 'It was decided to', and 'Sites were chosen', I immediately ask myself who assumes?, who decided?, and who chose? – the author?, or his boss?, his client?, or some overseeing committee? These actions lie at the heart of original research, and results depend crucially on them . . .
>
> "A second reason why I want to know who did what is that we scientists are human, and we are fallible therefore. We make mistakes, we take foolish decisions, we choose unrepresentative specimens, we overlook uncomfortable results, and we misinterpret what we see. We do so not necessarily, not usually, wilfully, but we are responsible, and readers must know that we are."

Use of "the author"

One way to avoid "I" while still being clear about who is the actor is to use the euphemism "the author" in the third person:

'The author selected five representative villages.'

'The authors, basing themselves on the general geomorphic map, partitioned the study area into three sampling strata: the lake plain, the volcanic plain, and the isolated volcanic hills.'

6.4 Punctuation

Punctuation is used to break words into groups with a related function, so that it is easier for the reader to understand the intent of the author.

The rules for the full stop or period (“.”), semi-colon (“,”) and colon (“:”) are fairly standard, but the comma (“,”) is used in a more varied manner. When in doubt, consult a style manual. Punctuation has changed over the years and also can vary substantially between countries or even styles within a country (journalistic, popular, academic, formal, . . .).[27]

Use of the period (“.”) This completes a sentence, which must be grammatically-correct. A speaker would come to a full stop at this point when reading the document aloud.

> ‘No man thinks more highly than I do of the patriotism of the very worthy gentlemen who have just addressed the House.’

Use of the exclamation point (“!”) This also ends a sentence, but indicates surprise or emphasis.

> ‘I repeat it, sir, we must fight!’

Use of the question mark (“?”) This also ends a sentence, but indicates a question which could be answered, either by the reader or the author.

> ‘Is life so dear, or peace so sweet, as to be purchased at the price of chains and slavery?’

Here the reader is expected to answer “No!”[28]

Use of the semi-colon (“;”) This joins two complete sentences where the second is closely related to the first. A speaker would come to a pause at this point when reading the document aloud. The semi-colon can be repeated.

> ‘We have petitioned; we have remonstrated; we have supplicated; we have prostrated ourselves before the throne, and have implored its interposition to arrest the tyrannical hands of the ministry and Parliament.’

Each sentence above is complete by itself. The author could have written:

> ‘We have petitioned.
> We have remonstrated.
> We have supplicated.

[27] The texts in this section are adapted from the speech given by Patrick Henry, 23–March–1775 to the Second Virginia Convention, commonly referred to by its closing statement “Give me liberty or give me death”.
[28] The author answers his own question: “Forbid it, Almighty God!”

> We have prostrated ourselves before the throne, and have implored its interposition to arrest the tyrannical hands of the ministry and Parliament.'

The author wanted to draw attention to the close relation between the four actions. Note that the last sentence must use a comma instead of a semi-colon because the phrase "and have implored ... " does not have an independent subject and so can not be an independent sentence.

Use of the colon (":") The colon is placed at the end of a complete sentence to introduce examples that are usually a list but can also be a phrase or sentence.

> 'I have but one lamp by which my feet are guided: the lamp of experience.'

Note that the part before the colon form a complete sentence; the part after expands the thought but is not necessary to complete the grammar.

Use of the comma (",") In American usage, the comma is placed where the speaker would make a brief pause to show the relation between words. One reason would be because the words form a list, and the speaker must clearly show that the items are separate:

> 'I am willing to know the whole truth, to know the worst, and to provide for it.'

Here the speaker has three closely-related items which are in a list:

1. to know the whole truth,
2. to know the worst, and
3. to provide for it.

In American usage there is a comma before the last clause, to avoid ambiguity (are there two or three items in the list?). In British usage the final comma is omitted.

Another use is to join an introductory clause to the main part of the sentence, at the point where a speaker would take a breath. It adds nothing to the understanding of the sentence.

> 'For my own part, I consider it as nothing less than a question of freedom or slavery.'

6.5 Some matters of style

There are many excellent style manuals [e.g. 5]; here I only point out some common style errors made by ITC students.

Repeated words

> 'We interviewed ten local farmers. Local farmers said that ...'

> 'We interviewed ten local farmers, who said that ...'

Superfluous words

> 'As a result of the field measurements, it could be observed that the average steady-state infiltration rate of the soils was 1.2 cm hr^{-1}. '

> 'The average steady-state infiltration rate was 1.2 cm hr^{-1}. '

In this example, it should be clear from the previous sections of the paper or thesis, specifically the 'methods', that infiltration was measured in the field. So it is not necessary to repeat this in the results.

Wordiness

> 'The correlation matrix between the NIR and IR bands was found to be 0.95 for LANDSAT TM7 and 0.96 for ASTER images. '

> 'The NIR and IR bands were highly correlated (LANDSAT TM7 $r = 0.95$, ASTER $r = 0.96$). '

6.6 For non-native speakers

Non-native speakers have special problems with writing English. Here we briefly discuss a few of these. For some languages there are specialist books available that discuss common pitfalls for that particular language (e.g. Burrough-Boenisch [3] for Dutch speakers).

If you are not completely comfortable with the English language[29], and you are serious about improving it, not just surviving the thesis-writing phase, I highly recommend the Oxford Advanced Learners Dictionary [6]. This has extensive notes on correct usage, discussions of difficult topics such as use of articles and prepositions, and suggestions to make your writing idiomatic (i.e., as if it were written by a well-educated native speaker).

Don't write in your native language

Some students like to write in their native language and then translate. This almost always is more work than **thinking and writing in English** from the beginning. It's OK to write notes to yourself in your own language, but not full texts.

[29] and who really is? Certainly not the current President of the world's most powerful English-speaking country!

Avoid Latinisms

For students whose native language is a modern dialect of Latin (e.g. Portuguese, Spanish, French), there are words with cognates in English that have a very different meaning in English. A well-known examples is *actually*, meaning "in fact" (rather than "now, currently"). More serious are using full Latin constructions, e.g. "Could not be sampled the soil" (correct: "The soil could not be sampled").

Latin-speaking students often try to write long, complex sentences with many dependent clauses. Short, simple sentences are easier to write and to understand.

Avoid non-English constructions

Other native languages give other typical problems in English. Some languages (e.g. Slavic) don't use articles, and many do not use articles the same way as English. So a student may write "Soil was red", which in English means that soil, in general, has a red colour, rather than "The soil was red", meaning that the specific soil in question was red. Dutch and German (as well as modern Latin) speakers often use gendered pronouns where English uses "it", e.g. "We sampled the soil. *He* [correct: *it*] was red". English does not use both the noun and pronoun in the same phrase, e.g. "The soil he was red" (correct: omit "he").

Do not use an automatic translator

Automatic translators such as BabelFish (available from Altavista) are interesting ideas, and can often cope with simple declarative sentences. You can use them to get an idea of what you might write, but I have yet to see one that gives acceptable results. Be especially careful of their translations of words with several meanings in either the target or source language.

For example, the verb 'to bear' in English may be translated into Spanish as:

- 'sostener' (to support, e.g. 'The beam can bear a heavy load'),
- 'dar' or 'producir' (to yield, e.g. 'The tree bears fruit'),
- 'llevar' (to carry, e.g. 'The wise men came bearing gifts'),
- 'tener' (to have, e.g. 'He bears a distinguished name'),
- 'soportar' (to stand, e.g. 'I can't bear his jokes'),
- 'odiar' (to hate, e.g. 'I can't bear spiders'),
- 'parir' (to give birth, e.g. 'She bore a child'),
- 'resistir' (to stand up to, e.g. 'That idea will not bear close inspection'), and
- various other idiomatic uses.
- In addition, the noun 'bear' (the animal) is translated as 'el oso' (or 'la osa').

Similar lists can be made when translating *into* English. This is difficult enough for a human translator, and well beyond the capabilities of computers. It's even worse when there are grammatical issues. Try the following:

- "Time flies like an arrow." (the comparative 'like', i.e. 'similar to'; the verb 'to fly'; a simile)
- "Fruit flies like an apple." (the verb 'to like'; the noun 'fly', a type of insect; a declarative sentence)

References

[1] American Society of Agronomy; Crop Science Society of America; & Soil Science Society of America. 1998. *Publications Handbook and Style Manual*. Madison, WI: American Society of Agronomy. URL: http://www.asa-cssa-sssa.org/publications/style/

[2] Booth, W. C.; Colomb, G. G.; & Williams, J. M. 1995. *The craft of research*. Chicago: University of Chicago Press. ITC 001.818

[3] Burrough-Boenisch, J. 1998. *Righting English that's gone Dutch*. The Hague: Sdu Uitgevers

[4] Gopen, G. D. & Swan, J. A. 1990. *The science of scientific writing*. *American Scientist* **78**(November-December):550–558

[5] Grossman, J. (ed.). 1993. *The Chicago manual of style : the essential guide for writers, editors and publishers*. Chicago: University of Chicago Press, 14th edition. ITC 001.811

[6] Hornby, A. S. (ed.). 1995. *Oxford advanced learner's dictionary*. Oxford: Oxford University Press. ITC 038.20.951 reference

[7] Katzoff, S. 1964. *Clarity in technical reporting*. Technical Report NASA SP-7010, National Aeronautics and Space Administration (NASA)

[8] Rossiter, D. G. 1995. *Economic land evaluation: why and how*. *Soil Use & Management* **11**:132–140

[9] Thompson, D. (ed.). 1995. *The concise Oxford dictionary of current English*. Oxford: Oxford University Press, 7th edition

[10] Webster, R. 2003. *Let's rewrite the scientific paper*. *European Journal of Soil Science* **54**(2):215–218. http://www.blackwell-synergy.com/links/doi/10.1046/j.1365-2389.2003.00515.x/

7 Assessing the quality of an ITC MSc thesis

Key chapter points

1. An ITC thesis is judged by a **Thesis Assessment Board** (TAB) consisting of an **external examiner** from a University, the **ITC Professor** in a relevant field, and several ITC or affiliated scientists competent in the fields covered by the thesis; the thesis **supervisor**(s) advise only (§7.1).
2. ITC has approved list of evaluation criteria (§7.2), classified as:
   - **Scientific scope and depth**: the research addresses a well-formulated relevant problem of sufficient scope and depth linked to relevant literature;
   - **Scientific method**: the research is undertaken with a clear and transparent methodology with proper use of concepts, methods and techniques;
   - **Reporting**: the thesis is a well structured and readable with a clear layout;
   - **Presentation & defense**: the research is well presented, followed by a discussion with proper argumentation.

The quality of the completed MSc thesis and the degree to which the candidate understands what was done and can defend it against other approaches are assessed by a Thesis Assessment Board.

! → The completed thesis is available from the *ITC library* for any interested party, and, if the mark received is 75 or higher, is also placed *on-line* as an Adobe PDF file for instant access from anywhere in the world. Thus it must be a reliable piece of work.

7.1 Examination procedure

An ITC thesis is judged by a **Thesis Assessment Board** (TAB).

For degrees conferred at ITC itself, this consists of:

- An **external examiner** of high academic rank, i.e. a Professor or Associate Professor from the Dutch university system;
- The **ITC Chair** or, in some cases, Associate Professor, responsible for quality control of the specialisation;
- The ITC thesis supervisor(s);

- One or two other ITC scientific workers, often with different scientific expertise than the supervisor and Chair, competent to judge the thesis.

For degrees conferred by ITC and a partner institute, i.e. Joint Education Programmes (JEP), the composition is slightly different, depending on the academic regulations in the partner's country and the logistics of the thesis exam. In general the exam is held at the partner institute, and the TAB consists of:

- The **partner institute Chair** of the academic department responsible for quality control of the specialisation;
- The **ITC Chair** or, in some cases, Associate Professor, responsible for quality control of the specialisation;
- The partner institute thesis supervisor;
- One other partner institute scientific worker, often with different scientific expertise than the supervisor and Chair, competent to judge the thesis.

The ITC supervisor is asked to give written comments on the thesis quality and the candidate's performance during the research and thesis writing phase.

In either case the TAB reads the thesis and hears the candidate's defense, and then grades the thesis on the following scale:

100 **Perfect**[30]

≥ 90 **Excellent**: publication quality, no flaws, quite innovative, could be a chapter in a PhD thesis;

≥ 80 **Very Good**; well above expectations, only minor flaws, innovative, research has no serious questions and can be incorporated into a journal article;

≥ 70 **Good**; meets expectations of a typical work within the time allowed and with the facilities available; nothing special but nothing really bad;

≥ 60 **Pass**: meets minimum standards, passing; not innovative, some serious flaws;

< 60 **Fail**: does not meet minimum standards.

The interpretation of terms such as 'good', 'well above expectation' etc. is **completely up to the discretion of the Board**. Most Boards give points

[30] The Dutch reserve this grade for the perfect Being; by definition no human is perfect, so draw your own conclusions about your chances of receiving this mark

in between, e.g. 75 for a thesis which is not "outstanding" but has features that make it more than simply "good". A grade of 80 or above is rare; it usually requires that the student not make any serious mis-steps during the thesis period.

The mark is composed from three parts (1) the written thesis; (2) the oral defense and (3) assessment of the learning process; of these the written thesis receives much the highest weight.

The Board assesses a thesis on the basis of quality criteria only; mitigating circumstances are not taken into account.

! → The thesis grade is adjusted **downwards** if, during the exam, the candidate does not appear to understand the work or is not able to defend it. In this situation the committee wonders if the student did the work and in case of serious doubt may fail the student even if the thesis is of sufficient quality.

! → The thesis grade is adjusted either **upwards** or **downwards** according to the candidate's independence, initiative, effective communication with supervisors, etc.; in short, the student's performance in the thesis period. This depends on the opinion of the supervisors, backed up by documentation. This adjustment is no more than a half-step (5 points).

ITC maintains quality equal to Dutch universities, so the opinion of the external examiner is of utmost importance. **The external examiner must sign the exam results in order for the student to pass**, so it is clear who is the primary audience for your thesis. ITC asks the external examiner to ensure that ITC grades correspond to those in the Dutch universities.

7.2 Evaluation criteria

This section is adapted from the "Criteria for MSc Thesis Assessment" approved by the ITC Degree Board in November 2005. The checklist is given to the TAB; however, the grade is **holistic** summary of the thesis, *not* the sum of points from the checklist.

Scientific scope and depth The research addresses a well-formulated relevant problem of sufficient scope and depth linked to relevant literature.

- Is the research problem clearly defined? (E.g. through well formulated research questions).
- Is a relevant research problem being addressed?
- Has the research problem been placed in the context of the scientific field concerned?

- Is there a critical discussion of and link to relevant contemporary literature?
- Is the research undertaken of sufficient scope and depth?
- Is there evidence of a thorough understanding and mastering of the subject and discipline?
- Is there an innovative part in the research?

Scientific method The research is undertaken with a clear and transparent methodology with proper use of concepts, methods and techniques.

- Were the research methods appropriate to answer research questions (conceptualization and operationalisation of the research questions)
- Is the research process and methodology clearly described and well structured?
- Are the methods and techniques for data collection and analysis properly selected and applied?
- Was the data collection and analysis performed using the correct methods and with proper reference to literature?
- Have the objectives been reached and/or are research questions answered?
- Are conclusions drawn correctly after analysis of data?
- Are the conclusions and statements supported by evidence?
- Is there a critical discussion and reflection on the research findings and awareness of the limitations of the research?

Reporting The thesis is a well structured and readable, with a clear layout.

- Is the thesis well and clearly written?
- Is the thesis well structured
- Is the thesis logically written?
- Is proper use made of literature references, and was proper referencing applied?
- Has effective use been made of visualization tools like maps, tables and graphics?

Presentation and defense The research is well presented, followed by a discussion with proper argumentation.

- Did the presentation provide a clear and concise summary of the research?
- Was the candidate capable to respond adequately to questions, criticisms and comments?
- Did the candidate make proper use of the thesis during the defense?

Process The candidate worked in a structured and rather independent way, while making adequate use of the guidance of the supervisor.

- Does the thesis reflect the candidatesâĂŹ own research ideas and efforts?
- Was the research planned and undertaken in an independent and structured way?
- Did the candidate take initiatives?
- Was there a good communication between the candidate and the supervisors/staff?

8 Abstracting a research paper or thesis

Key chapter points

1. The abstract is often **the only part of your work that will be read**; either because it is all that is available or because the reader is in a hurry.
2. The abstract is **the paper in miniature**; everything that is important in the paper must be included (in abbreviated form) in the abstract.
3. The abstract does not usually contain **citations** or **detailed reasoning**; there is not enough room to prove your case as you do in the thesis or paper.

Your thesis work fits into the larger enterprise of scientific progress. Others want to know what you have accomplished and what you have discovered, so they can verify or extend your work, or just use its results. The abstract is the only part of your work that most people read, and sometimes the only part they have available to them.

To quote American Society of Agronomy *et al.* [1, p. 12]:

> "An abstract has two typical uses. [It] helps readers decide whether to delve into the paper as a whole; abstracts are also published separately in outlets such as Web sites and secondary and indexing journals. Thus, **the abstract will be seen and read by many more people than will read the paper**.
>
> "With this in mind, a basic rule emerges: **Everything that is important in the paper must be reflected in the abstract**. . . . Be specific. In essence, an informative abstract . . . presents **the paper in miniature**, complete within itself. It moves from an introductory statement of the rationale and objectives or hypotheses, through materials and methods, to the results and conclusions."

The abstract is the paper in miniature

Different journals have slightly different rules, but in all cases the abstract must fit on one page. A general rule is that the abstract must be written as:

- one continuous paragraph;
- with a limit of 250 to 300 words.

The abstract can usefully be structured exactly as the thesis, with one or more sentences for each section:

1. Rationale
2. Hypothesis and objectives
3. Methods
4. Results
5. Conclusions

Note that these words do *not* appear in the abstract itself! They are listed here just to show you the structure.

Since space is so limited, writing must be very compact. There is very little need in such a short piece for connective text. Every word should count. Some sentences from the introduction and conclusion may be taken over almost verbatim, but usually they must be shortened.

Some style points

- The abstract should not contain **citations** unless they are absolutely necessary to understand the work; an example is if the main purpose of the paper is to follow up someone else's work.
- There is not enough room in the abstract for **detailed reasoning**; you are not expected to to prove your case as you do in the thesis or paper.
- Pick a voice (active or passive) and stick to it; do not change voices. Because it is assumed that the text refers to your work if not explicitly mentioned otherwise, passive voice is acceptable.
- Phrases such as "The results show", "The analysis reveals" etc. are rarely needed, for the same reason: these are usually wasted words.
- Don't refer to the authors in the 3rd person – you are the authors.

8.1 An example

Here is an example, written by me and loosely based on the thesis by Hengl [2] with the different sections labelled. Note that I've tried to put in as much *specific* information as possible.

Rationale — "Semi-detailed soil survey is a costly and time-consuming activity, requiring, among other, subjective expertise for photo-interpretation of soil landscapes.

Objective — "The aim of this study was to replace as far as possible subjective photo-interpretation in these surveys with a more rapid and objective procedure.

Method — "Conventional photo-interpretation maps were prepared for five sample areas totalling 111km^2, representing 10.5% of a survey area in Baranja County, eastern Croatia. Landform parameters extracted from a DEM (slope gradient, wetness index, relative elevation, plan curvature, profile curvature, and relative annual solar radiation) were used to train several supervised classifiers, and these were then used to extrapolate over the entire area.

Results — "Using all six predictor variables resulted in a reasonable classification ($\hat{k} = 0.4$) of soil-landscape units inside the training areas. Greatly improved accuracy was obtained by separately classifying high ($\hat{k} = 0.7$) and low ($\hat{k} = 0.5$) relief areas, using only units known to occur in each area. In both landscapes, a reduced predictor set of three variables, as well as the first three principal components, provided almost as much accuracy as the full set. Some photo-interpretation classes, representing about 15% of the area, were consistently mis-classified.

Discussion — "The extrapolation maps show fine details not achievable by photo-interpretation. Some of these may be artifacts but others represent small areas of contrasting soils. Areas where the method did not succeed were mostly in flatter areas where the DEM had insufficient vertical resolution.

Conclusion — "The technique is quite promising and, with some obvious refinements, should be operationalised."

Here is the final form, as one long paragraph of 261 words. It presents the entire article in miniature, and stands on its own as information:

> "Semi-detailed soil survey is a costly and time-consuming activity, requiring, among other, subjective expertise for photo-interpretation of soil landscapes.The aim of this study was to replace as far as possible subjective photo-interpretation in these surveys with a more rapid and objective procedure. Conventional photo-interpretation maps were prepared for five sample areas totalling 111km^2, representing 10.5% of a survey area in Baranja County, eastern Croatia. Landform parameters extracted from a DEM (slope gradient, wetness index, relative elevation, plan curvature, profile curvature, and relative annual solar radiation) were used to train several supervised classifiers, and these were then used to extrapolate over the entire area.Using all six predictor variables resulted in a reasonable classification ($\hat{k} = 0.4$) of soil-landscape units inside the training areas. Greatly improved accuracy was obtained by separately classifying high ($\hat{k} = 0.7$) and

low ($\hat{k} = 0.5$) relief areas, using only units known to occur in each area. In both landscapes, a reduced predictor set of three variables, as well as the first three principal components, provided almost as much accuracy as the full set. Some photo-interpretation classes, representing about 15% of the area, were consistently mis-classified. The extrapolation maps show fine details not achievable by photo-interpretation. Some of these may be artifacts but others represent small areas of contrasting soils. Areas where the method did not succeed were mostly in flatter areas where the DEM had insufficient vertical resolution. The technique is quite promising and, with some obvious refinements, should be operationalised."

References

[1] American Society of Agronomy; Crop Science Society of America; & Soil Science Society of America. 1998. *Publications Handbook and Style Manual*. Madison, WI: American Society of Agronomy. URL: http://www.asa-cssa-sssa.org/publications/style/

[2] Hengl, T. 2000. *Improving soil survey methodology using advanced mapping techniques and grid-based modelling*. MSc thesis, International Institute for Aerospace Survey & Earth Sciences (ITC)

9 Presenting a research proposal and results

Key chapter points

1. Researchers have a duty to **communicate** with those who fund, use, benefit from and evaluate their work.
2. In a spoken presentation, the aim is to **communicate the main points** in a short time and **engage** the audience in discussion; details are left for written reports or handouts.
3. Each **type of audience** must be approached in a manner appropriate to its level, background and interests. (§9.1).
4. **Text** should be short (keywords or phrases) and legible (§9.2).
5. Each main slide should be discussed for thirty seconds to a minute; this limits the number of slides.
6. **Do not repeat what is written on the slide**; instead, discuss it (§9.6).
7. **Graphics** greatly enhance a presentation; they must be uncluttered and intuitively communicate the main point (§9.3).

As a working research scientist, you have a duty to **communicate** with those who

- fund your work (donors, research councils);
- can use your work to further their work (colleagues);
- benefit from your work (society at large); and
- evaluate your work (superiors, external review commissions).

Written reports are important, but often you will be asked to make a *verbal* presentation, almost always with *graphical support*. This communication is part of your professional responsibility as a scientist.

You do not have to be a preacher, an entertainer, or a politician, only a competent public speaker. Your message is important (we hope!), and the presentation is a way to bring out its principal points to the appropriate audience.

Keep in mind key differences between a written and spoken presentation:

- With a written report, the emphasis is on completeness, whereas in the spoken presentation, the aim is to communicate the main points and convince the audience.
- In the written report, communication is one-way, but in a presentation it can be two-way (during question time); therefore the presentation should **engage** the audience and **encourage discussion**.

- The reader can take as much time as necessary to absorb what is written, the listener must understand in real time; therefore not too much can be presented.

Several good books have been written on this subject (e.g. Anholt [1], Rabb [2]); here I give some hard-won advice from these sources and my own experience.

9.1 Designing and writing the presentation

Before you begin

- **Know your audience**. At ITC this will be your scientific peers (fellow MSc candidates) and supervisors. At home it might be your superiors, non-technical decision-makers, technicians in a completely different field, or the local population at large.
- **Decide what you are trying to accomplish** with the presentation. At ITC this will be to communicate your research plans or results, and to convince your audience that you understand these. At home it might be to get funding, to build bridges to another work group, or to be allowed to do work in a local community.
- **Decide on your primary message**. Your audience can not absorb very much in real-time during an oral presentation; you want to **maximize** the information **absorbed** and **remembered**by your audience, in the time alloted.

These will allow you to choose a technical level, a level of formality or familiarity, a choice of vocabulary (specialist or common language), and so forth. You can then choose what goes in the presentation and what part must be left out.

Designing the presentation

- Think about the presentation **from your audience's point of view**. Simply put, why should they be interested in what you have done? Why should they spend their time listening to you? Tell them exactly that (not in those words, of course) to introduce the presentation.
- Plan to use **80% – 90%** of the allotted time. This gives you some margin for error. For example, if the total time is given as 12 minutes, plan and practice to use 10. This allows you to relax and not rush.
- A good rule of thumb is **one slide per minute**, not counting the title and "thank you" slides.

- Keep the presentation **short and to the point**. Your audience can only absorb so much; you can always explain more during question time or afterwards.
- You are **telling a story**. It must have a **beginning**, **middle**, and **end**. The various parts must be **linked**.
- Some people like to use an "outline" slide, but this time is generally wasted. It's best to get right to the purpose of the presentation.
- Start with the wider context, **"zoom in"** to your specific part of this, and at the end **"zoom out"** to the context again. This shows how your work relates to the large issues.
- In a geographical study, include a **location slide**: a map (and photo if possible) of the study area, to orient the audience.
- You must **repeat key points** in order to make an impact. "Tell them what you are going to tell them, tell them, and then tell them what you just told them".
- Finish with a strong and clearly-stated, simple **conclusion**, a 'take-home' message. The last thing you say will most likely lead to the first question, and it is also what the audience will most remember. For scientific work, the most useful conclusion is a one– or two–sentence summary of your main results.

hidden slides

- **Stick to the main points**, and hold extra information in reserve in case there are questions (for example, the 'Hidden Slide' feature in PowerPoint, or have a set supplementary extra overheads keyed to the main material).
- You can **provide supplementary material**, for example a reprint of the paper(s) you are discussing, or more complete results than you have time to present. Then you can refer those who want more details to this material, and you won't feel that you have to cram everything into the presentation. In the case of a thesis or other research proposal, you should supply the actual written proposal.

Details

- **Be very careful with your statements**; use precise and accurate language. **Never say anything that you can not defend** in case you are questioned.
- Avoid jargon and clichés.
- Be very clear about which statements are well-supported with data and which are speculative. Speculation belongs only in the conclusion.

9.2 Writing the text

Nowadays most presentations are PowerPoint[31] slides, but old-fashioned overheads are useful in low-tech situations.

- Each main slide or overhead should be discussed for thirty (30) seconds to about a minute. If it has too much material for this time, split into several slides.
- Everything you are going say should be at least reflected by some element on the slide. There should always be something you can point to that refers to what you are saying. This can be text or a graphic.
- Conversely, if you aren't going to say anything about something, don't put it on the slide.
- The above two points do *not* imply that you have to read the slide! Rather, that everything on the slide must refer to something that is mentioned, and vice-versa.
- Slides that introduce or link parts of the presentation can be scheduled for about ten seconds.
- Use few words (key words) and large fonts.
- Do not fill the slide; leave plenty of room around the sides.
- Use phrases instead of complete sentences, to save space.
- You are expected to explain more than is written. You can prepare speaker notes to go with each slide, from which you can read. These can be just notes to jog your memory, or, if you are really underconfident, a full text.
- Especially for non-native speakers, make sure every unfamiliar word is written either on the slide or speaker notes. This way you won't "go blank" and forget a word under stress.

9.3 Designing the graphics

- If it's not legible, eliminate it or enlarge it.
- Show simple graphs, figures, or tables. Every element should be legible. You may have to simplify or zoom in on one part of the graph.
- Do not put too much on one slide; however, you may need to show several graphs side-by-side for comparison.

[31] or similar programs such as Keynote

- Use extra graphic elements, e.g. arrows with brief text, to draw the audience's attention to the parts of the graphic you are going to discuss.

9.4 Sequencing

- Allocate your time to sub-topics according to its importance in the context of the talk.
- Give most of the time your own material (what is new is what you are doing) rather than background.

9.5 Using PowerPoint

- I intensely dislike animation and 'cute' little pictures, but some people find them attractive.
- Simple slide transitions (e.g. dissolve, window shade) can break up the visual monotony, but too much looks (to me) unprofessional.
- Break up a long series of text or tabular slides with some relevant photographs or even a work of art; this gives a visual break and clearly marks a new section.
- Bringing lines of text into view as you speak (equivalently: blocking lines of the overhead and revealing them as needed) can focus the audience's attention on the point at hand, and prevent them from reading ahead.

9.6 Making the presentation

- **Practice, practice, practice!** This is what a spouse or best friend is for. A painful but effective methods is to videotape the practice presentation and review it.
- Know what you are going to say at the very beginning (welcome, introduction etc.), word-for-word. This is especially true if you are nervous.
- Similarly, know what you are going to say to **conclude**. Finish on a positive note and show clearly that you are done; this allows the audience to applaud (notice how the great composers finish their musical works). A useful formula is something like "In conclusion, from this work I can state confidently that ... Thank you for your kind attention".
- **Relax!** Easy enough to say, I know ... but if your presentation is well-prepared, all you have to do is follow it. Just tell the story. Of course, if you are nervous because your material is weak, you really

do have something to be nervous about. The only solution is to do better work!

- **Speak slowly and clearly**; too few words are preferable to too many. Avoid words you can not pronounce well.
- For **native English speakers**, use as neutral an accent as you can; avoid a conversational or local accent;
- Also for **native English speakers**, use standard constructions and the simplest vocabulary you can while still explaining your message; avoid slang, catch phrases, or cultural references that might not be understood by everyone in the audience.
- Look at your audience, establish eye contact, make them feel that you are entering a dialog with them (even if you are doing all the talking).
- **Do not repeat what is written on the slide**; the audience can read it (to themselves) faster than you can speak. Instead, use what is written to support what you are saying.
- Especially **do not read every number in a table!** Instead, point out the key ones you want to discuss,
- A little humour and remarks for your specific audience will lighten a technical presentation, but be sure you know what is considered humorous and what in bad taste *in the specific situation*.
- Unless you know your audience extremely well, avoid casual remarks about religion, ethnicity, or politics.
- **Never apologise** for lack of time, or anything else. You can *explain* facts, for example, that you were only able to collect a limited number of samples, but don't apologise.

9.7 Internet resources

- Resources for making Scientific Presentations, from the NOAA Seattle Regional Library, has links to several good resources.
  http://www.wrclib.noaa.gov/lib/reference/scipresentations.html
- Excerpts from Anholt [1]:
  http://www-physics.mps.ohio-state.edu/~wilkins/writing/Supp/dazzle.html

References

[1] Anholt, R. R. H. 1994. *Dazzle 'em with style : The art of oral scientific presentation.* W H Freeman & Co

[2] Rabb, M. Y. 1993. *The presentation design book : tips, techniques & advice for creating effective, attractive slides, overheads, multimedia presentations, screen shows & more.* Chapel Hill, NC: Ventana, 2nd edition. ITC 003.6