AUTOMATED IDENTIFICATION AND NETWORK ANALYSES IN *ACACIA* POLLINATION ECOLOGY

A thesis submitted to the University of St Andrews for the degree of Master of Philosophy in the Faculty of Science

September 2005

Anna T. Watson

School of Biological Sciences

Funded by a NERC Case Studentship in association with the Natural History Museum, London.

CONTENTS **Page**

ABSTRACT

Acacia trees are an important source of pollen and nectar to a wide range of savannah insects. Identifying *Acacia*-visiting insects and the herb pollens they also carry is essential for pollination studies but problematic as expert entomologists and palynologists are overworked. Automated identification offers a solution to this problem.

The DAISY automated identification system can identify insects accurately but the pattern selection methodology has been time consuming and impractical for large applications. Two new methodological approaches are investigated that allow quicker processing or avoid wing removal. They showed great promise, identifying species with more than 90% accuracy.

Pollens provide very different challenges to computer vision, they are three-dimensional, often vary substantially within a single species and have complex shapes. In the initial DAISY identification of pollens accuracy was low (34% at best) so several methodological means to improve accuracy were investigated. The processing regime recommended for greatest accuracy is clean pollens, imaged with dark field microscopy and which had their relative sizes maintained. Cleaned and sized pollens were identified with 100% accuracy when only 5 pollens were considered but this fell to around 50% accuracy when 50+ pollens were considered. As most applications will require many pollens to be considered at once this is not yet a useful system for pollens.

Network analysis is a recent technique in pollination ecology, well suited to pollen load analysis. The most important insects to the pollination network were all bees: *Apis mellifera, Megachile* and the small halictids *Lipotriches* and *Patellapis.* The damage from removing a single insect genus was generally below 5% and never more than 10% of network integrity. Removal of combinations of insects showed great tolerance to extinction, this is likely to be due to functional redundancy and nestedness of the network, with specialists visiting plants that are also visited by generalists.

DECLARATION

I, Anna Watson, hereby certify that this thesis, which is approximately 40 000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

date 26.1.06... signature of candidate .

I was admitted as a research student in October 2003 and as a candidate for the degree of MPhil in 2004; the higher study for which this is a record was carried out in the University of St Andrews between 2003 and 2005.

date 26.1.06... signature of candidate

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Master of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date 26.1.06... signature of supervisor ..

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker.

date 26.1.06... signature of candidate .

THE AUTHOR

The Author was awarded a 1st class B.Sc. Honours in Ecology, from the University of Wales, Bangor in 2002. Having worked as a research entomologist in Belize, funded by The Natural History Museum and The Systematics Association, she recently published her first research paper (Watson *et al.,* 2004).

ACKNOWLEDGEMENTS

I wish to thank John, Liz, Mike and the Mpala *Acacia* team for all their support. I am grateful to Mark O'Neill for all the energy he has put into our DAISY and network collaborations. Pat Willmer, Ian Kitching and Mark O'Neill provided valuable supervisory guidance and detailed comments on the manuscript. Patrick Lenguya provided diligent field assistance, collecting and pinning insect specimens. Connal Eardley, the National Museum of Kenya, National Museum of Scotland, the Museums and Galleries of Wales and the Natural History Museum provided assistance in the identification of insect specimens.

ABBREVIATIONS

A	Attached wing method
ABIS	Automated Bee Identification System
AI	Artificial Intelligence
ALARM	Assessing LArge scale Risks for biodiversity with tested Methods
AMNH	American Museum of Natural History
ANN	Artificial Neural Network
ANSI	American National Standards Institute
API	Application Programming Interface
AToL	Assembling the Tree of Life
ATW	Anna Tamsin Watson
B	Boxed wing method
bfw	Bees, flies and wasps
bs	Bees and syrphids
C	Clean pollen
C	Coordination level, e.g. C3
CAT	Computer-Assisted Taxonomy
CC	Cross Correlation
CLIPS	C Language Integrated Production System
DAISY	Digital Automated Identification System
DF	Dark-field
DFE	DAISY Front-End
DIARES-IPM	DIagnostic Advisory Rule-based Expert System for Integrated Pest Management
DiCANN	Dinoflagellate Categorisation by Artificial Neural Network
DNN	Dynamic Neural Network
EU	European Union
FPTP	First Past the Post
GAA	Glacial Acetic Acid
GBIF	Global Biodiversity Information Facility
GUI	Graphical User Interface
HAB	Harmful Algal Bloom, e.g. HAB-buoy
HTML	Hyper Text Markup Language

ID	Identification
IEEE	Institute of Electrical and Electronics Engineers
INBio	Instituto Nacional de Biodiversidad
IR	Infrared
ISI	Institute for Scientific Information
jpeg	Joint Photographic Experts Group
KBS	Knowledge-Based System
LDA	Linear Discriminant Analysis
LF	Light-field
LINNE	Legacy Infrastructure Network for Natural Environments
MAO	Dr Mark O'Neill
MASDEA	MArine Species Database for Eastern Africa
MOSIX	Multi-computer Operating System for UNIX
MRC	Mpala Research Centre
NHM	Natural History Museum
NMGW	National Museums and Galleries of Wales
NMK	National Museums of Kenya
NMS	National Museum of Scotland
NNC	Nearest Neighbour Classification
NPT	Normalised Polar Thumbnail
NVD	Nearest Vector Difference
PBI	Planetary Biodiversity Inventory
PCA	Principal Component Analysis
PEET	Partnerships for Enhancing Expertise in Taxonomy
PFT	Partial Fault Tolerance
PolyROI	Polygonal Region of Interest
POSIX	Portable Operating System Interface
PSOM	Plastic Self-Organising Map
ROI	Region of Interest
S	Standard method
s.d.	standard deviation
SDS	Sodium Dodecyl Sulphate
s.e.	standard error
SEM	Scanning Electron Microscope

SOM	Self-Organising Map
SPIDA	SPecies IDentified Automatically
tiff	Tag Image File Format
TS	Training Set
UC	Unclean pollen
UCR	Universidad de Costa Rica
UK	United Kingdom
US	United States
VHTML	Virtual HTML

TAXONOMIC NOTE

Since this research was done the accepted taxonomy of *Acacia* has been changed. The genus formerly known as *Acacia* has been split into five genera. The original type species for *Acacia* was an African species. However, the Australian species are the most diverse (there are nearly 1000 'acacias' in Australia) so a botanico-legal request was made for the Australian species (subgenus Phyllodineae) to keep the same *Acacia.* This request was accepted and the change was ratified on 30 July 2005.

The African species have been split between two previously known genera. Those that were in the subgenus Acacia are now in the genus *Vachellia.* Those that were in the subgenus Aculeiferum are now in the genus *Senegalia.* Six of the seven former-acacias at Mpala Research Centre are now *Vachellia.* These are the trees at which most of the insects were gathered. The species that was formerly known as *Acacia mellifera* is now *Senegalia mellifera.*

Chapter 1 – Acacia pollination and computer-aided taxonomy

1.1 The aims of this research

Pollinators provide essential ecosystem services (e.g. Costanza *et al.*, 1997) but declines in some pollinator communities have been reported (e.g. Watanabe, 1994; Williams, 1982). These communities cannot be conserved effectively unless the fundamental components defining them are better understood (Potts *et al.*, 2003). For this reason, pollination ecology is important and the taxonomic shortfalls that are limiting progress (discussed in section 3.1.1) must be overcome.

Acacia pollination is a good model system to assess the limitations of the Digital Automated Identification System (DAISY) as it poses two distinct taxonomic challenges, insects and pollen. Acacia visiting insect species are very numerous, providing a large pool size for analysis, but many species visit infrequently so training sets have to be small. It is important for insect specimens to be identified in the field on the day of capture, so that this knowledge can inform fieldwork. Therefore, the equipment used must be available in any minimally equipped field laboratory and image pre-processing needs to be as time-efficient as possible. The pollen species that might be present in pollen loads are even more numerous than the insects. These images are cross-section views of colourful, three-dimensional objects so they may pose significant difficulties to computer vision.

The DAISY research goes beyond pollination ecology. The work on pollinator wing venation would be applicable to any insect with transparent wings and clear venation, such as crop pests or quarantined species. The pollen work is relevant not just to pollens in other contexts, such as air samples for allergies or soil cores in palynology, but to any small three-dimensional object, such as the rock foraminifers that indicate the presence of oil.

Once the identities of both insects and pollens have been made clear then *Acacia* pollination, with many inter-connections, is an interesting proof of concept trial for the network algorithms of Dr Mark O'Neill of University of Newcastle. Network analyses based on pollen loads allow the pollination community to be modelled so that the impact of single and combined extinctions can be predicted. This work makes comparisons with Memmott *et al.* (2004), in which much larger networks were produced based on flower visitation observations.

1.2 Pollination ecology of Kenyan acacias

Acacia (Fabaceae: Mimosoideae) is endemic in tropical and temperate regions worldwide (Kenrick, 2003). These shrubs and trees grow in dry regions and have been studied in East Africa (e.g. Stone *et al.*, 1998; Stone *et al.,* 1999), Central America (e.g. Janzen, 1974) and Australia (e.g. Bernhardt, 1989; Kenrick & Knox, 1989; Kenrick, 2003). This is a very large genus, with over 950 species described in Australia alone (Maslin, 2001).

In Kenya, acacias dominate large areas of savannah, providing food and living environments for a huge diversity of wild animals, including large mammals, birds and particularly insects. Their economic role is modest but important. They provide firewood, their foliage and pods are browsed by domesticated goats and cattle (Fig. 1.1) and their sharp thorny branches used to build cattle enclosures, known as 'bomas'.

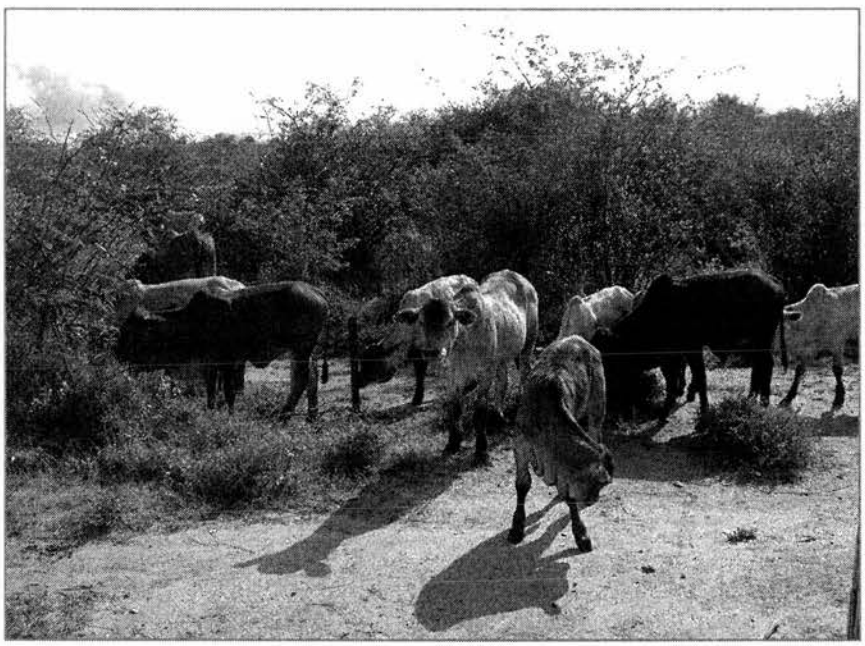


Fig. 1.1 – *Acacias* are a vital source of grazing for domesticated cattle, as ground vegetation is often scarce. This herd is browsing on *Acacia brevispica* at Mpala Research Centre. All the shrubs in this photo are *Acacia*.

The *Acacia* species studied at Mkomazi in Tanzania mainly coflowered after the long and short rains (Stone *et al.,* 1999). At Mpala, however, the *Acacia* species differed substantially in their phenology, flowering at different times of the year in a manner that was difficult to predict (the rains had become similarly unpredictable in recent years). For this reason it was necessary to work with whichever common species was flowering at the time. The *Acacia* species studied were *A. brevispica, A. etbaica, A. gerrardii, A. mellifera, A. nilotica* and *A. seyal.*

Acacia is a distinctive genus, characterised by bipinnate leaves, straight spines, sharp hooks, long pods and 'pom-pom' inflorescences (Fig. 1.2).

Fig. 1.2 – *Acacia nilotica* has 7-25 pairs of leaflets on each pinna, is armed with straight stipular spines as long as 8cm and bears spherical yellow inflorescences. Sketch by Rosemary Wise, taken from Coe & Beentje, 1991.

Acacia flowers are fundamentally all very similar in structure. The flowers are individually small (5-10 mm long and about 1mm across) but are collected together into inflorescences containing from 10-20 to 500 flowers (Fig. 1.3). The inflorescences are either spherical (*A. brevispica, A. etbaica, A. gerrardii, A. nilotica* and *A. seyal*) or bottle-brush-shaped (*A. mellifera*).

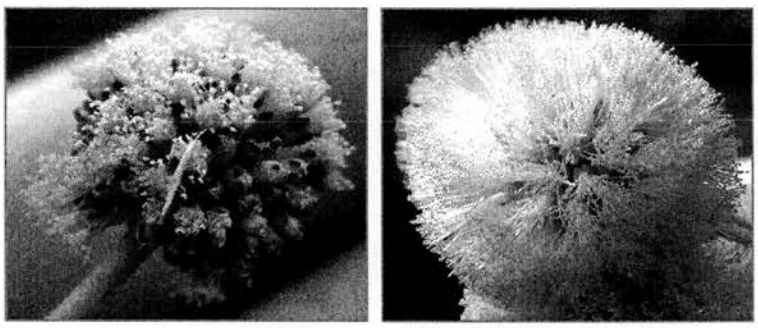


Fig. 1.3 – The separate flowers can be seen in a partially opened inflorescence of *A. nilotica* (left). Once an *Acacia* inflorescence, in this case *A. seyal*, is fully open the stamens meet such that separate flowers are no longer visible (right).

Research into the mechanisms of pollination and seed-setting of *Acacia* has been spasmodic and fragmentary (Kenrick, 2003). *Acacia* is hermaphroditic, each inflorescence going through a male phase and a female phase. In the Australian acacias the female phase comes first (Kenrick & Knox, 1989) but these Kenyan species seem to have their male phase first, with the styles extending only after the stamens have wilted. *Acacia* trees cannot fertilise their flowers with their own pollen (Kenrick & Knox, 1989), and so pollen transfer between trees of a given species is essential (Stone *et al.,* 1999).

Sherry (1971) considered that most *Acacia* pollen is not suited to wind transport as the pollen is too large. A limited amount of airborne dispersal is indicated for a few species (Kenrick, 2003) and *Acacia* polyads are found in aerial pollen counts wherever the genus grows (Smart & Knox, 1979). However, these airborne pollens are from species with unusually small compound grains (polyads), just 25 μm in diameter (Knox, 1979). The compound grains of most *Acacia* species exceed this size (Guinet, 1981), the species used in this study varied from 38 μm in *A. brevispica* to 57 μm in *A. mellifera* and *A. gerrardii* . Therefore, insect vectors are considered to be vital to *Acacia* pollination. The pollen is presented on the surface of the inflorescence (Kenrick, 2003), so they provide a potential resource to a wide diversity of insects: specialist pollen feeders (bees, beetles and some of the true flies, especially hoverflies), specialist nectar feeders (birds, butterflies and bee flies) and opportunistic foragers (other flies, ants and wasps) (Stone *et al*, 2003) (flower visitors are discussed further in section 3.1). Most pollinations occur by automimicry, when polyad laden insects attempt to forage on female phase inflorescences (Bernhardt, 1989). While entomological research at Mkomazi in Tanzania (e.g. Russell-Smith *et al,* 1999; Davis, 1999; van Noort and Stone, 1999) has shed some light on savanna insect communities, identification of insect specimens remains a major stumbling block for studies of *Acacia* pollination. Without the financial provision to pay overstretched taxonomists, many specimens remain unidentified and pollination papers are less effective than they could be.

As yet, little work has been done on the pollen characteristics of *Acacia.* Mimosoideae pollen grains are characteristically shed in permanent compound units called polyads (Guinet, 1981) (Fig. 1.4). *Acacia* species vary in the size and number of pollen grains they incorporate into each polyad (between 4 and 64 grains, most commonly 16) (Kenrick, 2003). Grain number is correlated with ovule number, which maximises the benefit from a single pollination event as many ovules can be fertilised from a single polyad transferral (Kenrick and Knox, 1989). Grains are generally arranged to form a biconvex disc that is two grains thick in the centre, with remaining grains arranged singly on the periphery (Kenrick and Knox, 1989) (Fig. 1.5).

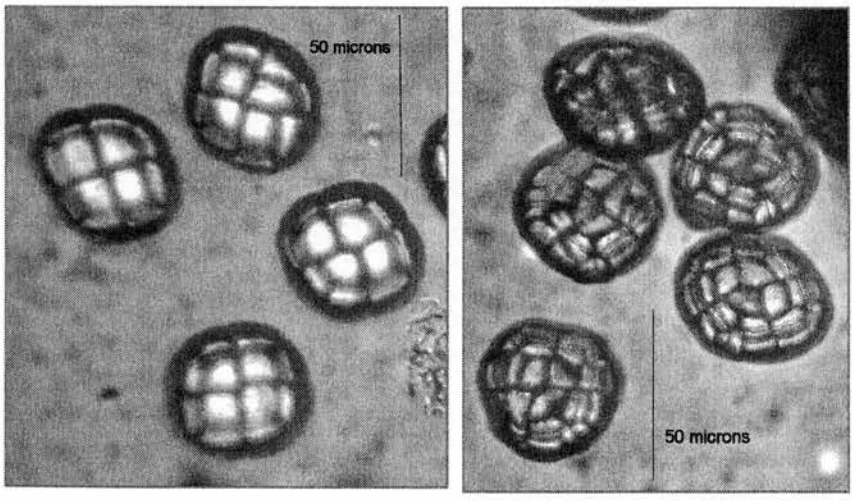


Fig. 1.4 – Optical microscope images of *A.brevispica* (left) and *A.nilotica* (right) polyads

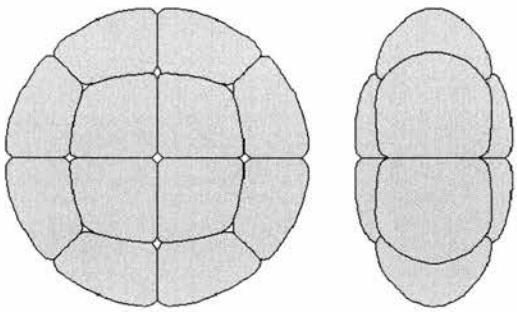


Fig. 1.5 – Diagrammatic representation of an *Acacia* polyad, a biconcave disk that is two grains thick at the centre. From www.bio.uu.nl

1.3 The Taxonomic Impediment to Pollination Studies

"Taxonomy provides the bricks, and systematics provides the plan, with which the house of the biological sciences is built."

Robert May, 2004.

Serious biological research in almost any biodiversity-rich country is limited by the ability to identify the species encountered accurately. This situation is often referred to as the "taxonomic impediment" (Wheeler, 2003). Names are the 'hooks' upon which information about organisms is stored, retrieved and exchanged (Maslin, 2002), so accurate taxonomy is essential to ecological research. Descriptive taxonomy is a 'facilitation science', comparable to genome sequencing or star mapping, yet it lacks the scale of funding that the Human Genome project or the Sloan Digital Sky Survey receives (Godfray, 2002). Wheeler (2003) estimated there were 8.3 million species still to be described, if work progresses at a rate equal to the average of the post-Linnaean period, we will need another 1196 years to complete the job and many species will become extinct before they are described (Wheeler, 2003). The 1992 Convention on Biological Diversity committed 167 nations to biodiversity assessment and protection, yet a lack of prestige and resources is crippling the cataloguing of biodiversity (Godfray, 2002). Several International organisations (generally funded by the U.S. National Science Foundation) have been set up to address different aspects of the taxonomic impediment (Wheeler, 2005), these are summarised in Table 1.1.

Table 1.1 – New organisations and projects to deal with the taxonomic impediment

Disappearing Expertise	Partnerships for Enhancing Expertise in Taxonomy (PEET) http://web.nhm.ku.edu/peet/ Rodman & Cody (2003)
Species description	Planetary Biodiversity Inventory (PBI) http://research.amnh.org/pbi/
Scattered data	Global Biodiversity Information Facility (GBIF) http://www.gbif.org/
Visual morphology knowledge	MorphBank http://www.morphbank.com/
Phylogenetic context	Assembling the Tree of Life (AToL) http://tolweb.org/tree/phylogeny.html
Integrated research	Legacy Infrastructure Network for Natural Environments (LINNE). http://www.flmnh.ufl.edu/linne/

Many authors have discussed possible ways forward (e.g. Lipscomb *et al*, 2003; Mallet and Willmott, 2003; Minelli, 2003; Tautz *et al*, 2003). These include:

- The use of local non-specialist taxonomic workers, known as parataxonomists, which has increased rates of specimen collection and processing in developing countries with highly diverse biota (Longino 1994). INBio (Instituto Nacional de Biodiversidad) in Costa Rica has been a pioneer of this approach (http://www.inbio.ac.cr/en/default.html).
- The use of morphospecies to circumvent the identification time lag, which is gaining acceptance for certain types of biodiversity assessment (Longino and Colwell 1997).
- The increasing use of bioinformatics technologies (as described in section 1.2.3), which is contributing to accelerated rates of accumulation and transfer of biodiversity information (Oliver *et al,*. 2000).

These complementary developments will remove much of the burden of mass sorting of field samples from the highly skilled but small taxonomic work force (Oliver and Beattie 1997).

The taxonomic impediment is felt strongly in the field of pollination biology. We need to understand pollination much better if we are to manage ecosystems successfully (Potts *et al*, 2003). Pollinators play a critical role in the maintenance of ecosystem diversity but pollination systems are under threat from habitat fragmentation, changes in land use, modern agricultural practices and invasions of non-native species (Kearns *et al*, 1998). Pollinator communities are in decline in many part of the world (Potts *et al*, 2003) and the Convention on Biological Diversity has recently approved the International Pollinator Initiative (Williams, 2003). Little information on pollination in Africa has been published. This is partly due to insufficient taxonomic knowledge of both plants and pollinators (Eardley, 2002). Even if the species involved have been described they may be unidentifiable to a non-specialist;

classification does not guarantee identification.

Recent advances in computing and electronics (with high speed, low cost microprocessors) are leading to new applications in biology and conservation. Computing algorithms can now automate the creation of low-cost identification keys (Reynolds *et al.*, 2003). Hypertext, multi-access keys and the Internet are making conventional taxonomic tools more user-friendly. Meanwhile, a range of automated identification systems is being developed (e.g. Hajdaoud *et al.,* 2005; Watson *et al*, 2004; Chesmore & Ohya, 2004), mainly using images and audio data and exploiting developments in artificial intelligence (Gaston & O'Neill, 2004).

1.4 Computer Assisted Taxonomy

1.4.1 Electronic resources

Traditionally, taxonomic descriptions and identification tools have been distributed in print, usually as long and complex dichotomous keys. Taxonomic journals are incredibly diverse and can be very specialised; most fail to reach the pages of the Institute for Scientific Information (ISI) Journal Citation Reports (Minelli, 2003). Journal page space and printing costs have selected for telegraphic jargon (terminology that is only readily understood by specialists [Weeks *et al*, 1999]) and against colour images; electronic publications (on CD-ROM, DVD or the Internet) have greatly reduced these restrictions (Godfray, 2002).

Electronic publication allows information to be linked in a non-linear manner, using hypertext. Text can be browsed and keywords or icons (known as 'buttons') clicked on to jump to another part of the document. Hypertext can make a dichotomous key less daunting as the user is presented with one couplet at a time, often illustrated with diagrams, photographs or sound clips. The University of Queensland website 'Bright Minds' has an aquatic invertebrates key that is a good example (http://www.brightminds.uq.edu.au/thechallenge/whatami/html/aquatics_key/couplet_1.htm). Transient buttons allow additional information, such as glossary definitions, to be added to the screen to expand on the text. Some hypertext keys include a decision map for easy backtracking and all have a 'back' button. On reaching an identity, species descriptions and photographs are obtained without having to move to a different part of the publication.

While hypertext can make dichotomous keys more user-friendly, the strict linear progression of a paper dichotomous key remains unchanged; a key that is difficult in wording or structure will remain difficult

to use (Edwards and Morse, 1995). Dichotomous keys are severely limited by this linearity. If a character is unobservable (e.g. antennae lost) a couplet may be unanswerable. Without a decision on this couplet the user cannot progress. Multi-access keys were designed to circumvent the 'unanswerable couplet problem' (Maslin, 2002). They were originally based on punched cards (polyclaves) but now are computerised as 'random access identification guides', exploiting the benefits of hypertext. The user decides the order in which character states are entered and can choose the characters they are able to distinguish. Many multi-access keys advise on this by ranking characters in terms of how well they discriminate between remaining species. If insufficient information is given then more than one species may be matched, or if a character is misidentified no species will match. In the latter case, a similarity index indicates which species is the closest match (Edwards and Morse, 1995). CABIKEY multi-access keys have been published on CD-ROM (e.g. Mosquito genera in Ramsdale & Ramsdale, 1998). Good multi-access keys can also be found on the Internet. The Natural History Museum (London) key to lichens (http://internt.nhm.ac.uk/cgi-bin/botany/lichen) and the Noctuid Search key (http://www.plant.cdfa.ca.gov/noctuid) illustrate very different styles that can be adopted. There is a multi-access key to Australian acacias available on CD-ROM, entitled 'WATTLE: Acacias of Australia' (http://www.worldwidewattle.com/infogallery/publications/wattle.php), which was coordinated by Maslin (2001b). He discussed the keys available for Australian acacias, concluding that electronic multi-access keys had advantages over conventional printed keys (Maslin, 2002). No such guide exists to the African acacias, which are served only by a traditional key (Coe & Beentje, 1991).

Tardival and Morse (1997) gave computerised dichotomous and multi-access keys, and a conventional paper key, to a group of zoology undergraduates, who had to use them to identify a species of woodlouse. The success rate was around 74% for all three keys but students commented that the computer keys were easier to use. While multi-access keys may be more user-friendly than dichotomous keys, specialist terminology can still impede use. Also, such keys still require the user to discriminate subtle differences in form between taxa, distinctions that may only be perceptible to an experienced taxonomist (Weeks *et al*, 1999a).

Many of the tools taxonomists produce for the identification of species, e.g., keys, have been largely ignored by the general public in favour of "field guides", which are essentially browsable picture guides. Stevenson *et al.* (2003) reviews the role of electronic field guides and discusses the application of a host of digital technologies to produce user-friendly tools for non-specialists. The first modern field guide, with colour plates, was produced in 1934 by R. T. Peterson and entitled 'A Field Guide to the Birds'. It had a huge impact, allowing the development of a large group of skilled amateur birders. The economics of publishing dictate that paper field guides must be commercially viable, so they tend to focus on popular taxa and cover wide geographical areas. Large online booksellers are best equipped

to cater for more specialised audiences. A search of www.amazon.co.uk in August 2005 using "field guide" as a key phrase produced 1915 books for birds and 623 for plants. As with paper field guides most electronic products deal with birds. Stevenson *et al.* (2003) lists more than 10 CDs of birding software. Their electronic format allows inclusion of bird songs, a search facility, and games and quizzes to help in learning birds. They may also include "listing" programs to keep track of the birds seen and output the user's bird list in useful formats such as maps. However, these electronic resources have yet to rival to portability of a book (Stevenson, 2003). A growing number of Internet sites offer field guide information about species.

A further development in electronic resources involves the addition of a unique barcode to each biological sample or mounted specimen (Oliver *et al.,* 2000; Thompson, 1994). The BioTrack laboratory at Macquarie University, Sydney, uses the Biota database management software over a local area network at workstations that consist of a computer, a microscope and a barcode scanner. Specimens are sorted, vials are barcoded, and data and images are entered into Biota at the time of processing, with no use of handwritten labels of datasheets. This approach increases the speed at which data are entered into and retrieved from biodiversity databases. Thompson (1994) estimates that the use of barcodes saves one biodiversity technician at least a quarter of a million keystrokes per year. Barcoding also improves data quality by eliminating transcription and typing errors (Oliver *et al.,* 2000). It requires initial investment in hardware (e.g. server and client computers, barcode scanners, camera, network hubs) but such technology is becoming more and more affordable. Oliver *et al.* (2000) recommends that all new biodiversity assessments adopt the use of barcodes and provides protocols for using barcode technology and virtual reference specimens for large scale invertebrate biodiversity assessment.

1.4.2 The Internet

Taxonomy is ideal for the Internet, being information-rich and requiring many illustrations. Yet taxonomic information on the web is minimal (Godfray, 2002). The Internet has the potential to revolutionise taxonomic practice, allowing global access to original literature and type specimens (through digital images). High-resolution image databases are critical for achieving the goal of an online identification network (Oliver *et al.,* 2000). The 'virtual herbarium' of the New York Botanical Garden comprises over 700,000 herbarium specimens and 100,000 high-resolution specimen images (http://sciweb.nybg.org/science2/VirtualHerbarium.asp). The type collection of University of Gottingen (GOET) is also online (Schmull *et al.*, 2005). Specimen identification by comparing real specimens to an on-screen reference specimen of each species is fast, can be undertaken simultaneously by different

investigators working on the same taxon, and minimizes the need for repeated handling of valuable reference specimens (Oliver *et al.,* 2000).

Internet "telemicroscopy" takes this a step further, since remote users not only view high-definition images but also remotely control the microscope at which the images are being captured (Wolf et al. 1998; Brauchli *at al.*, 2002; Wheeler, 2005). The 'Envision' system of remote microscopy will be based in three large entomological collections (London, Paris and Washington) and will allow users to examine, manipulate, image and archive specimens. This will make around 100 million specimens available from anywhere with an Internet connection. 'Envision' will include a conference facility, so experts in different places can discuss a specimen as it is being examined. Such conferences could be webcast in museums or live on-line to give the general public a 'fly-on-the-wall' experience of taxonomists at work (Wheeler, 2005).

Websites often host free photo guides, the largest and most complete deals with US wildlife, www.enature.com. As well as field guides to over 5500 species, enature is also able to offer facilities not present on CD-ROMs, e.g. to have questions answered by an expert and to get regular natural history notes by email. One of the most innovative approaches to Internet field guides comes from The Royal Ontario Museum and lets users build their own field guide (http://www.rom.on.ca/ontario/fieldguides.php). The software uses a three step process in which the user selects one of the 54 regions in Ontario, chooses one of three groups (birds, amphibians or fish) and either a field guide or a checklist output format. This tailoring of resources to user needs is a great advantage of the Internet (Stevenson, 2003).

Internet photo galleries are useful for African pollen species. The best overall site is the African Pollen Database website hosted by medias.obs-mip.fr. The University of Arizona concentrates its site on US pollen but it is still a useful reference for *Acacia* polyad images (http://www.geo.arizona.edu/palynology).

Specialist websites such as the Australian Biodiversity Information Facility (http://www.deh.gov.au/biodiversity/digir/), AlgaeBase (http://www.algaebase.org/), Antbase (http://antbase.org/) and FishBase (http://www.fishbase.org/) provide global biodiversity catalogues (Knapp *et al*, 2002). Other online databases are more geographically specific, such as MASDEA, the Marine Species Database for Eastern Africa (Vanden Berghe, 2005). Online meta-databases such as Species2000 (which has now collated 500,000 species, http://www.sp2000.org/) and web-based biodiversity projects such as the Tree of Life (http://tolweb.org/tree/phylogeny.html) aim to gather all this information together. Currently, the taxonomic knowledge relating to a species is spread amongst

the accumulated literature, cross-referenced by biological nomenclature and position in the taxonomic/phylogenetic hierarchy. Patterson (2003) suggests we need a biological names register. Oliver *et al.* (2000) propose that virtual entomological collections should be linked by search engines. This online invertebrate identification network would be analogous to that used by molecular biologists searching their distributed databases with new molecular sequences (see the National Centre for Biotechnology Information, http://www.ncbi.nlm.nih.gov/).

Godfray (2002) called for web revisions of taxonomic groups to be maintained online, bringing a unitary taxonomy into a single place. This idea has been developed by other taxonomists. Wilson (2003) made the case for a single-portal electronic encyclopaedia of life with a page for each species, summarising everything known about that species and linking to databases. These revisions could include information not currently required in a formal description (such as images and keys) and links to relevant sites, becoming information hubs. Wheeler (2005) suggested Curated Virtual Monographs of taxa groups, these would be accessed online and would allow the user to choose their output format (e.g. monograph, field guide, checklist or map). A unitary taxonomy would require extensive administration, with its associated costs. While taxonomists continue to disagree on its worth, such funds are unlikely to be found (Knapp *et al*, 2002).

1.4.3 Expert Systems

The most accurate identification comes from an experienced taxonomist, who knows not only specialised information but also how to best apply it. "Expert systems" are computer programs that employ the knowledge and reasoning processes of human experts (Galbraith & Bryant, 1998). They have been likened to 'an expert on the end of a phone', as they often obtain information by asking questions. At the end of the 'dialogue', the system should produce an identification that it can justify. These systems have an intelligent search strategy based on 'rules of thumb', which differ among systems (Edwards & Morse, 1995). Expert systems are able to handle uncertain and incomplete information. If the information supplied is inconsistent with the knowledge of the system, an uncertain conclusion or several competing conclusions can be reached. The most recent systems are capable of explaining their reasoning (Lacave & Diez, 2004). Liao (2005) provides a review of expert system methodologies and applications from 1995 to 2004. Liao (2005) distinguishes rule-based systems from knowledge-based systems. Rule-based expert systems represent information in the form of simple rules, such as IF-THEN. A knowledge-based system (KBS) takes the neural network approach (explained in section 2.1.7), trying to imitate human knowledge in computer systems.

To set up an expert system specific knowledge is placed into an expert system shell. The search strategy and 'intelligent behaviour' are provided by the shell itself (Edwards & Morse, 1995). Expert system shell packages, such as Babylon, CLIPS (C Language Integrated Production System) and Frulekit are available either for free or cheaply in a CD-ROM collection (e.g The Prime Time Freeware for AI CD-ROM collection costs $60). Expert systems are better than paper materials written by human experts; they apply rules rapidly and without bias and can quickly check for errors, and they are less intimidating and unlikely to suffer from information overload (Galbraith & Bryant, 1998). However, they may be less trusted by the general public than human experts, especially for critical decisions such as medical diagnoses.

Expert systems have the potential to ease the huge taxonomic workload. They have been developed for living objects (plants, mammals and micro-organisms), formerly living objects (fossils and fossil pollen), and non-living natural objects (minerals, geologic formations, soils) (Galbraith & Bryant, 1998). Galbraith, Bryant & Ahrens (1998) developed a prototype expert system for soil taxonomy (a complex and intimidating classification system), this was successful, with each of the data sets identified to the correct soil order within minutes. SPONGIA is another example, which helps to identify marine sponges; it can identify to any taxonomic level starting from class, depending on the information available (Domingo & Uriz, 1998). When faced with 82 sponge species SPONGIA obtained similar quality results to those of experts in Porifera systematics (Domingo *et al.,* 1999).

No expert systems of direct relevance to *Acacia* pollination have been developed to date. The closest are the 'pest management' systems of Mahaman *et al.* (2003) and Kaloudis *et al.* (2005). DIARES-IPM (Mahaman *et al.,* 2003) serves as a diagnostic and educational tool for Integrated Pest Management of Solanaceous crops. It includes the most economically important diseases, insects (both noxious and beneficial), and nutritional deficiencies that affect these crops, suggesting treatments to problems. The Kaloudis *et al.* (2005) system can identify more than forty distinct insects, either from some stage of their lifecycle or from the damage they cause to trees. Once an insect identification is completed, the system can recommend an appropriate treatment, aiming at reducing spread of insects and minimizing forest damage.

1.4.4 Automated Identification – an introduction

While computerised keys still require the user to compare subtle character states, taxonomic experience is not a problem with an automated system. Automated identification systems involve the application of general pattern recognition, in which an unknown is placed into a class on the basis of extracted features (Chesmore, 1997). Pattern recognition already has many applications including the recognition

of human faces (e.g. Turk & Pentland, 1991, Ekenek & Sankur, 2005), fingerprints (e.g. Tian *et al.*, 2004), palmprints (e.g. Connie *et* al, 2005) and handwriting (e.g. Gunter & Bunke, 2005). Automated identification can be either fully or semi-automated. When fully automated, no user interaction is needed but the system must be capable of highly reliable identification. Semi-automated identification may be more realistic; here the system prior-sorts, quickening the process, but leaves the final identification to a human operator. This may serve to reduce the human mistrust of computers, which has been seen as a potential barrier to Computer-Assisted Taxonomy (CAT) (Chesmore, 1997).

Although automated identification has barely progressed beyond the prototype phase, its potential is huge, both for biodiversity assessment and as a mechanism to overcome the taxonomic impediment to ecological studies. Multidisciplinary research is needed, because new applications can only be designed appropriately if biologists have some knowledge of computing and vice versa (Chesmore, 1999). It is also essential that the technological cost associated with CAT does not limit its global application. The availability of equipment and infrastructure in developing countries is a particular concern. However, in developing countries in which biodiversity assessment is a high priority, high levels of funding have tended to be made available for this purpose (Oliver *et al.*, 2000). Joint projects with tropical institutions, such as the implementation of DAISY at the Universidad de Costa Rica (UCR) (O'Neill, pers. comm.), will help ensure that applications are made widely available.

Gaston and O'Neill (2004) address the notions that automation is too difficult, too threatening, too different or too costly. The primary difficulties are threefold. Firstly, individuals of a given species may vary greatly in their morphology (due to genotypic variation, age, environmental conditions experienced or accidents); so a single name corresponds to a range of similar but different images. Secondly, closely related species may be very similar in morphology and detailed patterns may not be captured in, for example, digital images. Finally, the number of possible identifications may be huge; many taxonomic groups comprise tens of thousands of species all exhibiting a similar body plan (Gaston & O'Neill, 2004). While recognising very real technical obstacles Gaston & O'Neill argue that progress so far is encouraging and that vision and enterprise may be more limiting than practical constraints.

1.4.5 Limitations of Computer-Aided Taxonomy

When we consider the potential of CAT, we must also bear in mind its limitations. All tools must be accessible and effective when used by their intended audiences. Handheld computers have yet to rival the accessibility of a book in the field. The user retains the responsibility to check the identification

against a species description. As new species are discovered these tools will need to be regularly updated (as do paper keys). The Internet could take on this role (Edwards & Morse, 1995; Wheeler, 2003).

Computer-assisted taxonomy is still in its infancy but developing fast. With financial support and scientific commitment, it has the potential to provide pollination ecologists with the means to do routine identifications themselves and allow taxonomists the time to focus their efforts on the most challenging specimens and on species not yet described.

Chapter 2 – Automated identification and the DAISY project

2.1 General Concepts in Automated Identification

To understand the DAISY system some concepts that are central to automated identification must first be understood. In this section, data capture, image enhancement, image segmentation, feature selection, pattern recognition, nearest neighbour classification and artificial neural networks are introduced.

2.1.1 Data capture

Data capture generally involves images (e.g. moth wings in Watson *et al*, 2004; spider genitalia in Do *at al.,* 1999) or acoustics (e.g. grasshopper calls in Chesmore & Ohya, 2004; bat echolocation in Obrist *et al.,* 2004; Corncrake calls in Terry & McGregor, 2002). However, DNA barcodes (e.g. Tautz *et al.,* 2003; Hebert *et al.,* 2003), radar (e.g. Chapman *et al.,* 2003), infrared (e.g. Wythoff *et al.* 1991), flow cytometry (e.g. Boddy *et al.*, 2001) and movements (e.g. wingbeat patterns in Moore & Miller, 2002) have also been used. Different taxa lend themselves to different data formats. As this research used digital images I will concentrate on digital image capture.

Image analysis techniques have seen huge advances in recent years (Gaston and O'Neill, 2004). Appearance can be used to identify to species (e.g. wasps in Yu *et al.*, 1992; Weeks *et al.*, 1999b), to recognise individuals (e.g. sperm whales from fluke pattern in Huel & de Haes, 1996; cattle from their faces in Kim *et al.*, 2005) or to quantify genetic influences on phenotype (e.g. *Drosophila* wing venation in Houle *et al.,* 2003). Gaston and O'Neill (2004) tabulate examples of automated species identification based on morphological characteristics.

A digital image is a simplified representation of a three-dimensional object. Digital images can be monochrome, colour, infrared (IR) or thermal IR. The image is created by **sub-sampling** to produce a rectangular array of picture elements (or pixels), each with distinct colour properties. If **spatial resolution** is set to be high (e.g. 2272 X 1704 pixels) each pixel sub-samples a very small area so the sub-sampling is not visible. If the spatial resolution is set to be low (less than 640 X 480 pixels) each pixel sub-samples a much larger area and the image may appear 'blocky' (Fig. 2.1). A low resolution image has a much smaller file size so it is quicker and less-demanding to process. High resolution devices are also significantly more expensive than their lower resolution counterparts. If images are to be further sub-sampled by the automated identification system, as is the case when DAISY produces

normalised polar thumbnails of spatial resolution around 32 X 32 pixels (see section 2.3), the resolution of lower resolution devices is more than adequate.

Fig. 2.1 - Image of a moth on *Acacia brevispica* at high (100 pixels per cm, left), moderate (10 pixels per cm, middle) and low (5 pixels per cm, right) spatial resolution. As resolution decreases the image appears more blocky.

Thousands of subtly different colours are perceived by the human eye. Not all of these can be recreated in a digital image. **Brightness resolution** determines how finely colours are discriminated (e.g. 256, 4096 or 65536 grey levels in Weeks & Gaston, 1997). This can be specified before images are taken.

Digital image capture may be achieved using an analogue video camera and image grabber (e.g. Weeks *et al.*, 1999), a digital video camera, a digital stills camera (e.g. Watson *et al.*, 2004), or a flatbed scanner (O'Neill, pers. comm.). The scanner was found to be well suited to two-dimensional structures such as wasp wings, allowing many specimens to be imaged in a single scan (O'Neill, pers. comm.). Cameras may be mounted on optical (e.g. Hajdaoud, 2005) or scanning electron (e.g. Treloar *et al.*, 2004) microscopes or equipped with a macro facility, for imaging objects as close as 2cm (e.g. Watson *et al.*, 2004), such that virtually any image a taxonomist may observe may also be acquired by an image-analysis system (Weeks & Gaston, 1997).

2.1.2 Image enhancement

Once an image has been taken it can then be modified. Image enhancement operations aim to increase the visibility of certain aspects of an image (Russ, 1995). Only the most basic operations will be mentioned.

Brightness and **contrast** values can be rescaled so that colours of the image cover the full colour range of the display device, thus giving an image that is easier to interpret visually (Weeks & Gaston, 1997). Colour intensity can vary between images due to inconstant illumination. In **intensity normalisation** it is assumed that, as the intensity of the lighting source increases by a factor, each RGB component of each pixel is scaled by the same factor. The effect of this intensity factor is removed by dividing by the sum of the three colour components, such that:

$$r_{norm} = r \,/\, (r+g+b)$$
$$g_{norm} = g \,/\, (r+g+b)$$
$$b_{norm} = b \,/\, (r+g+b)$$

Heseltine *et al.* (2002).

In **histogram equalisation** the colour intensity at each pixel is changed so that a wider range of intensity values are used. This increases visual contrast. Often the colour intensity values of an image are clustered in a small range of values, so that a histogram of intensity has a large peak. For example, even though the integer range of [0, 255] was possible when Il Duomo was imaged, the image was dark, with the intensity values grouped towards zero rather than spread evenly across the range of values. After histogram equalisation the histogram of intensities are evenly spread and the features in the image are much easier to see (Fig. 2.2) (Nottingham University, 2005).

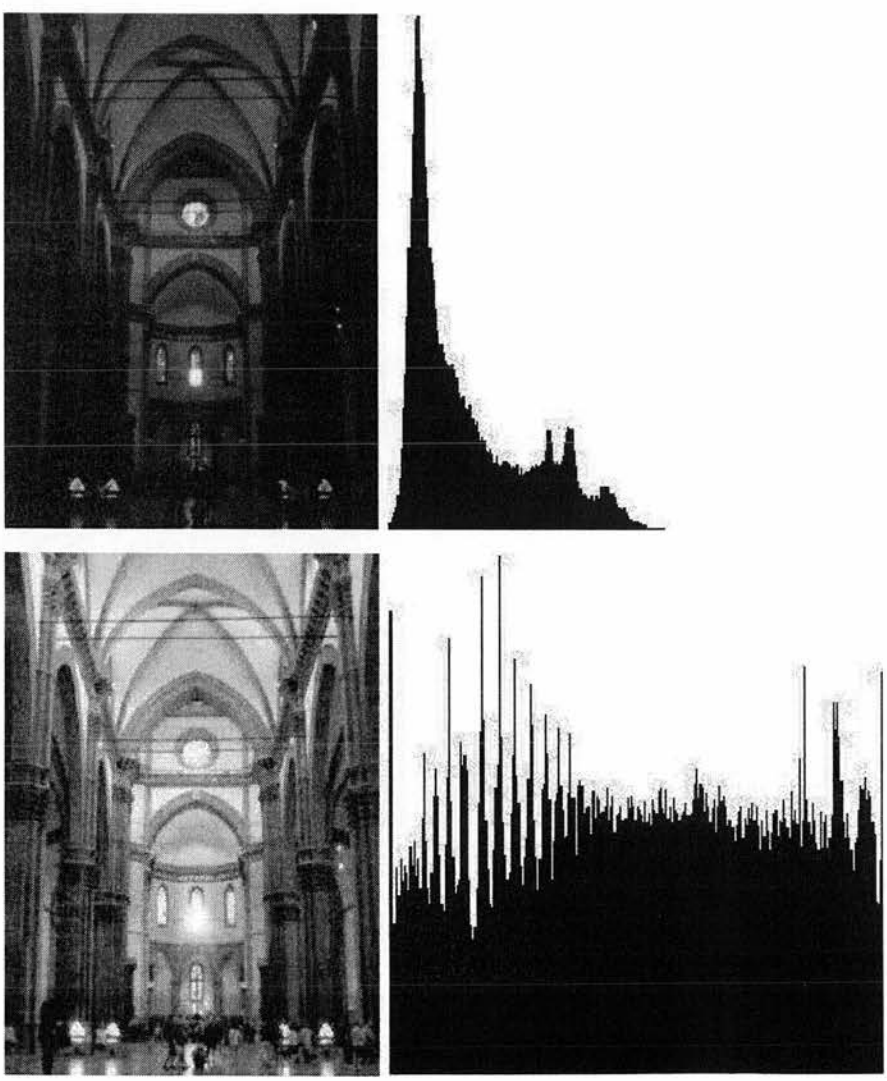


Fig. 2.2 – The interior of Il Duomo, Florence before (top) and after (bottom histogram equalisation. Nottingham University, 2005.

Images can be **flipped** about the x axis, **mirrored** about the y axis or **rotated**. The simplest rotation is 90°, either clockwise or anti-clockwise. If more precision is required an angle can be specified by typing an angle or by adding 'start' and 'end' lines. A rectangular region of an image can be cut out and retained by **box-cropping**. Images can be **re-sized** or spatial resolution decreased (it is not possible to increase resolution at this point). These operations can be performed using any standard graphics software.

2.1.3 Image segmentation

Before an object can be identified, it first has to be located within the image. This process is known as 'segmentation'. Human vision achieves this without conscious thought but it is a major challenge for computer vision. 'Pop-out' must occur against a background of distracters, difficult when many small objects are mounted together on a single slide (Fig. 2.3) (Culverhouse, 2005).

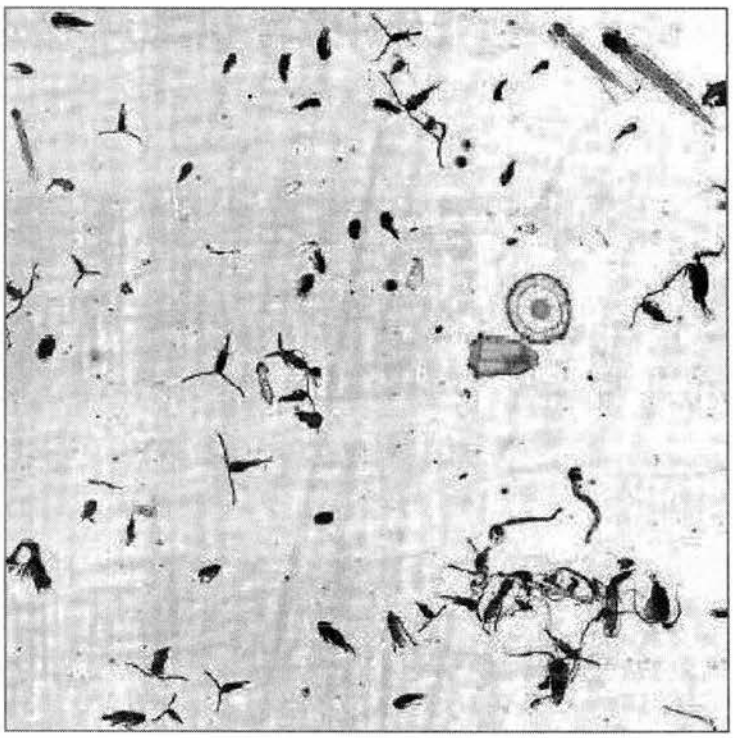


Fig. 2.3 – Segmenting zooplanton specimens in a seawater sample can be a slow and demanding manual task. Taken from Culverhouse & Williams (2003).

The simplest way to select an object is to manually draw a box around it and box-crop this region (Fig. 2.4). This is the approach adopted here for pollen segmentation and for wing venation in the Boxed method. While it is a relatively quick manual operation it is difficult to automate.

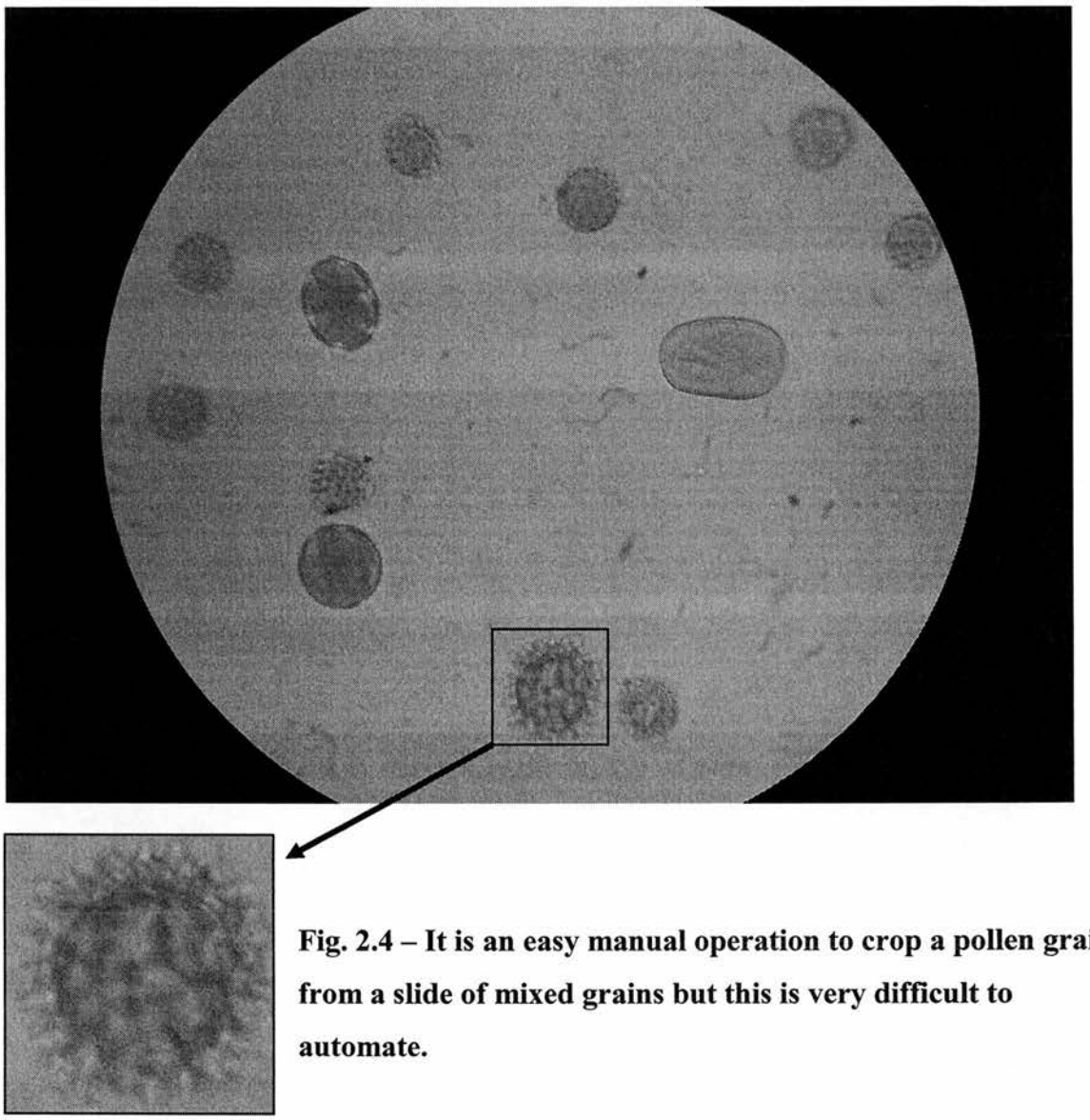


Fig. 2.4 – It is an easy manual operation to crop a pollen grain from a slide of mixed grains but this is very difficult to automate.

Culverhouse has incorporated automated segmentation as an important part of his hierarchical identification systems for phytoplankton (Culverhouse *et al.*, 1996; Culverhouse *et al.,* 2003; Wang & Culverhouse, 2004). The HAB-Buoy system (described at http://www.cis.plym.ac.uk/cis/projects/HABBuoy.html) aims to detect harmful algal blooms (HAB) in coastal waters. The underwater microscope can be connected to a buoy or mussel-producing raft and images everything in a filtered seawater sample, including detritus and plankton. Each object is assessed and rejected if not relevant. Harmful algae go to the DiCANN (Dinoflagellate Categorisation by Artificial Neural Network) natural object recognition software for identification.

When there is only one object in an image it is still worth segmenting it from the background. Firstly, it is very difficult to standardise the background when creating images. Watson *et al.* (2004) found that even if the same piece of coloured card was placed behind each moth imaged, the card colour looked different in different images, possibly due to subtle changes in illumination. If the background differs

between images it will provide non-informative 'noise', decreasing the likelihood that the images will be identified correctly. Secondly, separating the representation of an object from its background allows the standardisation of object size and orientation. This allows direct comparisons to be made between parts of objects.

Specialised computer algorithms aim to automate image segmentation. **Snake algorithms** (Tong *et al*, 2002; Yoon *et al.,* 2004) follow the outline of an object (the 'object envelope'), assuming that the direction that gives the smallest colour change is likely to be the continuation of the envelope. Such algorithms are being used increasingly in medicine and food technology. Saxena *et al.* (2002) extracted the shapes of tibiae and fibulae from amputee computer tomography data. Lindo *et al.* (2004) found that Active Contour analysis could recognise muscles from MRI scans of Iberian hams at different maturation stages and calculate growth rate as well as the traditional destructive method of weighing volumes.

Automated segmentation by snake algorithms may fail when faced with complex images, or situations where the background is very similar in colour to the object. The live moth images of Watson *et al.* (2004) are an example of this. The aim was to separate the right fore wing from the rest of the image. This wing was still attached to the rest of the moth, which was generally the same colour as the fore wing. This colour similarity caused snake algorithms to miss the envelope where the posterior edge of the wing overlapped the thorax and abdomen. Where the background is poorly differentiated from the area of interest it may be necessary to use **statistical edge detection** (Pedersen & Lee, 2002). These techniques are being developed but there is not yet a generally applicable method for automated segmentation in complex images.

Where automated segmentation is not feasible then the user must highlight the objects of interest. This repetitive task can be time-consuming, potentially undermining attempts to reduce the labour involved in routine identifications. It can also introduce errors as users become tired and less precise. Watson *et al.* (2004) attempted to quantify the impact of user precision. Three different users repeatedly highlighted a single fore wing of a hawk moth species using different (subjective) levels of precision. These efforts produced significantly different identification success (one-way ANOVA, $P < 0.0005$).

2.1.4 Feature selection

In this section the morphometric approach, subsampling, wavelet and other tranformations and principal component analysis will be introduced.

The **morphometric approach** to feature selection requires that images are supplemented with landmark points. Each training image is measured for the size of landmark features and the distances and angles between them, producing a grid of values for each species. The set of values for each image is a vector. Images for identification are landmarked and measured in the same way and their landmark vectors compared with those from the training set. This approach has been used for numerical description of insect wings (e.g. Yu *et al.*, 1992; Hajdaoud *et al.,* 2005) (Fig. 2.5).

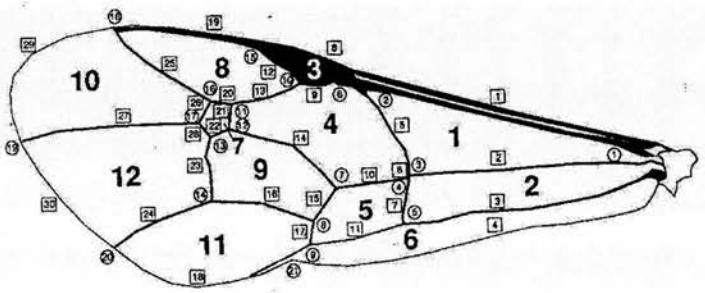
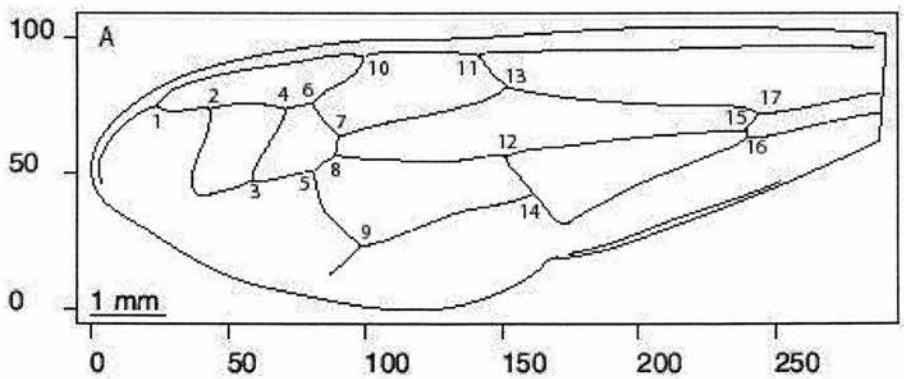


Fig. 2.5 – Yu *et al.* (1992) analysed a wasp wing as measurements between landmark features.

While the morphometric approach reduces a pattern to a sequence of numbers (ideal for computation) it ignores much of the available information and requires the user to be familiar with informative regions (Gaston & O'Neill, 2004). Hundreds of linear distance measurements would be needed to create a linear discriminant space (a decision space in which classes are separated by straight lines) for a moderately diverse group (such as planktonic foraminiferal species). This is problematic for two reasons. First, it can take a lot of time and effort to manually extract landmark coordinate data and distance measurements. Wing venation has been automatically extracted by the DrawWing software of Tofilski (2004) (Fig. 2.6) and by the Automated Bee Identification System (ABIS) (Hajdaoud *et al.,* 2005). Second, some taxa exhibit very simple adult morphologies so there are few landmarks to use as the basis for measurements (MacLeod *et al.*, in prep.).

Fig. 2.6 – Wing diagram of the wasp species *Dolichovespula sylvestris* generated by DrawWing. Tofiliski, 2004.



Sub-sampling is a very different method, requiring no user knowledge of landmark features. Instead the image is further sub-sampled (as in 2.1.1) to a smaller pixel grid. This lower level of spatial resolution better maximises the signal-to-noise ratio (MacLoed *et al.*, in prep.). The optimal number of pixels to make up this grid will vary from case to case, and may need to be determined empirically. The colour properties of each pixel in this grid provides a separate feature for analysis, so the 36 X 36 pixel grid of Fig 2.7 would correspond to a single point in 1296 (36 X 36) dimensional decision space (Heseltine *et al.*, 2002).

Fig. 2.7 – A 36 X 36 pixel grid to represent the 'moth on *Acacia*' image.

A disadvantage of the sub-sampling approach is that taxonomic features can be confounded by non-informative variation (e.g. variation in illumination and posture) (Watson *et al.*, 2004). These problems will be discussed in section 3.2

Wavelet transformation is another way to compress an image, preserving gross shape information while eliminating noise (e.g. high frequency data) and unnecessary details. This technology has been used in the identification of individual whales from fluke pattern (Huel & Haes, 1998). Daubechies / Gabor wavelet transformation is important to the SPIDA (SPecies IDentified Automatically)-web automated identification system of Russell *et al.* (2005), in which images of spider genitalia are identified better if hairs and spines are not pronounced (Ness, 2005). Wavelet transformation works as follows. Since an image generally contains structures of different sizes there is no single optimal spatial resolution for analysing them. Stéphane Mallet (the pioneer of multiresolution theory) wrote that, "a multiresolution decomposition enables us to have a scale-invariant interpretation of the image". A series of pictures represent the image at a series of resolutions, each twice as fine as the previous image. Wavelets encode the differences of information between two successive resolutions, i.e. the details that must be added to one image to obtain the image at the resolution twice as great. Wavelets are often used to compress information, as areas of an image that change very little can be ignored (Hubbard, 1998). However, in SPIDA-web the highest frequency resolution (representing very fine-scale changes) is discarded. This blurs the image such that 'noise' is reduced and general features, more

characteristic of its taxon, become apparent (Russell *et al.*, 2005)

Transformations may also be performed to each pixel to reduce inter-image variation due to variable illumination (intensity normalisation) and to increase contrast within the image (histogram equalisation). These processes were described in section 2.1.2.

The grid may also be transformed from **Cartesian to polar** format. The Cartesian system locates objects on a plane by measuring the horizontal and vertical distances from an arbitrary origin to a point. This is the most familiar coordinate system, illustrated in Fig. 2.8.

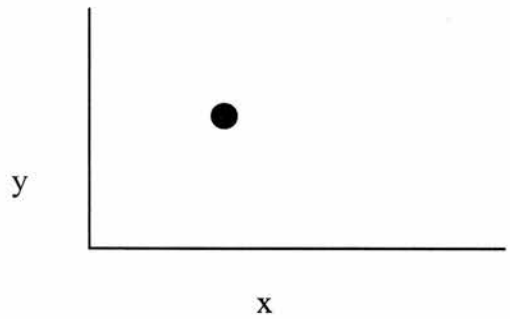


Fig. 2.8 – Two dimensions, x and y, in the Cartesian coordinate system.

The polar system (also known as the circular system) is defined by the origin, O, and a semi-finite line leading from this point, L. In terms of the Cartesian system, one usually picks the origin (0,0) to be O and the positive x axis to be L. In the polar system a point is represented by a tuple of two components (r,Θ).

- r (radius) is the distance from the origin to the point
- Θ (azimuth) is the counterclockwise angle between the positive x axis and the radial line.

The polar system is illustrated diagrammatically in Fig. 2.9

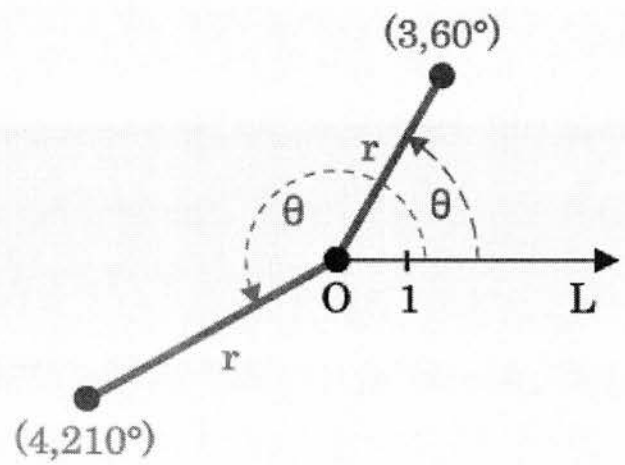


Fig. 2.9 – Diagrammatic representation of the polar (or circular) coordinate system. Taken from Wikipedia (http://en.wikipedia.org).

Simple algorithms are available to automate the Cartesian to polar conversion. This conversion allows an analysis to utilise spatially irregular regions of interest in addition to the more traditional rectilinear image boundaries. Once all transformations are complete the colour values of the final pixel set will form a vector for each image. These vectors can be analysed as before.

Principal Component Analysis (PCA) was made popular by the face recognition work of Turk & Pentland (1991). It achieves feature selection by dimensionality reduction (Haykin, 1999). Images of a single object class should map to nearby points in image space and images of different object classes should map to far apart points. PCA aims to reduce the dimensionality of the image space without reducing the variation spread. Each set of training images is taken and the average image and the difference of each training image from the average image is computed. This covariance matrix is used to determine principal components, such that the first principal component, p_1, explains the largest amount of variation, p_2 the second largest, and so on (Heseltine *et al.*, 2002). The average image and first five principal components for a face are shown as Fig. 2.10. This comes from Heseltine *et al.* (2002).

Fig. 2.10 – Average face image and first five principal component images. Heseltine *et al.* (2002).

A training image caricature can be created from these first few components, and a caricature for each object class projected into the image space (Heseltine, 2002). An unknown image is then projected into the image space using the same principal components and compared to the training caricatures Fig. 2.11). The class of the closest match is then given as the identification.

Fig. 2.11 – An unknown image and its image space projection. Heseltine *et al.* (2002).

PCA has been criticised for being computationally slow and error prone. It is slow because the caricatures have to be recomputed every time a new specimen is added to a training set (Gaston & O'Neill, 2004).

2.1.5 Pattern recognition

Pattern recognition is the process whereby a pattern is assigned to one of a prescribed number of classes. Humans are good at pattern recognition. When we receive sensory data from the world around us we are able to recognise the source of the data, often almost immediately and with no conscious effort. For example, we can recognise the face of a familiar person who has aged since our last encounter or recognise a familiar voice on a telephone line with a bad connection. Humans perform pattern recognition through a learning process; automated identification systems also need to learn (Haykin, 1999).

A pattern recognition system must first be trained. This is done by repeated presentation of input patterns along with the class to which each pattern belongs. Later, a new pattern is presented to the network, which belongs to the same population of patterns as those used to train the network. The network identifies this unknown pattern using information it has extracted from the training data.

2.1.6 Nearest Neighbour Classification (NNC)

There are many methods of statistical pattern recognition but the nearest neighbour rule has many benefits. It is simple to understand and achieves consistently high performance without *a priori* assumptions about the distributions from which the training examples are drawn (Oxford University, 2005).

Firstly, the simplest case is explained, that of two-dimensional, single neighbour NNC. The image for identification will be referred to as the 'unknown'. If two parameters, X and Y, are measured from all training images and from the unknown they can be plotted in Cartesian space and the unknown put into the same class as the most similar training image. This is represented in Fig. 2.12, in which the unknown vector, ?, is classified as a 2.

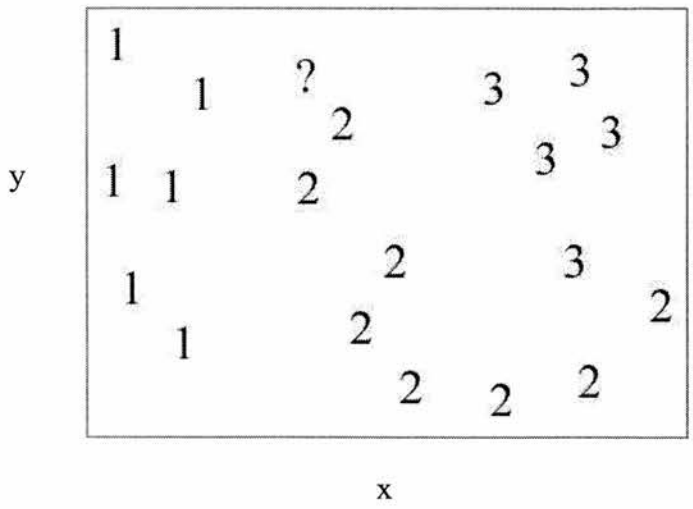


Fig. 2.12 – An unknown vector, ?, is assigned the label of the training set vector which is nearest. In this case, the unknown will be classified as a 2. Adapted from Barber (2001).

The key word in the strategy is 'similar' but similarity is a rather subjective notion. The simplest way is to quantify dissimilarity, based on the squared Euclidean distance between vectors, $(x - y)^2$ (Barber, 2001). There are also more complex ways to quantify dissimilarity such as the Mahalanobis Distance (Barber, 2001).

The decision boundary is the boundary in decision space, such that a decision as to the class of an unknown changes as the boundary is crossed (Barber, 2001). In Linear Discriminant Analysis (LDA) (the analysis generally combined with the morphometric method of feature selection) the boundary is a straight line (MacLeod *et al.,* in prep.). When dissimilarity is quantified by the squared Euclidean distance the decision boundary is piecewise linear, with each segment the perpendicular bisector of two datapoints from different classes. This is illustrated in Fig. 2.13 and Fig. 2.14.

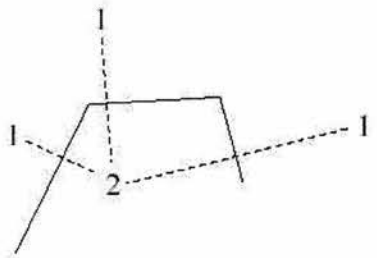


Fig. 2.13 – The decision boundary for squared Euclidean distance is the piecewise linear perpendicular bisector. Barber (2001).

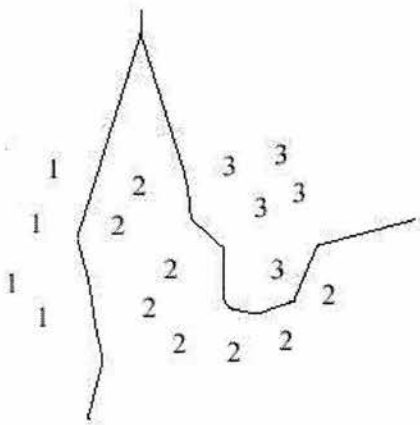


Fig. 2.14 – $(x - y)^2$ decision boundary for NNC with three classes. Barber (2001) .

So far just two parameters, X and Y, have been used, as this is easiest to visualise on paper. If more parameters are measured then the vectors can be more complex and decision boundaries formed in n-dimensional space.

There are two problems with the single-NNC approach.

1. Two or more training vectors with different class labels may be equidistant from the unknown. If one label is more common than the other then this class can be chosen. If there is no one most numerous class then this approach fails.
2. It is sensitive to outliers. An outlier is a 'rogue' data point with a strange label. If every other point close to this outlier has a consistently different label then the unknown should not be given the label of the outlier, just because it is the least dissimilar vector point.

Both of these problems can be overcome if more than one neighbour is included in the class decision. This approach is known as 'K-NNC', with K being the number of neighbours considered. It is useful to visualise K-NNC in terms of a hypersphere diagram (Fig. 2.15). A hypersphere is centred around the unknown. The radius of the hypersphere is increased until K neighbours fall within it. The class label is then given by the most numerous class in the hypersphere. In Fig. 2.15 the inner circle corresponds to single-NNC, the nearest neighbour, whereby ? would be classified as 1. If K-NNC is used instead, where K has been set as 3, reference is made to the outer circle. There are two 2's and only one 1 so the unknown will be classified as a 2. Using more than one neighbour means that the influence of outliers is out-voted and there is less likely to be equally dissimilar classes (Barber, 2001).

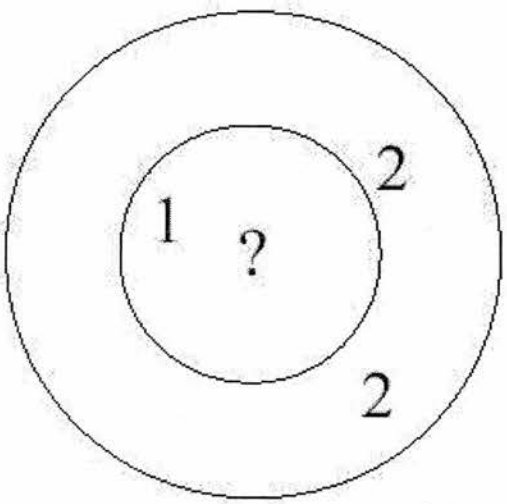


Fig. 2.15 – Hypersphere diagram in which the inner circle corresponds to single-NNC and the outer circle to *K*-NNC (in this case 3-NNC). From Barber (2001).

If K nearest neighbours are to be used, the value of K first needs to be decided. Larger values of K better reduce the effect of outliers but if K becomes very large (close to the overall number of training

points) then the classifications will all select the most numerous class. There is an optimal intermediate setting of K, which could be determined empirically.

K-NNC provides greater certainty than single-NNC that an identification is correct. If yet more certainty is required then '**coordination**' is a useful approach as coordination often provides over 95% certainty (MacLeod *et al.,* in prep.). Coordination is an output metric to quantify accuracy in this version of DAISY (see section 2.3). The number of neighbours considered is the '**coordination level**'. In K-NNC the K nearest neighbours voted. In coordination they all have to agree, i.e. be of the same class. If this class is correct then the unknown is said to have been identified successfully to that coordination level. The higher the coordination level the greater the certainty.

2.1.7 Artificial Neural Networks (ANNs)

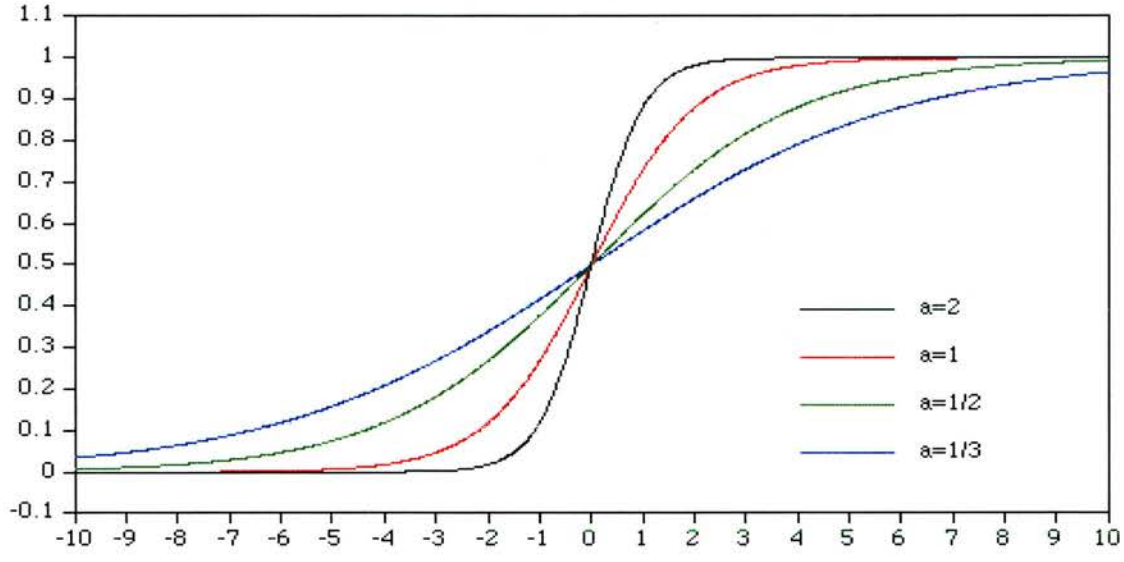
Neural computing provides an approach which is closer to human perception and recognition than traditional computing (Newman, 1998). ANNs are information-processing structures modelled on the massively parallel structure of the brain. Conventional computer algorithms follow explicit rules; neural networks 'learn' by being shown examples. Benefits of ANNs include generalisation, non-linearity, partial fault tolerance (PFT) and graceful degradation:

- The **generalisation** ability of an ANN is dependant on its architecture. An ANN with the correct architecture will learn the predictive task presented by the training set but also gather enough general rules to correctly predict outputs for unseen test set examples. Generalisation is essential for a classification system (Roadknight, 1997).
- ANNs made up of **non-linear** neurons (with sigmoid transfer functions, see Fig. 2.16) are able to perform non-linear discrimination, this is important as the physical mechanisms responsible for the generation of many input signals (e.g. speech) are inherently non-linear (Haykin, 1999; Chesmore, 1997).
- **Partial fault tolerance** (PFT) is due to distributed memory. Memory is distributed over many units so if a few units are destroyed or their connections are altered the behaviour of the network as a whole is only slightly degraded. PFT can be obtained by redundancy (replicating a smaller network), by design (specific training to increase PFT) or a combination of the two (Tchernev, 2005).
- If damage continues and the network can no longer function, it will reduce processing power gradually. This is known as **graceful degradation** (Weeks & Gaston, 1997). Graceful degradation makes neural computing systems well suited for applications where sudden failure of control equipment would be disastrous (Newman, 1998).

The human brain contains about 10 billion (10^6) neurons and 60 trillion (10^{12}) connections (Shepherd & Koch, 1990). Each is a relatively slow information processor, operating in the millisecond (10^{-3} s) time range, five to six orders of magnitude slower than silicon logic gates (10^{-9} s) (Haykin, 1999). The brain achieves complex tasks quickly because of the vast numbers of neurons, the complex interneuron connections, and the way many simple operations are carried out simultaneously (Weeks & Gaston, 1997).

The connecting links of ANNs, analogous to synapses, are the **nodes**. Each node receives weighted inputs, which are summed to give its internal activation level. Once a node is activated it modifies the summed input by a **transfer function** and passes the resultant signal on to other nodes.
The sigmoid transfer function, the graph of which is s-shaped (Fig. 2.16), is by far the most common form of transfer function used in ANNs (Haykin, 1999) as it allow non-linear relationships (Montague & Morris, 1994).

Fig. 2.16 – An example of a sigmoid function showing the variable slope parameter, a. From http://astronomy.swin.edu.au/~pbourke/analysis/sigmoid/sigmoid1.gif



Individual nodes are connected to form a highly structured network. Knowledge is bound up in the arrangement of nodes and interconnection weights (Weeks & Gaston, 1997). The learning algorithms (rules) used in the design of neural networks are structured. There are three fundamentally different classes of network architecture.

- A **single-layer feedforward network** is the simplest form of layer network, having just an input layer of source nodes and an output layer of computation nodes.
- A **multilayer feedforward network** has one or more hidden layers. The architectural graph in Fig. 2.17 illustrates the layout of a multilayer feedforward network with a single hidden layer. This example would be referred to as a 10-4-2 network because it has ten source nodes, four

hidden neurons and two output neurons (Haykin, 1999). Hidden neurons intervene between the external input and the network output in a useful manner, may evolve to function as feature detectors (Boddy & Morris, 1997) and may enable the network to extract higher-order statistics, giving a global perspective despite local connectivity (Haykin, 1999).

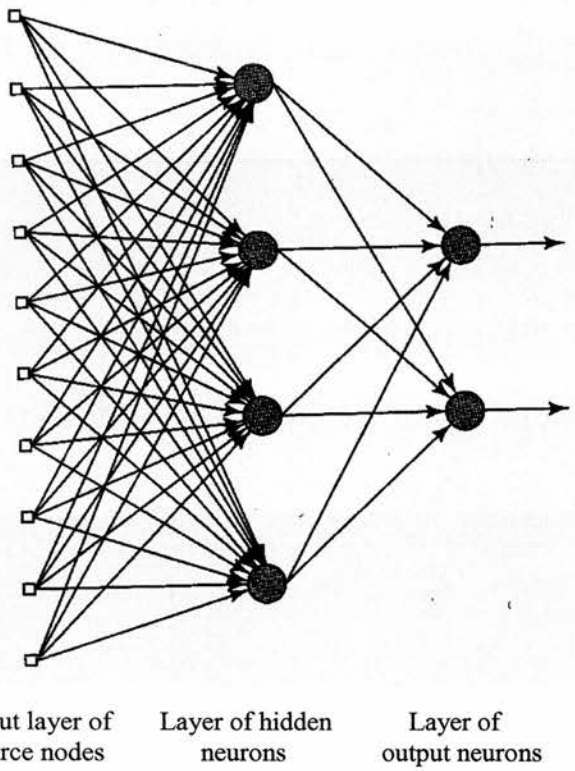


Fig. 2.17 – Architectural graph of a 10-4-2 multilayer feedforward network. Signal flow progresses from left to right on a layer by layer basis.Taken from Haykin (1999).

Data from the hidden layers propagate through to the output layer, where the output nodes learn to associate particular features with the groups that produced them (Boddy & Morris, 1997). If there are only two output neurons these may give 'yes' or 'no' answers, so each taxa needs its own network. This is the approach adopted in the SPIDA-web system of Russell *at al.* (2005) (Ness, 2005). If there are more than two output neurons they may each correspond to a possible taxa. The output with the largest number is taken as the identification.

- A **recurrent nctwork** has at least one feedback loop. These loops can have a profound impact on the learning capability of the network (Haykin, 1999).

Feedforward networks are often trained using the **error back-propagation algorithm**. This is a **supervised learning** technique. Input layer nodes are repeatedly presented with randomly selected input patterns. At each presentation the network's output is compared to the desired output and an error produced. (Weeks & Gaston, 1997). The output neurons produce an error signal which propagates backwards (layer by layer) through the network. The network neurons compute this error signal with an error-dependant function, updating the network's weights (Haykin, 1999). Training continues until the actual output is sufficiently close to the desired output. Network training is usually computationally

intensive, but new data can then be processed rapidly (Weeks & Gaston, 1997). New classes can only be added if the network is retrained, so a large and changing number of classes will result in high computational load (O'Neill *et al.*, 1997). This has been recognised as a major limitation of the SDIDA-web system of Russell *et al.* (Ness, 2005). Retraining can take as much as 30 hours for a two species, 2000 image dataset (MacLeod *et al.*, in prep). One way to resolve this is to employ a hierarchy of networks. Boddy *et al.* (1994) used a natural hierarchy to identify 40 species of phytoplankton, with the first network assigning specimens to a major phytoplankton group and the second assigning to species. Similarly, the SPIDA-web system (Russell *et al.*, 2005) classifies first to genus then to species.

Another solution to a nonstationary environment is the use of **Unsupervised ANNs**. In unsupervised ANNs there is no external critic to oversee the learning process. **Adaptive identifiers** change their synaptic weights so that their decision boundaries better reflect each new data point (Haykin, 1999). In **positive enforcement learning** when a new image is identified correctly (i.e. with a high level of certainty or confirmation from the user) **adaptive identifiers** alter the decision boundary for that class so it better encompasses the new data point (Haykin, 1999). In **negative enforcement learning** the user declares that an identification was incorrect so the decision boundary is shifted away from that data point.

A **Dynamic Neural Network (DNN)** has no defined training phase as every pattern it encounters is accepted as training input (**'perpetual novelty'**). A DNN has been described as a 'living network' (Lang, 2000). From the start, the network grows to meet the learning needs of the input data. Once no (or few) new patterns are being presented the network is allowed to settle. A series of 'virus' algorithms may then attack parts of the network to make it more efficient, such as a node virus removing unused nodes (Lang, 2000).

The **Kohonen Self Organising Map (SOM)** (Kohonen, 1989 and 2001) is one of the best known DNNs. It was inspired by the way in which human sensory impressions are neurologically mapped. The neurons are arranged in an *n*-dimensional grid (where *n* is usually two). Data is compressed, so highly dimensional data may be represented in a much lower dimensional space. Each competitive unit corresponds to a cluster, the centre of which is called 'the focus'. The SOM arranges these clusters so that any two similar clusters are spatially close. When an input is presented to the network its nearest focus is found. All the nodes of the nearest competitive unit (within a neighbourhood radius) are moved closer to the input. The distance each node is moved depends on the learning rate and the distance from the focus. The neighbourhood radius and learning rate are set by the user, it is normal to start with large values and decrease them during training. Problems could occur where a focus occurs near the edge of a SOM, as some of the neighbour nodes disappear off the edge of the network. This problem can be

overcome by making the SOM 'wrap around' on itself, creating a doughnut-shaped construct instead of a flat plane (Lang, 2000). Kohonen (2001) notes that SOM are not intended for pattern recognition but for clustering, visualisation and abstraction.

Lang was inspired by SOM to create the **Plastic Self Organising Map (PSOM)** (Lang, 2001; Lang & Warwick, 2002; Lang, 2005). The main differences between PSOM and SOM are that PSOM are **free-floating** and are able to improve network efficiency by **'pruning'** any unused links and nodes. At all times PSOM are enlarging and shrinking their topology to better reflect the input data. As the network gets more developed (i.e. at higher iterations) the similar nodes become grouped into distinct clusters (Fig. 2.18). The classification in a standard PSOM is similar to NNC with squared Euclidean distance to quantify similarity (see section 2.1.6). A PSOM is **able to generalise with noisy signals**. It assumes that every input is a valid input corrupted by noise, learning everything it encounters. It will sometimes learn pure noise. The difference between patterns and noise is recurrence. When the noise fails to recur it is pruned from the system. A real world example of PSOM is in radar pulse monitoring, where a system needs to detect new radar sources, classify them and forget them when the radar source leaves the system (Lang, 2005).

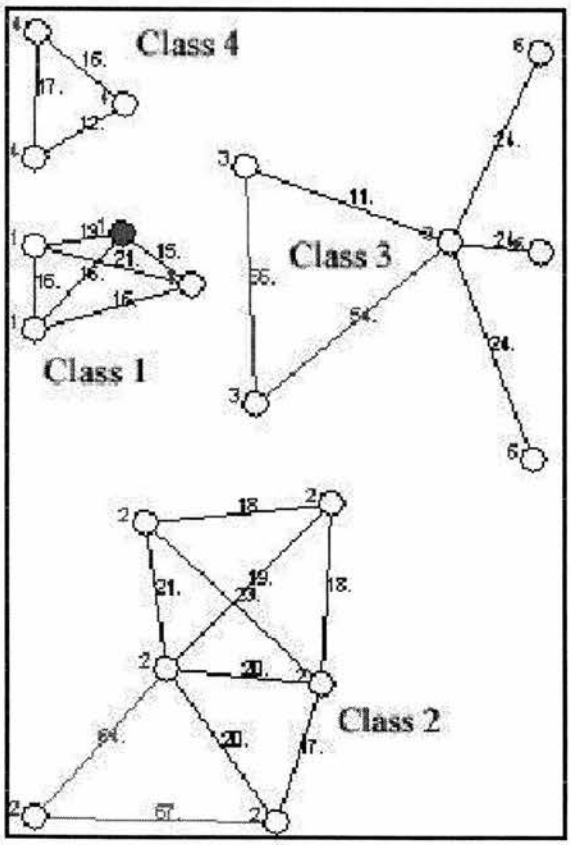


Fig. 2.18 – After 545 iterations this PSOM network has broken into four clusters that can be classified. Lang & Warwick, 2002.

If PSOM are to be used in automated identification then it may be more desirable for all nodes to be maintained, in this case the pruning function can easily be turned off. The PSOM would initially be shown the training images, allowing it to settle into a topology suited to these patterns. However,

training would continue for as long as the system is used, as every new input (i.e. unknown for identification) causes the system to develop. Classifications can be made by freezing the topology in its current position.

Unsupervised ANNs tend to be **modular** (i.e. a separate training pool for each species), so that when extra images are added only the pools they relate to need recalibration. Even within a given species training pool it is only the part of the network in the immediate vicinity of the new training image which is affected.

Modularity makes possible **manifold reduction**. Some regions of an image will be more informative than others. Manifold reduction is a stochastic optimisation algorithm that runs in the background. It seeks out an optimal image region that is distinctive for a particular species. The weighting (i.e. importance) of pixel values in this region is increased to recognise this.

2.2 DAISY history

DAISY (Digital Automated Identification SYstem) was the brainchild of insect biologists. When Ian Gauld (expert in ichneumonid wasps, NHM) and Kevin Gaston (ecologist, University of Sheffield) found themselves stranded in a Costa Rican airport they discussed at length how the Japanese would overcome the taxonomic impediment, using pattern recognition software to automate taxonomy. That discussion produced the seminal idea of a computer based automated identification system. Weeks & Gaston (1997) discussed the challenges facing taxonomy and the potential contributions of image analysis and neural networks to this field. For the first 18 months of the DAISY Project, Paul Weeks prepared specimens while Mark O'Neill designed the codes, working jointly on testing the system.

"Our objective at the start of this research programme was to use computer vision to develop an automated system to compare images of insect wings in order to be able to mechanize the identification process and remove observer error."

Weeks *et al.*, 1997

The first journal publication of the DAISY Project was Weeks *et al.* (1997). Weeks *et al.* (1999a) went into greater computational detail. Early development of DAISY was based on principle component analysis (PCA), the approach taken in face detection (Turk & Pentland, 1991; Pentland *et al.*, 1994) (explained in section 2.1.4). In this case, each caricature was constructed using the first five to ten principal components. A threshold certainty value served to ensure that 'outsider' species, excluded from training, were not erroneously identified as one of the training species.

The use of PCA broke away from the traditional landmark approach (e.g. Yu *et al.*, 1992), where distances and angles are measured between features, instead working directly on the sub-sampled pixels of an image. There are no landmarks to mark, so the sub-sampling approach allows greater automation. Information can be derived from the statistical structure of the whole image, eliminating the difficult problem of feature selection. However, PCA is more sensitive to variation 'noise' in illumination and pose. Weeks *et al.* (1997) avoided this problem by working exclusively with insect wing images, which were two-dimensional and less likely to present pose-related visual recognition problems. The study organisms were five Costa Rican species of the ichneumonid wasp genus *Enicospilus*. Ichneumonids are notoriously difficult to identify but species are known to differ in wing venation. Gauld had recently revised this particular genus (Gauld, 1991) so large numbers of specimens from a variety of localities were available for examination. Yu *et al* (1992) had achieved 100% accuracy when they separated five ichneumonid species using the landmark approach, so this selection would allow some degree of comparability.

Fifty specimens of each species were imaged. Fifteen of each were used to train the system and the remaining 175 images were used as unknowns. The mean level of correct identification was 94%. When the misidentified 6% were re-imaged and re-analysed half were then identified correctly. In blind tests using only wing slides, Gauld (an expert in ichneumonid identification) achieved a lower rate of accurate identification than DAISY. However, this comparison is unrealistic as an expert would also make use of non-wing features when making an identification.

While initial results were promising some limitations were apparent. Weeks *et al.* (1997, 1999a) recognised that screening specimens against a large number of classifiers would be computationally intensive and suggested a hierarchical approach including genus classifiers. They also acknowledged that their automated approach continued to place significant demands on the user, as wings have to be slide mounted to a high standard, aligned prior to image capture and the images manually edited to increase information content.

O'Neill *et al.* (1997) described two main ways in which the PCA in the DAISY classifier differed from the face recognition PCA of Turk & Pentland (1991). Firstly, a different function was used to correlate an unknown image with its principal component projection. Turk & Pentland (1991) used a simple distance function. The DAISY classifier used Kendall's rank order correlation (Kendall's Tau), a non-parametric statistic, to measure association, the statistic of which, *T*, ranges from –1 (perfect negative correlation) to 1 (perfect positive correlation) (Dytham, 1999). Only later was it realised that Kendall's Tau would process larger images (> 16 X 16 pixels) too slowly to be practicable. Kendall'd Tau is an order of N^2 algorithm (O'Neill, pers. comm.), processing effort is the square of the input so scaling problems were inevitable. Secondly, Turk & Pentland (1991) put all their faces in one big training set, which had to be recomputed every time a new face was added. The DAISY system was modular, with a separate classifier for each object (i.e. species), so one object could be modified without recomputing the entire set and new objects could be added without altering any existing classifiers.

Weeks *et al.* (1999b) introduced a second set of study organisms to DAISY, biting midges (Diptera, Ceratopogonidae), and a substantially larger species pool of 49. Twenty specimens of each species were imaged. When 11 of each were used in training and nine treated as unknowns 86% of species were identified accurately.

The simplest measure of identification success is first-past-the-post (FPTP), i.e. the proportion of cases in which the closest match is correct. While FPTP is a useful metric for DAISY it is poorly suited to applications in which a high level of certainty is vital. Weeks *et al.* (1999b) explored accuracy metrics

with different levels of certainty. The first accuracy metric required greater certainty, as it demanded a specific winning margin. The test image was only deemed to have been correctly identified if its correlation with the closest species classifier was a pre-determined magnitude greater than its correlation with any other classifier. Increasing the required winning margin decreased the number of correct identifications but increased confidence in these identifications. The percentage of correct identifications dropped from 86% to 60% when a winning margin of just 0.05 correlation units was stipulated, confirming that the wings were indeed extremely similar. The second performance monitor demanded less accuracy than FPTP. A test image was deemed to have been correctly identified if the correct species classifier was above a certain rank in the list of closest matches. Identification success increased from 86% to more than 90% when the correct species only had to be in the top three matches. This showed that DAISY had considerable potential to reduce the number of possible identities from 49 to very few, eliminating most species as 'improbables', hence suggesting that DAISY could function effectively as a screening engine.

Some apparent limitations were cause for concern in Weeks *et al.* (1999b). First, as the number of species classifiers increased performance decreased, the addition of species leading to increasing overlap of characters; this may have been an artefact of the very difficult 'species cloud' datasets tested (O'Neill, pers. comm.). Second, the narrowness of the 'winning margin' for a typical correct identification suggested that difference between a correct and an incorrect individual was very small, which was considered an undesirable property for an identification system.

Any automated system to facilitate the routine identification of a group of insects needs to be expandable to accommodate a large number of taxa (Gauld *et al.*, 2000). While DAISY had performed well in small trials, it was apparent that principal component analysis would lead to huge computational loads in larger implementations. Nearest neighbour classification (NNC) (Alexander, 1984; Lucas, 1997) works better at large scales than PCA. The new version of DAISY described by Gauld *et al.* (2000) could classify either by PCA or by NNC.

The unpublished Masters project of Pajak (2001) evaluated the merits of "skeletonisation". Skeletonisation involved the "pulling out" of the veins of the wing as white lines on black. This was intended to accentuate the veins but led to some blurring. His results suggested that skeletonisation made no significant difference to DAISY performance. Pajak (2001) also compared two measures of similarity in NNC, Nearest Vector Difference (NVD) (the squared Euclidean distance between vectors, explained in section 2.1.6) and Cross Correlation (CC) (described as a standard correlation coefficient between the pixel matrices of the two images). Again his results were inconclusive.

O'Neill *et al.,* (2002) detailed the open source 'P3M' computing environment on which DAISY is built. P3M enables DAISY to make use of parallel machines and computing clusters to achieve both speed and accuracy (P3M will be covered in greater depth in section 2.3). This is also briefly discussed in Gaston & O'Neill (2004), a review of automated species identification and the challenges it faces.

By the time of work by Watson *et al.* (2004), DAISY was using modified CC to quantify NNC similarity, as influenced by recent work on plastic self-organising maps (Lang & Warwick, 2002). This paper aimed to take DAISY out of the laboratory and into the field. British butterflies had already been identified successfully from wing scale pattern but previous trials had involved professionally pinned or wing mounted specimens, all undamaged and imaged with optimal lighting. In the study of Watson *et al.* (2004), live moths were allowed to settle into a normal resting posture, worn moths were included, and lighting was inconsistent. The mean accuracy of identification for the 35 species (using FPTP as the success metric) was 83%.

MacLeod *et al.* (in prep.) compares the morphometric approach (described in section 2.1.4) to the most recent version of DAISY for the identification of 202 planktonic foraminifa specimens representing seven modern species. In the morphometric approach, 11 landmark features were added to training and unknown foraminifa images, landmark measurements from the training set plotted in *n*-dimensional space, decision boundaries set by distance-based Linear Discriminant Analysis and the unknowns identified according to those decision boundaries. DAISY identification relied heavily on unsupervised neural networks (see section 2.1.8). It is this version of DAISY that is described in section 2.3. The DAISY results were superior to those obtained through the morphometric approach both in terms of the raw (FPTP) numbers of correct identifications (0.99 vs. 0.91) and in terms of the ratio of well supported (>95% certainty) identifications (0.93 vs. 0.56).

O'Neill *et al.* (2005) looked ahead to DAISY operation on wireless hand-held machines, with a Java-based GUI (graphical user interface) allowing images to be taken using palmtops and mobile phones. This poster also emphasised that DAISY is very versatile software that need not use digital images of morphology, giving the example of SDS (Sodium Dodecyl Sulphate) protein gel identification of hawkmoths.

Other projects currently using DAISY include:

- The EU **ALARM** (Assessing LArge scale Risks for biodiversity with tested Methods) project (Simon Potts, George Else, Paul Williams and Andy Polaszek) may be using DAISY for the routine identification of the European bee fauna.
- The **Department of Paleontology** of the NHM (Norman MacLeod and Stig Walsh), is using DAISY as a tool for routine identification of forminiferal faunas. This has significant

commercial potential as fossil forminiferal faunas are used by the oil industry to identify potential oil bearing strata.

- The **University of Costa RICA** (UCR) (Daniel Briceno and Paul Hanson) is in the process of obtaining funding to install DAISY with a view to using it a *screening tool* for agricultural pests (e.g. *Anastrepha* sp.) and invasive plant species. The UCR system may also be used as a general purpose *screening tool* for plants, insects and other invertebrates. (O'Neill, 2005).

2.3 How DAISY works

Different approaches in computing provide different advantages and disadvantages. These approaches tend to be discussed as though they are mutually exclusive but this is not always true. DAISY combines cluster with parallel computing and supervised with unsupervised neural networks, reaping the benefits of each and compensating for the weaknesses.

2.3.1 The computing system used

It is important that DAISY is **portable**, able to function on all the common operating systems (Linux, Apple Mac and Windows), so the system has been implemented using the Portable Operating System Interface (POSIX) environment. POSIX, produced by the Institute of Electrical and Electronics Engineers (IEEE), is the open operating interface standard accepted world-wide. (Lynuxworks, 2005). A POSIX compliant application programming interface (API) is offered by all modern versions of UNIX (including the MacOS X operating system used by the Apple Macintosh). A POSIX API (Cygwin) is also available for the Microsoft Windows (NT/XP) family of operating systems. The version of POSIX used is the **POSIX.1b environment** (Portable Application Standards Committee, 2005); this differs from POSIX.1a and POSIX.1c in having real-time extensions (Lynuxworks, 2005).

DAISY was initially designed for use on UNIX systems. UNIX server systems tend to be **fast**, so a standard readily implemented on UNIX has the potential to process large numbers of Internet-based identification queries in a reasonable time (< 1 minute per enquiry) (Watson *et al.,* 2004). While the MS-Windows operating systems have less power to process large datasets quickly they are better known to non-specialists, making DAISY **accessible** to a wider audience (O'Neill, pers. comm.).

DAISY is programmed in the **ANSI** (American National Standards Institute) **C** programming language. ANSI recognises POSIX (Lynuxworks, 2005). ANSI C is also one of the most commonly used languages, so this choice also promotes portability.

The **P3M** distributed programming environment (O'Neill *et al.,* 2002), used by DAISY, combines the P3 distributed programming environment with MOSIX (**M**ulti-computer **O**perating **S**ystem for UN**IX**) extensions. Both P3 and MOSIX are **open source** software, available without charge on the understanding that programmers will use them, improve them and make the improved versions available for free.

The **P3** distributed programming environment allows clusters of low cost computer hardware, such as networked PC's, to work in parallel. This cluster of computers (known as nodes) functions like a single computer with multiple processors. This parallel computing makes DAISY seamlessly scalable and able to process very large data sets, considering thousands of taxa in a short period of time.

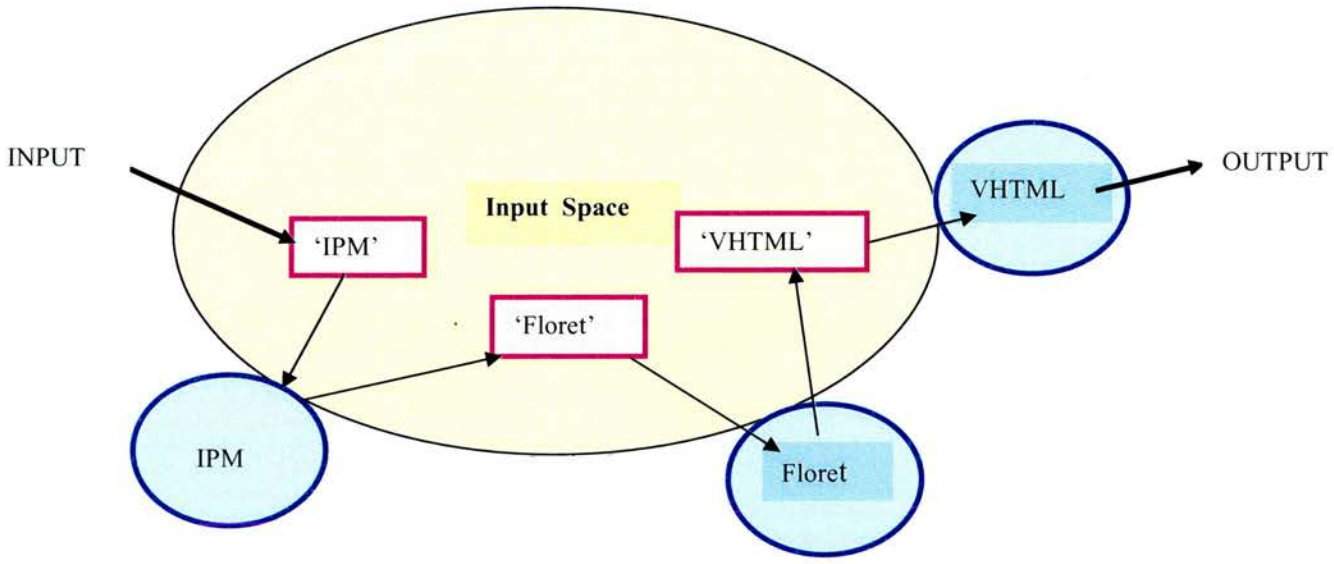
MOSIX was similarly developed to manage a single cluster of computers, e.g. those in a university department. Recently, it was extended with new features and these **MOSIX extensions** have been used for DAISY. MOSIX extensions make independent clusters of nodes run as a federated grid of cooperating nodes. These cooperating clusters can be anywhere in the world and their operators need never meet. The goal of a MOSIX grid is to allow owners of such nodes to share their computational resources from time to time, while still preserving the autonomy of each user to disconnect their nodes from the grid at any time. The sharing of computational resources gives huge processing power, so tens of thousands of taxa could be considered in seconds. For further details of the MOSIX project see www.mosix.org.

P3M enables DAISY to make optimal use of both parallel machines and computing clusters in a dynamic fashion (Gaston & O'Neill, 2004). Aspects of parallel computing enable DAISY to handle very large tasks in a reasonable time; the clusters ensure that quality remains high. The distributed processing of P3M has the added benefit of homeostatic mechanisms, making DAISY fault tolerant and able to make optimal use of the resources within the environment in which it finds itself. An application will not stop running as a result of hardware failure; it will fail gracefully and simply run more slowly (Watson *et al.,* 2004).

P3M also provides new ways to build networks of cooperating identifiers using methodologies derived from biological systems. Inter task communication is modelled loosely on the **lock-and-key** mechanisms of protein-protein signalling networks in eukaryotic cells. Information to be processed by the four DAISY components is tagged with key codes. Components sample an input space hundreds of times a second, they have the locks to read and process those data and if necessary re-tag them for downstream processing by other DAISY components. A basic schematic diagram of the DAISY

processing system is given in Fig. 2.19. The processor units will be explained in section 2.3.2. Any component of DAISY can be multi-threaded, i.e. use cooperating processors, running on different computers, and this speeds up their performance.

Fig. 2.19 – A basic schematic of the DAISY processing system (first used in Watson et al., 2004). Tagged images (rectangles) are taken in, processed and expelled from the processor units (circles). In the real system many processor units of each type may work in parallel.



2.3.2 Structural roles in DAISY

The four basic components and their roles are:

1) **DAISY Front-End (DFE)** - Digital images are inputted, modified and the region of interest selected.
2) **IPM** - The region is transformed and sub-sampled to increase information value and minimise noise.
3) **Floret** - Pixel grids are compared and an identification made.
4) **Virtual HTML (VHTML)** - The identification is displayed as a webpage and links made to relevant resources.

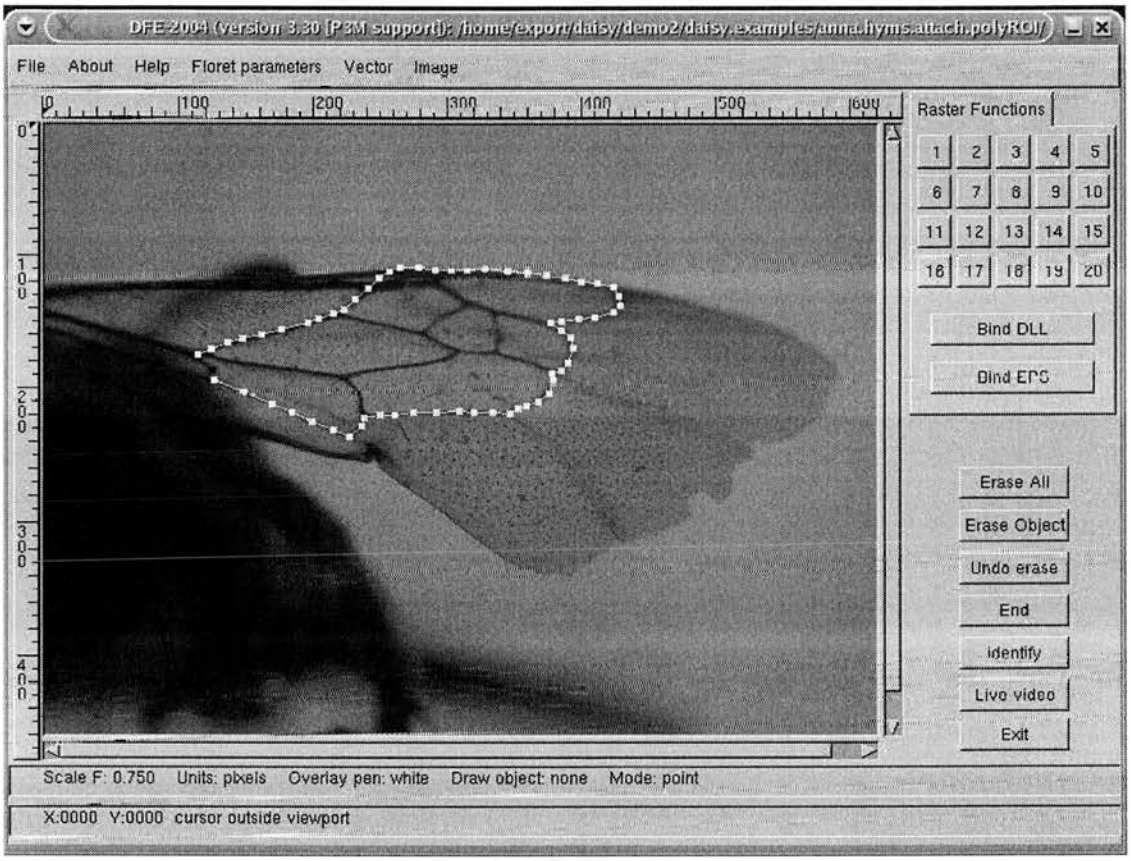
DFE and VHTML are the interface components; IPM and Floret work behind the scenes.

DFE (DAISY Front End)

DFE is a graphical user interface, implemented using the GTK+/Gnome X toolkit (Pennington, 1999). A screencapture is shown in Fig. 2.20. It is through DFE that a user instructs the other components to build training sets and perform identifications. It has some of the functionality typical of a graphics package, e.g. rotation and zoom. However, its main function is the selection of a

region of interest (ROI), either rectangular or polygonal (polyROI). A polyROI is added by evenly spacing landmark points around the edge of the region. DFE tags the ROI region either as "training" or "unknown" imagery; the tagged imagery is written to the input space (Fig 2.19) and recognised for processing by IPM.

Fig. 2.20 – A screencapture from DFE showing a polygonal region of interest (polyROI). The wing cell envelope is drawn by adding landmark points, which are automatically joined by straight lines.



IPM

In IPM, the polyROIs are standardised in size and pose. Each image is sub-sampled to a small pixel grid. This has an empirically-determined optimum resolution to maximise the signal to noise ratio. The default grid size is 32 X 32 pixels but some patterns are better processed as smaller (min. 20 X 20) or larger (max. 80 X 80) grids.

Pixel-based transformations aim to eliminate problems caused by variable lighting and poor contrast. Intensity normalisation aims to standardise the colour intensity of images taken with inconsistent illumination. Histogram equalisation increases image contrast. These have been explained in section 2.1.2. Each normalised pixel grid is then transformed from a Cartesian to a polar format (see section

2.1.2), producing a normalised polar thumbnail (NPT). For en examples of a NPT see section 3.3.1, Fig. 3.18. These NPTs are tagged for identification by Floret.

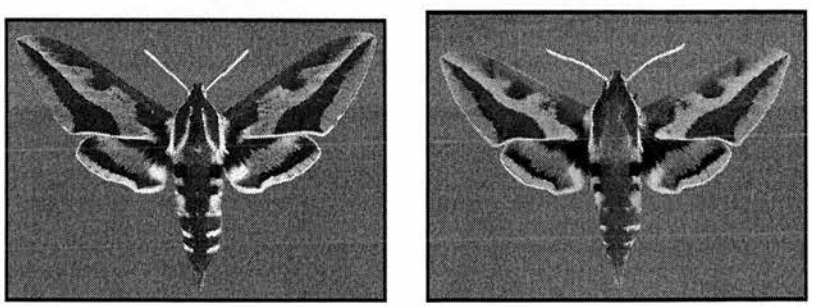
Floret

Floret is the Artificial Neural Network component of the DAISY system. Supervised and unsupervised ANNs have been introduced in section 2.1.7. DAISY has a **hybrid approach**, combining unsupervised foreground training with supervised background training. While **Plastic Self Organising Maps (PSOM)** (Lang & Warwick, 2002; Lang, 2005) (see section 2.1.7) run in the foreground, a set of background stochastic optimisation processes adjust training set composition and pixel parameters to maximise information content (these are coded but, as yet, minimally tested). These background processes are similar to the training of supervised ANNs. Effectively, this means that DAISY trains in a supervised manner when it has nothing else to do, rather like a human using spare time for study. When an identification is required it comes on-line and uses its best training set to date.

DAISY's ***n*-tuple/PSOM** foreground system responds well to the modelling of non-linear regions, which tend to be problematic to morphometric and Principal Component Analysis approaches. It also has the advantage of being **modular**, reducing the scale of recalibrations and making possible adaptive identifiers and manifold reduction (the advantages of modularity have been discussed in section 2.1.7).

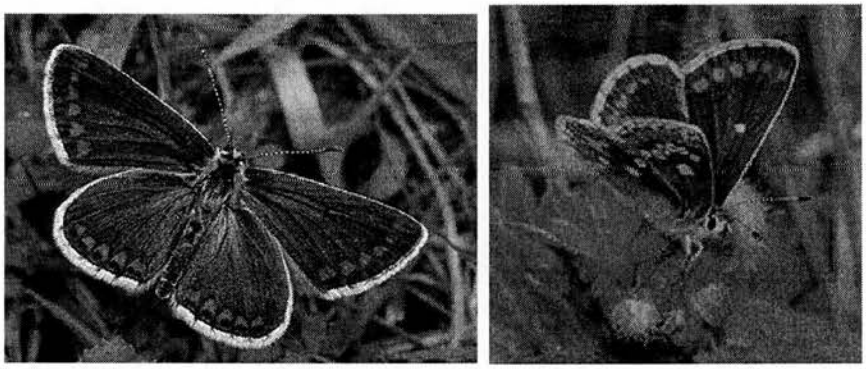
While the positive enforcement learning of **adaptive identifiers** allows DAISY to 'learn from experience', adaptive identifiers can lead to over-training. For example a DAISY system trained to recognise *Hyles tithymali* using only specimens of *Hyles tithymali tithymali* may ultimately reject *Hyles tithymali mauratanica* as "something else" when presented with it as an unknown (Fig. 2.21) (O'Neill, pers. comm.).

Fig. 2.21 – *Hyles tithymali tithymali* (left) and *H. t. mauratanica* (right). Images from Tony Pittaway's website 'Sphingidae of the Western Palaearctic, http://tpittaway.tripod.com/sphinx/list.html



DAISY discrimination of the Brown Argus (*Aricia agestis*) and the Northern Brown Argus (*Aricia artaxerxes*) have demonstrated the importance of **manifold reduction**. The *Aricia* species look almost identical except that the Northern Brown Argus has a white spot in the centre of each forewing (Fig. 2.22). This may get drowned in the essentially identical signal from the rest of the wing unless manifold reduction is used to extract the feature (O'Neill, pers. comm.).

Fig. 2.22 – Brown Argus (left) and Northern Brown Argus (right). Images taken by Alan Barnes and Anon., Butterfly Conservation website, www.butterfly-conservation.org



The most important **output metrics** of DAISY are first-past-the-post (FPTP) and coordination. In the simplest form of *n*-tuple nearest neighbour classification, single-NNC, an unknown is put in the same class as its nearest neighbour in feature space. If this nearest neighbour is of the correct class then the **first-past-the-post (FPTP)** identification is said to be correct. Identifications by this method can be led astray by atypical training images so certainty is low. The higher certainty output metric used to quantify accuracy is **coordination**. Coordination provides a very conservative rule base for making identification decisions, commonly giving certainty as high as 95%. Both single-NNC and coordination are have already been introduced in section 2.1.6.

MacLeod *et al.* (in prep.) found that a coordination level of eight (Coord8) gave good DAISY accuracy for foraminifa identification without sacrificing speed. The lower coordination level of three (Coord3, i.e. if the three nearest neighbours were of the correct class then an identification was considered successful) was necessary for this work as some training sets were as small as four images. This was likely to be acceptable as Coord3 is the default coordination level for DAISY, found empirically to be a useful metric in many different identification scenarios, including Hymenoptera identification. The choice of Coord3 was supported when it corresponded to 95% certainty in the identification of these Hymenoptera wing images.

Virtual HTML (VHTML)

This takes the identification data and attempts to access extra information on that taxa.

- If local web pages have been prepared in advance, VHTML will start a browser to display them.
- If no information is available on the local system, VHTML can:
    - Simply display a list of names (as a HTML file); or
    - Dispatch a search agent to seek out appropriate information on the Internet.
- VHTML is also capable of more complex data mining operations. For example, it can extract information from remote databases and transform it into virtual HTML.

Chapter 3 - DAISY identification of flower visiting insects

3.1 Introduction

3.1.1 *Acacia*-visiting insects and the problem of identification

A great range of insects visit *Acacia* flowers (Stone *et al,* 1996). Over 70 species of bees, 70 species of wasp, and 100 species of fly were caught for this project alone. These came from 29 different families (Appendix 1). Identifying this diverse assortment of insects has been a major challenge ever since the research group began work on acacia pollination (Stone, pers. comm.). Species-level identification has largely been avoided by discussing only higher taxonomic groups. For example, Stone *et al.* (1998) referred to flower visitor assemblages, e.g. megachilid bees, other solitary bees, honeybees or wasps. If identification could be automated then information of species could be accessed in real-time, informing the fieldwork decisions of researchers at Mpala and enabling insect behaviour to be understood at species level. This would transform *Acacia* pollination studies. The DAISY automated identification system has great potential here, as it has previously identified insects accurately (Weeks *et al.,* 1997; Weeks *et al.,* 1999b; Gauld *et al.,* 2000; Pajak, 2001; Watson *et al.,* 2004). However, the methodology used to extract the region of wing pattern for analysis, known as 'standard' or 'S', has been time consuming and has generally required wings to be removed from dead specimens. So that these two limitations do not prevent the application of DAISY in *Acacia* pollination studies two new methodologies have been investigated (introduced more fully in section 3.1.5). In the first, known as 'boxed' or 'B', the region of pattern is selected with less precision using a rectangular box-crop. In the second, known as 'attached' or 'A', the wing remains attached, making possible the use of live and museum specimens for DAISY training.

The range of visitors is known to differ between *Acacia* species that offer different amounts of nectar. In Stone *et al.* (1999), *A. senegal* produced abundant nectar, attracting honeybees, leaf-cutter bees, butterflies, spider wasps and sunbirds. *A. tortilis* and *A. zanzabarica* produced minute quantities of nectar and were mainly visited by leaf-cutter bees, small halictid solitary bees and pollen-feeding flies, especially hoverflies (Syrphidae) of the genus *Eristalinus* and blowflies (Calliphoridae) of the genus *Rhyncomya. A. nilotica* and *A. drepanolobium* produced no nectar and received almost all their visits from leaf-cutter bees. Megachilid leaf-cutter bees in the genera *Creightonella, Chalicodoma* and *Megachile* visited all the coflowering *Acacia* species, structuring their pollen-collecting activity to correspond with the sequence of pollen release in the *Acacia* species (Stone *et al.,* 1998; Stone *et al.,* 1999).

There was insufficient time to train DAISY on the full range of inflorescence visitors. It is most useful to identify the insect groups that are directly involved in pollination so this was investigated by reviewing the literature and observing flower visitation. The frequency with which different taxa landed on inflorescences was recorded, so that the most likely pollinators could be noted and their inclusion in DAISY training ensured. Insects that actively collect pollen or nectar and more likely to be pollinators of *Acacia* than insects that visit inflorescences to mate, eat the leaves and flowers or predate other visitors.

Psyllid bugs are commonly found in the inflorescences of *Acacia* and polyads are often collected in the wings of mature individuals (Bernhardt, 1989). However, as psyllid maturation requires the immature psyllid to eat the sexual organs of the flower they are best regarded as floral predators (New, 1984).

Butterflies are frequent visitors to nectar-producing acacias but are thought to be poor pollinators. Their long legs raise their bodies such that anthers rarely contact the wide surfaces of the wings and abdomen while the butterfly attempts to forage (Bernhardt, 1989). Hawkeswood (1985) has shown that fewer than half of the butterflies captured on *A. bidwillii* carried polyads.

The role of beetles is unclear, they may act as pollinators but generally seem to function as floral predators. Beetles removed from *Acacia* inflorescences can carry heavy loads of pollen (Hawkeswood, 1983). Many genera have modified mouthparts and digestive tracts for the consumption of pollen and nectar (Crowson, 1981), especially the families Buprestidae, Cleridae and Chrysomelidae which are frequent visitors to *Acacia* (Bernhardt, 1989). Bernhardt (1989) describes a convincing example of beetle pollination, in which clerid beetles covered the inflorescences of *A. mearnsii.* They foraged on anthers, did no damage to carpels and flew from tree to tree. Twenty-one of the 26 beetles caught carried polyads of *A. mearnsii.* However, the case against beetle pollination is strong. Some beetles strip away whole stamens and possibly consume carpels (New, 1984) so they often function as floral predators. Although *Phlogistus* species (Carabidae) were often numerous on *A. myrtifolia,* they could only be collected during the last two weeks of the flowering season when they foraged principally on inflorescences with dying, withered stamens (Bernhardt, 1989). The preference of beetles for old inflorescences was also observed in this study. Beetles also tend to stay on a single *Acacia* tree for a long time, so they are more likely to effect self-pollination than cross-pollination (Hawkeswood, 1983).

The true flies appear to play a larger role in pollination than the beetles. In Bernhardt (1989) they were found on every *Acacia* taxon studied and a far greater number of flies were collected on the inflorescences than were beetles. When observing and catching flies in this project the families that occurred most frequently were the blow-flies (Calliphoridae) and the hoverflies (Syrphidae). While the

calliphorids were common they carried too few pollen grains to make a significant contribution to pollination. Adult hoverflies carried larger pollen loads. They are known to eat pollen and are considered legitimate pollinators of Australian angiosperms. They have often been observed and captured on plants that offer pollen but lack floral nectar (Armstrong, 1979). Syrphids visit *Acacia* inflorescences when no other taxa are actively foraging, early in the morning on cool, cloudy days and even in light rains (Bernhardt, 1989).

Wasps are less frequent visitors than flies and are rarely observed to probe inflorescences (Berhardt, 1989). Less than a third of the wasps collected on inflorescences of *A. terminalis* carried polyads (Knox *et al.,* 1985). Most of the wasp genera observed in this project are known predators, so were probably attracted by pollinators rather than floral resources.

Ants in the genus *Chromatogaster* can be very numerous on Kenyan *Acacia* trees, especially *A. drepanolobium*. They live in hollow 'pseudogalls' (actually the swollen bases of thorns) formed by the tree, and deter herbivores of all sizes from browsing on the foliage. They are unlikely to contribute to pollination for three reasons. First, they tend not to move between trees (Willmer & Stone, 1997; Willmer *et al.,* 1999). Second, acacias have evolved mechanisms to keep ants away from flowers during the pollination period (Ghazoul, 2001; Raine *et al.,* 2002; Stone *et al.,* 2003). Finally, many ant species secrete antibiotic substances onto the integument, making any polyads transported inviable (Wagner, 2000).

The bees are the most important group of *Acacia* pollinators (Bernhardt, 1989), as they actively collect pollen with which to rear their larvae. This pollen is carried in regions of dense hair, known as scopa. The bee families thought to play the largest role in the pollination of East African *Acacia* are the Apidae, Halictidae and Megachilidae (Stone *et al*, 2003). Most bees sampled for pollen carried large numbers of *Acacia* polyads. Therefore, the bees were selected to be the focal group for DAISY training, along with a few wasp species that could be mistaken for bees by an non-expert.

A collection of authoritatively identified *Acacia*-visitors needs to be available for initial training of the DAISY system. Such a collection was not available so it was constructed from scratch. Three stages were necessary to achieve this:

1. Specimens of *Acacia*-visitors were collected by hand netting.
2. These specimens were "roughly" identified in the field.
3. The taxa to be used for DAISY identification were identified by an expert.

To obtain identities for the training specimens it was first necessary to identify them using conventional taxonomy. There are no published taxonomic works of direct relevance to the *Acacia*-visiting insects of Kenya. The National Museum of Kenya has a wide range of bee and wasp specimens but these are badly in need of curation, with incorrect and out-dated determinations. The Natural History Museum collections in London are well curated but Kenyan specimens are distributed throughout the entire collection, so it was necessary to have provisional identifications at genera-level to use them effectively. Dr Connal Eardley (Agricultural Research Council, Pretoria, South Africa) is working on a key to the East African bees but this is still in development. As it was not possible to identify specimens using published resources and reference collections aculeate specimens were sent to Eardley for identification. He identified bees to species or morphospecies and sphecid wasps to genus. This dependency on expert taxonomists for routine identifications had clear disadvantages: such experts are overstretched, specimens must be transported around the world and specimens are unavailable for reference for the duration of identification.

3.1.2 Non-automated insect identification

The groups most often caught in studies of *Acacia* pollination are the Hymenoptera, Diptera, Coleoptera and Lepidoptera (Stone *et al.*, 2003). Within these taxa, some closely related species are almost indistinguishable to the human eye, and identification requires microscopic examination of characters and subtle distinctions only appreciated readily by an experienced specialist.

The easiest characteristics to observe in the Hymenoptera (bees, wasps and ants) are size and colour. However, these " simple" characteristics can be unreliable. Size and colour vary both within and among species, and coat colour may fade during life while many species show some polymorphism, including melanic individuals (Prys-Jones and Corbet, 1991). Structural characters are less variable but may require magnification and specialist experience to interpret, making their use in the field more problematic. The main areas for examination are the head, wings and legs. The Hymenopteran head has many diagnostic characters, a frontal view of the head of a honeybee, *Apis mellifera* is shown in Fig. 3.1.

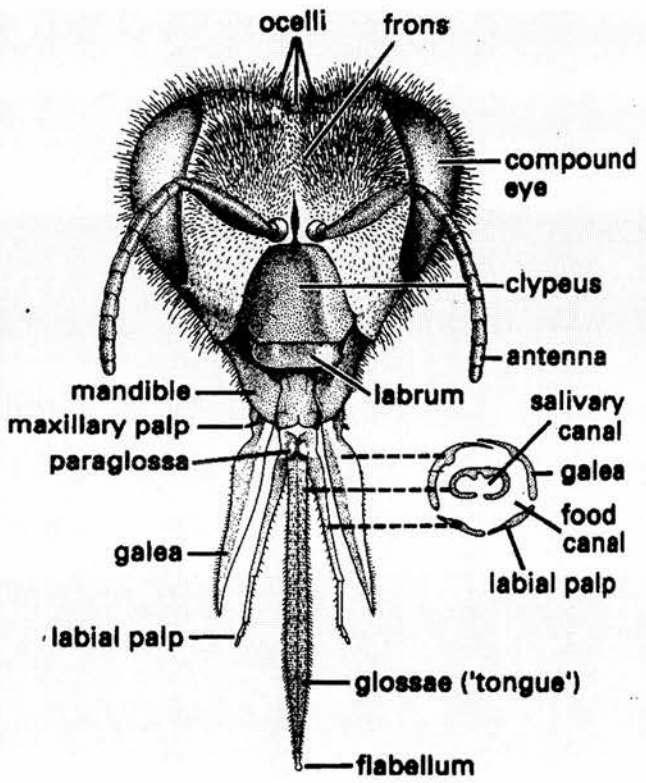


Fig. 3.1 – The head of a worker honeybee, *Apis mellifera*, shown in frontal view (Gullan & Cranston, 1994).

The antennae, the organs of smell, vary between taxa in their segment number. The mouthparts are used for chewing and lapping (Gullan & Cranston, 1994). The proboscis as a whole is especially useful for family level identification (Michener, 2000). The glossae ('tongue') may be short (e.g. Colletidae, Halictidae) or up to twice the length of the head. Long tongued bees (e.g. Apidae) are able to take nectar from flowers with longer corollae. The glossae may be forked (in the Colletidae) or pointed at its tip. The labrum may have a tuft of erect hair or a marginal fringe. It is broader than long in the Apidae but longer than broad in the Megachilidae. The mandibles ('jaws') may be simple or have sharp or rounded teeth, separated by acute notches or rounded emarginations (Scholtz & Holm, 1985; Eardley, pers. comm.).

The wings have many useful characters. The tegula is a flap that covers the base of a wing and it is unusually large in the genus *Pseudapis*. The spatial arrangement of wing veins provides several important characters. These are discussed and illustrated in section 3.1.3 and wing venation, used for automated identification in sections 3.2 to 3.4. The pterostigma, an opaque wing spot at the anterior of the wing, is not always present and varies in its relative size and shape.

Tibial spurs are long, pointed projections from the distal end of the hind leg femur. There may be none (e.g. Apidae), one or two. Terminally on the leg, there may or may not be a lobe, known as an arolium, between the lateral claws but this can be difficult to see (Gullan & Cranston, 1994).

Vital to pollen transportation is the location of the scopa (the hairs that hold pollen). Parasitic bees and many wasps have no scopa. Some of the eusocial apids (e.g. *Apis* and *Plebeina*) have scopa on their hind leg only, modified to form pollen baskets, or corbiculae. The outer hind leg of an *Apis mellifera* worker is shown in Fig. 3.2.

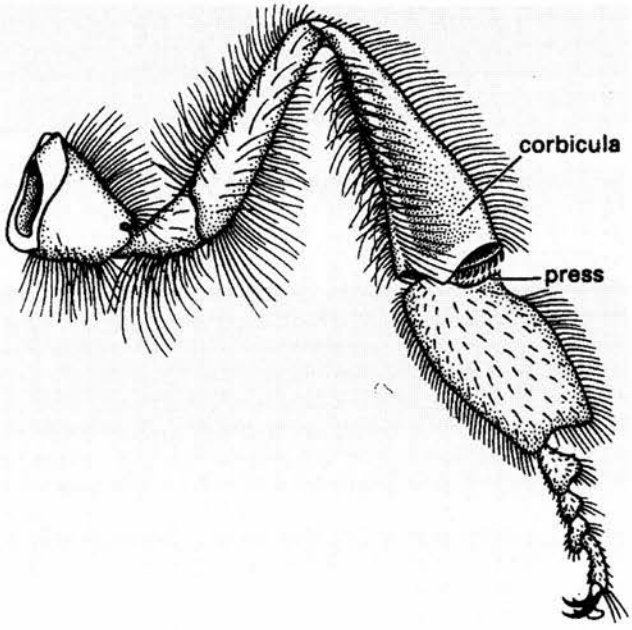


Fig. 3.2 – The outer surface of the hindleg of worker *Apis mellifera*, showing the corbicula, a depression fringed by stiff setae, and the press that pushes pollen into the basket (Gullan & Cranston, 1994).

Other bees have no concave corbiculae but masses of branched hairs, either mainly on the hind legs (e.g. *Amegilla*) or on the ventral surface of the abdomen (diagnostic of the family Megachilidae).

The genitalia may be the only way to separate very similar species of Hymenoptera (e.g. Koeniger *et al*, 1991). A specimen must be dissected to examine these genital structures so these characters are less accessible to the non-specialist than other features.

The Syrphidae are thought to be the most important flies for *Acacia* pollination (Stone *et al*, 2003). Syrphids, commonly called hoverflies, resemble bees (e.g. *Eristalis, Merodon* and *Volucella)* and wasps (e.g. *Syrphus* and *Chrysotoxum*) (Smith & Vockeroth, 1980) (Fig.3.3).

Fig. 3.3 – *Chrysotoxum festivum*, a syrphid that mimics a wasp.

Hoverflies can be recognised from other flies by their distinctive wing venation. They have two outer cross veins close to the wing margin, other flies have one or none, and a false vein running down the middle of the wing (a simple thickening of the membrane) (Fig. 3.4).

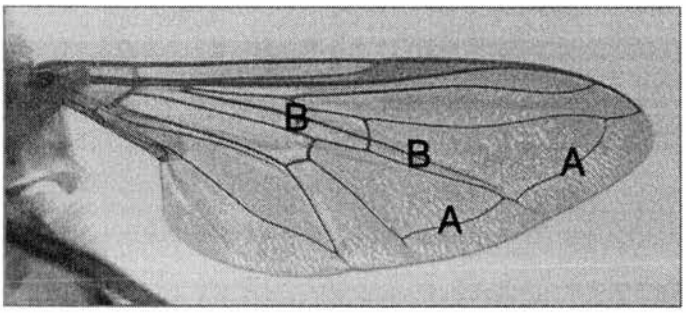


Fig. 3.4 – Wing of *Eupeodes corollae,* a syrphid, *showing* the two outer cross veins (A) and the false vein (B). Adapted from www.bioimages.org.uk

Wing venation is also useful when identifying below family level. For example, the genus *Eristalis* (and a few closely related Eristalini genera) has a distinctive venal loop. Species can also be identified by abdominal colour patterns, but again these can vary intraspecifically and dark forms occur. Other

important characters require microscopic examination, including the distribution of hairs over the body, head structure and antennal shape (Gilbert, 1993).

Pollen and nectar are used by many adult beetles, especially members of the families Cerambycidae, Lacanidae, Dascillidae, Melyridae, Oedemeridae, Mordellidae, Scrapriidae, Nitiduliidae and Meloidae (Evans, 1975). Size is often an important character, so identification will generally begin with calliper measurement of the length of the beetle. Elytra colour varies between species but newly emerged adults are often paler and colour may change in older specimens. Useful characters include antennal form (threadlike or more complex, with or without fine hairs towards tip); tarsal segment number (1-5), shape and fringing; pronotum shape; and mouthparts (especially mandible shape). Occasionally, the only reliable characters are related to the male genitalia or elytral sculpturing (Forsythe, 1987).

The colourful wing markings (on both the dorsal and ventral surfaces of fore wing and hind wing) of butterflies can be very informative and most butterflies can be identified easily from photographs. The patterns often vary intraspecifically in extent but generally not in position. The principal areas of the wing are defined by venation and include the costa (front edge), the inner margin (rear edge), the outer margin and the subdivision of the wing surface into basal, discal, postdiscal, submarginal, marginal and apical areas (Fig. 3.5).

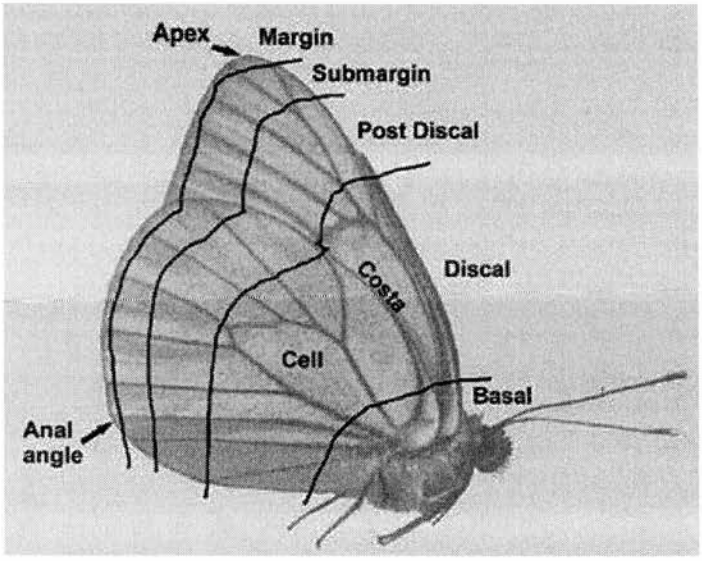


Fig. 3.5 – Subdivision of the butterfly wing surface. By Simon Coombs, www.butterfly-guide.co.uk

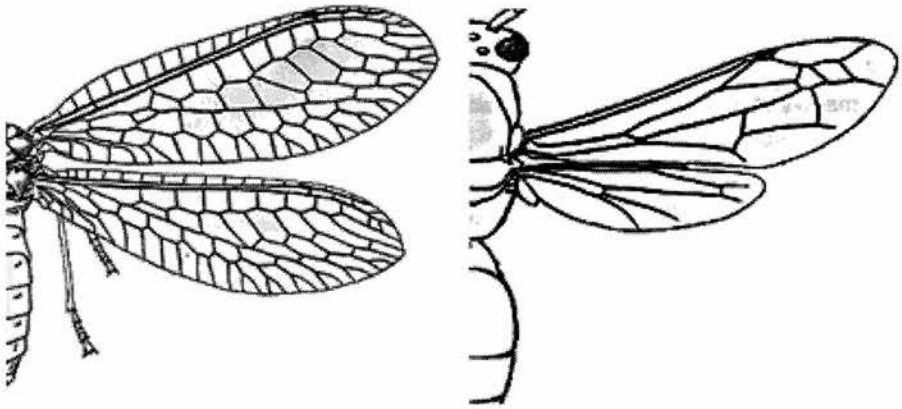
Lepidoptera taxonomy also relies on numerous other structures and, at the species level particularly, on features of the genitalia. Size is too variable within species (depending on larval diet) to form a reliable character (Higgins and Hargreaves, 1983).

3.1.3 Wing venation for identification

Insect wings are flap-like folds of the integument, composed of two sheets of cuticle supported by tubular thickenings called veins. Cells are areas of wing delimited by veins. The pattern of veins in a wing is known as wing "venation". Hymenoptera wing venation provides many important characters for identification and phylogenetic discussion (e.g. Sharkey & Roy, 2002; Schulmeister, 2003; Rasnitsyn *et al.,* 2004).

Most insects use two sets of wings in flight, fore wings and hind wings. They may be coupled together, improving the aerodynamic efficiency of flight. The commonest coupling mechanism (and that seen in the Hymenoptera) is a row of small hooks along the anterior margin of the hind wing that engage a fold along the posterior margin of the fore wing (Michener, 2000). The wing venation of more primitive insects is netlike but there has been a general tendency in wing evolution towards fewer, more highly sclerotized veins. This reduced venation is seen in the Hymenoptera (Fig. 3.6). Groups also differ in wing pigment patterns and colours, hairs and scales.

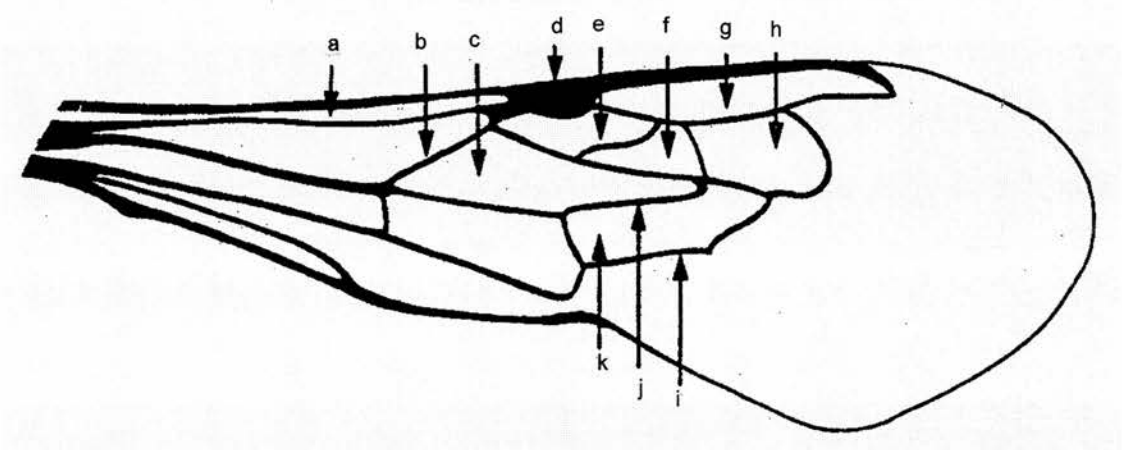
Fig. 3.6 – Extensive wing venation (left) and reduced wing venation (right), as illustrated in the CSIRO online insect key. www.ento.csiro.au/.../couplet_20.htm



The forewings tend to provide wing characters for insect identification and these will be discussed from now on. The major veins are longitudinal, running from the wing base towards the tip. These major veins open into the body, circulating blood and containing sensory nerve fibres. Additional supporting cross-veins are transverse struts (Michener, 2000).

The arrangement of veins in a wing is very specific in certain genera and families. The most important cells and veins for bee taxonomists are known by special names. Several naming schemes have been proposed, one of them is shown in see Fig 3.7.

Fig. 3.7 – Generalised bee fore wing with the most informative cells and veins for taxonomy labelled: a costal cell; b basal vein; c first discoidal cell; d pterostigma; e first submarginal cell; f second submarginal cell; g marginal cell; h third submarginal cell; i second recurrent vein; j first recurrent vein; k second discoidal cell. Adapted from Scholtz & Holm (1985).



In bees the submarginal cells (e, f & h in Fig. 3.7) are especially informative. There may be either two or three and the relative sizes of these cells are used in bee keys. The pterostigma (d in Fig. 3.7) is an opaque spot on the anterior side of the wing. The size and shape of this pigmented spot is used in identification. The basal vein (b in Fig 3.7) is also described in keys, as either straight or bent.

Insect wings are essentially two-dimensional so the venation pattern is well suited to computer vision. The area of densest wing venation is half way between the base and the apex. The basal region can easily get torn in wing removal and the distal edge often frays during the life of the insect. In addition, the wing can become folded at either end during mounting. For these reasons it is preferable to base venation analyses on the mid-region only.

3.1.4 Automated identification

Many insect wings, with their transparent, two-dimensional structure and clear venation pattern, are ideal for image analysis. The most exciting image analysis systems that are relevant to *Acacia* visiting insects are the Automatic Bee Identification System (ABIS) of Bonn University, SPIDA (SPecies IDentified Automatically)-web of the American Museum of Natural History (AMNH) and the Digital Automated Identification SYstem (DAISY) of The Natural History Museum (NHM), London.

ABIS identifies bees on the basis of wing venation landmarks (see section 2.1.4 for a description of the morphometric approach), the points at which veins meet. In earlier versions of ABIS, the user was required to mark vein junctions prior to analysis and 60+ individuals were required to represent each species. ABIS could then identify closely related species with 98% accuracy (Roth *et al,*. 1999; Arbuckle *et al*, 2001). The Bonn group has since reduced the demands for user experience. Three key wing cells (labelled 1 on Fig. 3.8) are located automatically and used to identify the bee to genus, and then genus templates are employed to locate the other informative landmarks (Hajdaoud *et al*, 2005) (Fig. 3.8). Once landmark data has been measured classification is statistical, using a variant of discriminant analysis. Whilst ABIS performs well for bees, this system cannot be used for other insects.

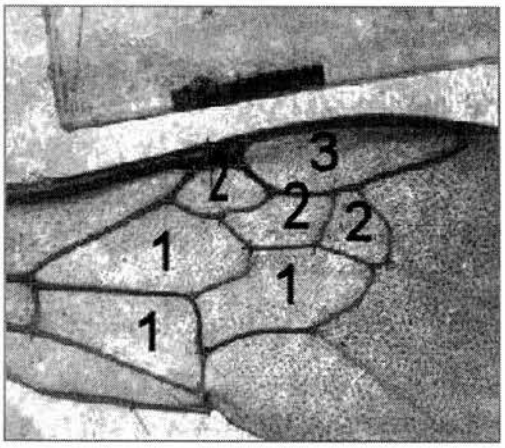


Fig. 3.8 – When ABIS analyses a bee wing it automatically locates the no. 1 cells. The shape of these cells provides a template to locate the submarginal cells (no.2). The shapes of the nos. 1 and 2 cells produce a template to locate the marginal cell (no.3) and all these cells then provide landmark characters to identify the specimen.

SPIDA-web does not use landmark features, instead simplifying the whole image by Daubechies / Gabor wavelet transformation and removal of the highest frequency resolution (see section 2.1.4). This compresses an image and eliminates some of the non-informative 'noise'. These wavelet transformed images are then classified in a hierarchical manner, first to genus and then to species. This is done using cascading artificial neural networks. Each taxa has its own neural network, each network gives a 'yes' or 'no' output and they all have to be run for an identification to be made. They are trained by back propogation (see section 2.1.7), so once trained the system is static and to add new images the entire system must be retrained. This is a major weakness of SPIDA-web. The SPIDA-web project began by identifying spiders from images of their external genitalia but the software has more recently

been applied to bee wing venation, separating 12 bee species (with 20 individuals per species). The accuracy of positive determinations was above 90% but positive identifications were rare, i.e. generally all the taxa networks outputted 'no'. They concluded that bee wings were insufficiently diagnostic to be the only pattern source for identification (Russell *et al.,* 2005).

DAISY is a generalised pattern-matching system, already described in sections 2.2 and 2.3. Not only is it applicable to bees, wasps, flies, beetle and butterflies, but it could also be adapted to a wide range of non-taxonomic, pattern matching tasks, such as medical physics (e.g. routine screening of cytological preparations), security systems (e.g. identification of individuals using retinal, iris and/or fingerprint analysis) or the location and subsequent tracking of tropical cyclones (O'Neill, 2005). When DAISY was tested on the 23 UK bumblebee species (for comparison with ABIS) it achieved 96% accuracy with a training set of 30 individuals per species (O'Neill, pers. comm.). It has already been used to identify live moths (Watson *et al.,* 2004), biting midges (Weeks *et* al., 1999a) and ichneumonid wasps (Weeks *et al.,* 1997 and 1999b). Whilst this work has relied on images, DAISY is capable of working with any pattern data, including DNA barcodes (such as those used in Barrett & Hebert, 2005). Further explanation of how DAISY works is given in section 2.3.

Other systems have also been used to extract pattern from and identify taxa that are relevant to *Acacia* pollination. The DrawWing software of Tofiliski (2004) simplified an image of vespid wasp wing venation into a venation diagram (see section 2.1.4). It would be just as applicable to bee or fly venation. Yu *et al.* (1992) identified ichneumonid wasps using vein junctions as landmark features and landmark coordinates as a numerical description of a wing. Honeybees have often been the subject of feature recognition (e.g. Dupraw, 1965; Daly & Balling, 1978; Rinderer *et al.*, 1993). Daly *et al.* (1982) used image analysis to measure 25 morphometric characters of honeybees and then by discriminant analysis determined whether they were European or Africanised.

Insect wings have been the subjects of many published automated identification trials (e.g. Gauld *et al.,* 2000; Hajdaoud *et al.,* 2005). However, it is uncertain how often wing venation alone can characterise a species. In non-automated identification of bees and wasps many other parameters are considered and the final decision is based on all of these. Head and body structures could potentially by used in image analysis but their three-dimensionality may lead to pattern distortions (Weeks & Gaston, 1997). However, DAISY has successfully identified caterpillars from the head capsules they shed when they moult (O'Neill, pers.comm.).

3.1.5 Work proposed

DAISY has successfully identified wasp and bee wings before (e.g. Weeks *et al.*, 1997; Weeks *et al.*, 1999a). However, in this published Hymenoptera work the identifications demanded a high level of skill and time input. Each wing was removed from the insect, mounted on a microscope slide and wing cells selected with a polygonal overlay (polyROI; see section 2.3.2.1 for polygonal overlays). This established method was used in this study as a baseline for new methods. It is referred to as "Standard" or "S". All closed cells (i.e. those surrounded by veins and not tapering to the wing base) are included in the polyROI for analysis: the pterostigma, marginal cell, submarginal cells, discoidal cells (labelled in Fig 3.7) and the cell posterior to the first discoidal cell. The poyROI region is shown in yellow in Fig. 3.9.

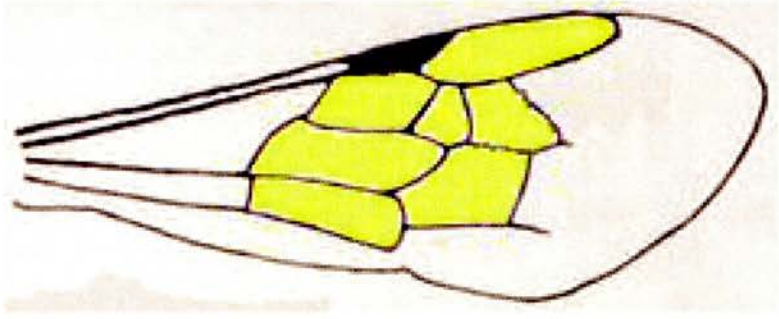


Fig. 3.9 – Yellow shading highlights the region selected by polyROI for the baseline method, S.

Two new methods have been developed, more applicable to field identification:

1. "Boxed", abbreviated to "B", in which the wing cells are box-cropped;
2. "Attached", abbreviated to "A", in which the wing is not removed from the body for imaging.

While introduced in this section, these methods are investigated in section 3.2. Chapter three also includes analysis of two methodological parameters, Normalised Polar Thumbnail size (NPT size, section 3.3) and training set size (TS size, section 3.4), as changing these parameters may increase identification accuracy.

Adding a polygonal overlay to a wing takes about two minutes, which is feasible for small-scale, proof-of-concept trials, but becomes impracticable when large number of specimens are involved. However, if the wing is imaged with the basal half of the anterior margin horizontal, then a standard region can be selected quickly by box-cropping the focal wing cells. All closed cells were included, as in S (pterostigma, marginal cell, submarginal cells, discoidal cells and the cell posterior to the first discoidal cell) so the sides of the box were delineated by the following wing features:

A	anterior wing margin
B	proximal, anterior corner of cell posterior to first discoidal cell
C	distal apex of marginal cell
D	distal, posterior corner of cell posterior to first discoidal cell

Fig. 3.10 – Box-crop of *Amegilla fallax*, a bee in the family Apidae, with cell box-crop delineated by A, B, C and D.

A

B  C

D

A box-crop approach was taken in the unpublished *Bombus* work of Pajak (2001), but it was not compared with any other method nor even discussed as an approach. Furthermore, the pool size was just four bumblebee species and it used a different output metric to other DAISY studies. This was the mean number of partitions (the rank of match where one is the closest match, necessary to identify 80% of the unknown specimens) rather than the proportion of correct identifications.

Both S and B methods use images of detached wings. The damage this inflicts to specimens is restrictive for two reasons. First, all insects for training and testing have to be killed. This makes DAISY unsuitable for identifying threatened species or those where the population contains few individuals. Second, museum collections cannot be used for training, as damage allowed to specimens is strictly limited. Catching and identifying the same species again is a huge investment of time that could be better spent on other work.

The second new method proposed (A) involves imaging wings still attached to insects. Watson *et al.* (2004) applied DAISY to live moths in varied resting postures, identifying 35 species with 83% accuracy. The transparent nature of bee wings makes attached imaging problematic, as body parts behind the wing can be seen and can obscure the wing cell pattern. The ABIS system (Hajdaoud *et al.*, 2005; see section 3.1.4) circumvented the problem by working with live bees; the bee limbs were flexible so a fore wing could be stretched out from the body using a wing clamp.

Method A is the first realistic solution for the use of museum specimens, in which pinned specimens are dried so body parts cannot be manipulated without damage. To move an obstruction (such as a leg) out of view and see all the wing cells it is often necessary to orientate the specimen so that the wing is

no longer horizontal; this is how wing cells would be observed for non-automated identification. However, such orientations introduce perspective errors into the venation pattern viewed, which could create problems of pattern distortion and focus. The selection of a polygonal region of interest (polyROI), as in S, should allow minor pattern distortion to be counteracted (for a screencapture from DFE showing a polygonal region of interest (polyROI) see section 2.3.2.1, Fig. 3.10).

3.2 Comparison of three methods for wing identification

3.2.1 Methods

3.2.1.1 Study organisms and site of capture

The study organisms were four families of bees (superfamily Apoidea) and three families of predatory wasps (superfamilies Vespoidea and Sphecoidea), classified as follows:

BEES

Superfamily *Apoidea*

- Family Apidae
  - *Amegilla*
  - *Apis*
  - *Ceratina*
  - Macrogalea
  - *Tetralonia*
  - *Tetraloneilla*
  - *Xylocopa*
- Family Colletidae
  - *Colletes*
- Family Halictidae
  - *Lasioglossum*
  - *Lipotriches*
  - *Patellapis*
  - Pseudapis
- Family Megachilidae
  - *Heriades*
  - *Megachile*

WASPS

Superfamily Vespoidea

- Family Eumenidae
  - *Delta*
- Family Tiphiidae
  - Meria

Superfamily Sphecoidea

- Family Sphecidae
  - *Ammophila*
  - *Bembix*
  - *Cerceris*
  - *Philanthus*
  - *Sphex*

The study organisms were caught by hand netting at flowering acacias at Mpala Research Centre, Kenya. Mpala is situated on the Laikipia Plateau, 1800m above sea level. It is just north of the equator, with grid reference 0°17'N, 37°52'E. The research centre occupies 1,200 ha but researchers have access to the adjacent Mpala Ranch comprising 17,000 ha (Young *et al.*, 1998). There are two major soil types on Mpala, known as 'black cotton' and 'red soil'. The black cotton is poorly draining (impassable after rains) and dominated by *Acacia drepanolobium.* However, most of Mpala is red soil, supporting a more diverse tree flora, which has been characterised as "open *Acacia brevispica* thicket" (Taiti, 1992 cited in Young *et al.*, 2003). The woody vegetation is dominated by *A. brevispica, A.*

etbaica and *A. mellifera*, but also includes *A. gerrardii*, *A. nilotica, Croton dichogamous*, *Grewia* spp. and *Rhus vulgaris* (Young *et al.*, 1995).

3.2.1.2 Capture, killing and pinning

Killing jars were made by adding a 5-10 mm layer of plaster of Paris to the bottom of 20 glass tubes and allowing it to dry fully. Before a catching session the killing tubes were cleaned thoroughly with ethanol, the plaster of Paris saturated with ethyl acetate and an air-tight lid used to seal the tube. Specimens were caught by hand-netting around flowering branches of *Acacia brevispica, A. etbaica, A. gerrardii, A. mellifera, A. nilotica* and *A. seyal.* They were then transferred to individual killing tubes, where they died quickly from the ethyl acetate vapours. Specimens were pinned on the evening of capture, before their wings and legs lost their flexibility. Each specimen was held between thumb and forefinger and a size two or three entomological pin pushed through its thorax, entering from the dorsal side. The insect was carefully pushed along the pin so that 1 cm remained between the dorsal surface of the insect and the pin head. This provided enough finger (or forcep) space to hold a specimen, whilst allowing as much room as possible below the specimen for data labels. The fore wings were decoupled from the hind wings (wing coupling was described in 3.1.2), the fore wings extended away from the body (to minimise obstruction to the venation pattern from other body parts), supported in this posture by bracing pins and left overnight to set in this posture. This is a standard posture for the pinning of Hymenoptera, seen in many museum specimens. The next day, two printed data labels were added to each specimen. The first (highest on the pin) was a place label, giving the locality details of Mpala Research Centre. The second gave ecological information specific to that specimen including the date and time of capture and the *Acacia* species they were visiting. Once specimens had been identified by Dr Connal Eardley (see section 3.1.1) a third data label was added, giving the identification of the specimen. The labels were spaced evenly on the pin using a mounting block, so that all three labels could be easily read without moving them on the pin.

3.2.1.3 Imaging

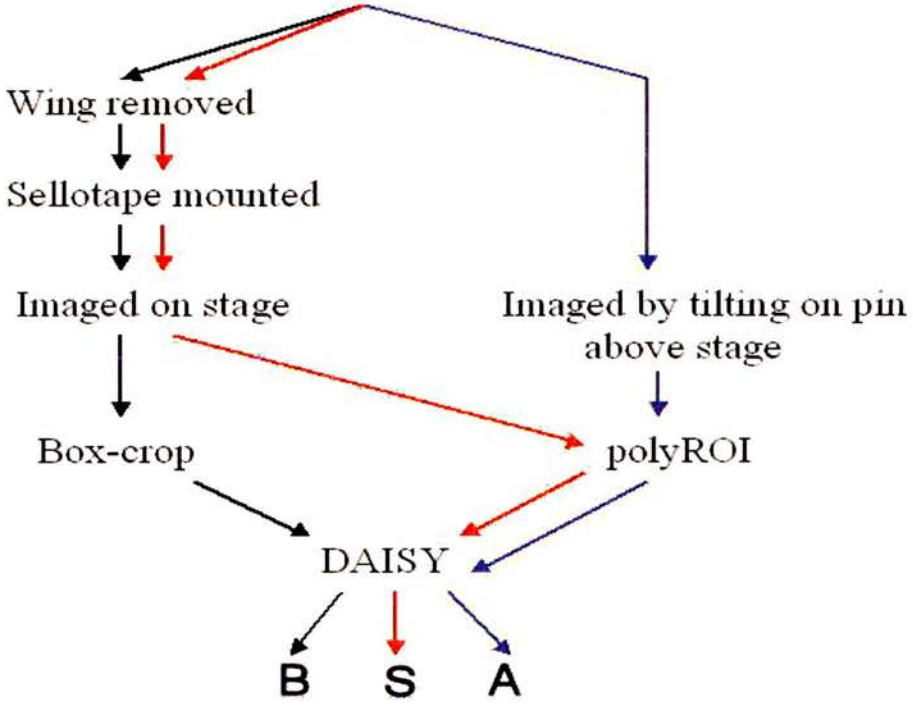
The three methods required different imaging methods. All images were taken with a Nikon Coolpix 4500 digital camera attached to a dissecting microscope, with no digital zooming, a background of thin white paper and illumination from beneath. Method A required the wing to be attached to the insect so this work was done first; the same specimens then had a wing removed for methods B and S.

There were two points at which method decisions had to be made (Fig. 3.11).

1) The forewing could either remain attached to the pinned insect, which was held above the microscope stage and tilted to bring venation into view (A), or be removed with forceps, sandwiched between two pieces of Sellotape® and imaged flat on the microscope stage (S and B).

2) The wing cells were then selected from the image either by rotating the wing so that the anterior margin was horizontal and box-cropping the area (B) *or* by adding a Polygonal Region of Interest (polyROI) with the DAISY front-end tool (DFE) (S and A).

Fig. 3.11 – Schematic diagram for the three image acquisition methods: Boxed (black), Standard (red) and Attached (blue).



Then:

- Images formats were changed from jpeg to tiff.
- The images were pasted into three folders in the LINUX file system (family, genus and species), and access settings modified if necessary.
- The images were renamed with their family, genus or species name (depending on their folder).
- Taxa with fewer than four images were deleted as these provided too few specimens for DAISY training.
- Image sets were built to training pools (generating NPTs) using the DAISY BuildTool.

- Each image was in turn identified with reference to all other images and mean values calculated. This process is known as jack-knifing and was done automatically using the DAISY JackTool.

3.2.3 Results

Tables showing the identities and numbers of all specimens caught (not just bees and wasps) are given as Appendix 1.

In the following tables:

- *FPTP (%)* is the percentage of images where the First Past the Post (i.e. the closest match) was correct. This corresponded to >90% mean certainty that the identification was correct.
- *Coord3 (%)* is the percentage of cases in which the closest three matches were correct (explained further in section 2.3). This more rigorous metric corresponded to >95% mean certainty that the identification was correct.
- Results are presented at family, genus and species levels.

Family level identification was very accurate, with the mean FPTP match over 95% correct in all three methods (Table 3.1). If FPTP was the identification metric then B and A performed almost as well as S (less than 2% reduction in accuracy). If 95% certainty was required (Coord3) then A and B led respectively to 5% and 8% lower accuracy overall than S.

Table 3.1 – Wing identification at family level using Standard, Boxed and Attached methods.

Family	genera #	specimen #			FPTP (%)			Coord3 (%)		
		S	B	A	S	B	A	S	B	A
Apidae	6	129	129	115	98	98	97	96	89	85
Colletidae	1	14	14	13	100	100	77	100	93	69
Eumenidae	1	19	19	25	84	95	96	84	79	92
Halictidae	4	57	57	41	95	93	95	91	74	83
Megachilidae	1	56	55	41	100	95	98	100	87	98
Sphecidae	5	67	67	59	100	100	97	90	87	95
Tiphiidae	1	16	16	15	100	100	100	100	100	100
All families		**358**	**357**	**309**	**97.5**	**96.9**	**95.8**	**94.4**	**86**	**89**

If the new methods are compared statistically to the standard method, using Wilcoxons' signed ranks test, there is generally no statistically significant difference in accuracy. The only significant accuracy difference is between the standard and boxed methods at 95% certainty (P = 0.027).

Apidae (6), Halictidae (4) and Sphecidae (5) were represented by more than one genera, and were therefore more variable than the single-genera families. This does not seem to have had a direct impact on their identification accuracy. Apidae and Sphecidae were identified quite well whilst Halictidae had relatively poor accuracy. This may be because these more variable families had more training images than the less variable families, such that each of the genera comprising that family was as well represented as the single genus in the other families.

The most accurately identified family (in all three methods) was Tiphiidae, achieving 100% accuracy with 95% certainty. Although there were relatively few training specimens for this family they all came from a single species. The flower wasps (tiphiids) have quite distinctive 'compact' venation (Fig. 3.12).

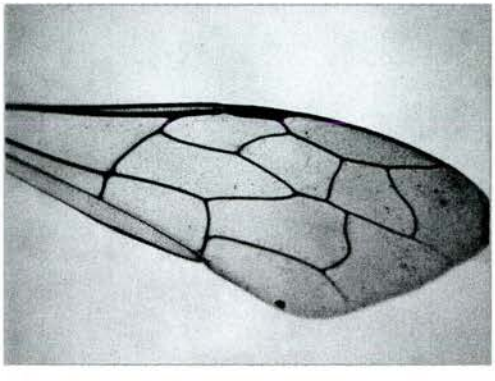


Fig. 3.12 – The wing venation of *Meria* sp.1, the single species to represent the family Tiphiidae.

The most poorly identified families (in all three methods) were Eumenidae and Halictidae. Eumenidae was represented by a single genus, *Delta.* The cell shape of *Delta,* with the first submarginal cell extending most of the way to the wing base, looks as though it should be distinctive (Fig. 3.13). However, *Delta* included three species which shared cell shape but differed in wing pigmentation, so a single pattern had very few training individuals.

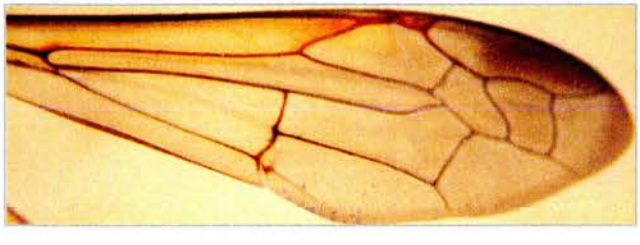


Fig. 3.13 – *Delta lepeleterii*, one of three *Delta* species that differed substantially in wing pigmentation.

The halictid bees, the other family to be identified with relatively low accuracy, are one of the most problematic groups to identify manually, often causing identification problems for expert taxonomists.

Genus-level identification was almost as accurate as family level, all three methods achieving overall FPTP accuracy above 94% (Table 3.2). For FPTP, the three methods differed in accuracy by less than 2% overall. When 95% certainty was required (Coord3) S and A out-performed B by 5% overall.

Table 3.2 – Wing identification at genus level using Standard, Boxed and Attached. Those genera that were identified with less than 90% accuracy FPTP are highlighted.

Genus	Family	#			FPTP (%)			Coord3 (%)		
		S	B	A	S	B	A	S	B	A
Amegilla	Apidae	38	38	30	100	100	100	100	100	97
Apis	Apidae	22	22	18	100	100	100	100	100	100
Ceratina	Apidae	8	8	9	75	75	100	63	38	100
Macrogalea	Apidae	21	21	20	95	95	95	95	95	95
Tetraloniella	Apidae	7	7	6	100	100	83	100	43	0
Xylocopa	Apidae	21	21	27	100	95	96	95	95	81
Colletes	Colletidae	14	14	13	100	100	92	100	86	72
Delta	Eumenidae	15	15	30	87	93	97	87	87	97
Lasioglossum	Halictidae	12	12	12	58	42	83	8	8	50
Lipotriches	Halictidae	15	15	15	87	93	87	67	60	60
Patellapis	Halictidae	12	12	13	83	83	100	33	42	100
Pseudapis	Halictidae	10	10	13	100	90	92	70	50	46
Megachile	Megachilidae	43	43	54	98	100	98	98	91	98
Ammophila	Sphecidae	10	10	6	100	80	67	70	60	33
Bembix	Sphecidae	18	18	20	100	100	100	94	89	100
Cerceris	Sphecidae	9	9	7	100	100	86	100	100	86
Philanthus	Sphecidae	19	19	17	100	100	94	89	89	94
Sphex	Sphecidae	7	7	22	100	100	100	100	86	100
Meria	Tiphiidae	16	16	15	100	100	100	100	100	100
All genera		**317**	**317**	**347**	**95.3**	**94.3**	**95.7**	**87.1**	**82**	**87.6**

When the accuracy of the three methods are compared using Wilcoxon's signed ranks test the result is the same as that for family-level. The attached wing method gives accuracies that are not significantly lower than those of the standard method. The boxed wing method is significantly less accurate than S only when 95% certainty is required ($P = 0.016$).

A few genera were identified less accurately than the others. If values are taken from the method S FPTP results (highlighted in green on Table 3.2) these are: *Ceratina* (75%), a small apid; *Delta* (87%), trained with three species that differed in pigmentation; and three halictids, *Lasioglossum* (58%), Lipotriches (87%) and *Patellapis* (83%), a family of small bees that has already been described as taxonomically challenging. The other methods were sometimes more accurate for these genera but had problems with others, e.g. method A achieved 100% accuracy with 95% certainty for *Ceratina* and

Patellapis and was 25% more accurate FPTP for *Lasioglossum* yet it identified the Sphecid wasps *Ammophila* (67%) and *Cerceris* (86%) less accurately (method S was 100% accurate in both cases).

Where there were several genera of a single family it might be expected that they would be confused. However, only around half the mis-identified images were identified as genera in the same family (S = 46%, B = 63% and A = 46%). This suggests that genera in a family have little in common so little was gained by family-level identification. *Lasioglossum,* the least accurately identified genus for methods S and B, must have been highly variable as it was misidentified as many different genera (*Lipotriches, Megachile, Patellapis* and *Macrogalea*) in three different families. *Pseudapis* mis-identifications were always *Lipotriches,* another genera in the Halictidae.

There were only 22 species with four or more images to represent them. Many of these Connal Eardley (expert bee taxonomist) was able to separate but not name, e.g. *Megachile* sp.4. Again the three methods gave very similar accuracies, identifying on average over 90% correct FPTP (over 90% certainty) and over 72% correct Coord3 (over 95% certainty) (Table 3.3). However, some species responded very poorly, e.g. accuracy dropped to 40% FPTP and 0% Coord3 in *Delta hottentottum* using method S.

S performed slighter more accurately on average than the new methods (6.5% is the biggest difference). B achieved greater accuracy than A FPTP but this was reversed for Coord3. However, looking at individual species there are often huge differences in accuracy from method to method. These are especially pronounced in the Coord3 accuracies where the species most affected by method choice were those with small training sets. For example, *Amegilla calens* (trained with five specimens) and *Megachile gratiosa* (five) varied by 100%, *Lasioglossum* sp.1 (six) varied by 78% and *Tetraloniella* sp.2 (four) varied by 75%. This largely suggest that methods B and A are more severely affected by low training set size than method S (see section 3.4 for further support for this idea), although in the case of *Tetraloniella* sp.2 the greatest accuracy came from method A (100% compared to 75% from method S).

The species identified with 100% Coord3 accuracy for all three methods were *Apis mellifera, Meria* sp. 1 and *Philanthus triangulum* (Fig. 3.14) . *Apis mellifera* is known for having distinctive venation, especially the long and thin 3rd submarginal cell. It is harder to see what may make the other two distinctive. All three had relatively large (> 15) training sets so they were well represented in morphological space.

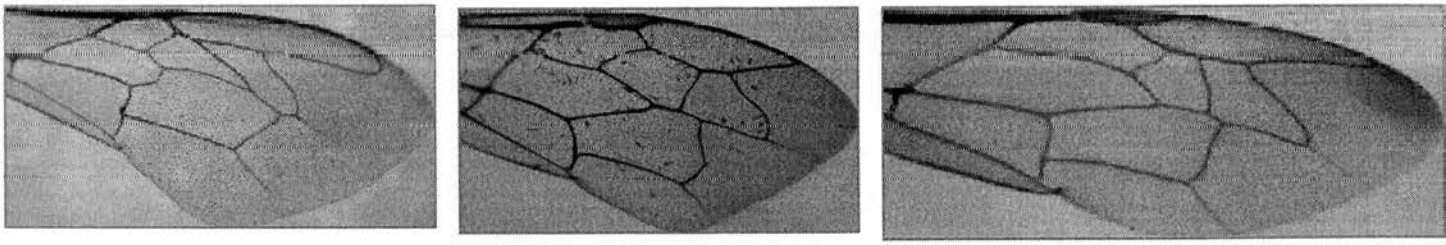
The three species that were identified with less than 100% accuracy FPTP in all three methods, were *Delta hottentottum, Delta lepeleterii* and *Heriades* sp.1, shown in Fig. 3.15. The *Delta* species resemble each other quite closely so were sometimes confused. Heriades is one of the smallest bees so it was more difficult to get good quality images of its wing venation, especially in method A (Fig. 3.16).

Table 3.3 – Wing identification at species level using Standard, Boxed and Attached methods. FPTP accuracies < 90% and Coord3 accuracies < 70% highlighted.

pecies	#			FPTP (%)			Coord3 (%)		
	S	B	A	S	B	A	S	B	A
megilla calens	5	5	5	100	80	80	100	0	20
megilla fallax	24	24	20	79	83	100	71	63	90
pis mellifera	22	22	18	100	100	100	100	100	100
embix forcipata	13	13	16	100	100	100	100	92	100
eratina moerenhouti	6	6	9	83	83	100	83	33	100
olletes sp.1	12	12	11	100	100	82	100	75	82
elta hottentottum	5	5	14	40	60	57	0	40	14
elta lepeleterii	7	7	11	71	71	64	57	43	9
eriades sp.1	4	4	6	75	75	67	0	25	17
asioglossum sp.1	6	6	9	100	100	78	100	100	22
lacrogalea candida	21	21	20	95	95	95	86	95	95
legachile gratiosa	5	5	4	100	80	50	100	80	0
legachile sp.1	7	7	5	86	86	100	86	86	100
legachile sp.4	8	8	7	88	75	100	88	13	57
legachile sp.6	9	9	8	100	100	88	89	89	75
leria sp.1	16	16	15	100	100	100	100	100	100
atellapis sp.1	12	12	13	92	83	100	67	58	100
hilanthus triangulum	17	17	16	100	100	100	100	100	100
seudapis sp.1	10	10	13	100	100	92	80	70	69
etralonia nigropilosa	4	4	4	100	100	100	100	75	75
etraloniella sp.2	4	4	4	100	100	100	75	25	100
ylocopa somalica	14	14	18	100	100	94	100	100	94
ll species	**231**	**231**	**246**	**93.1**	**92.2**	**90.7**	**78.8**	**72.3**	**74.8**

When the accuracies of A and B are compared to those of S using Wilcoxon's signed ranks test again it is only B at 95% certainty that is significantly different (P = 0.038).

Fig. 3.14 – The three species that were identified perfectly with 95% certainty in all three methods. L – R: *Apis mellifera, Meria* sp.1 and *Philanthus triangulum.*



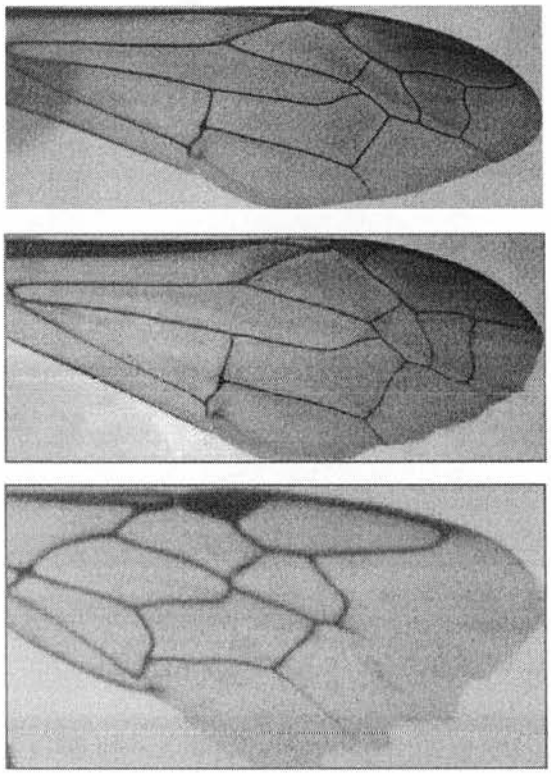


Fig. 3.15 – The three species that no method could identify with 100% accuracy. Top to bottom: *Delta hottentottum, Delta lepeleterii* and *Heriades* sp.1.

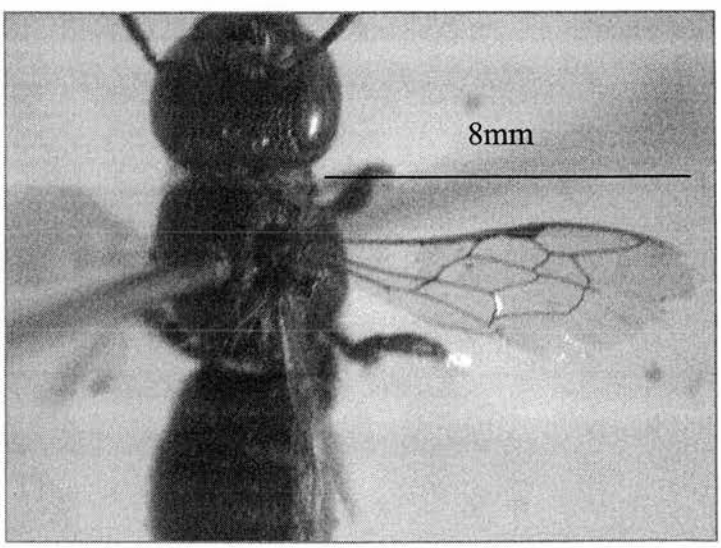
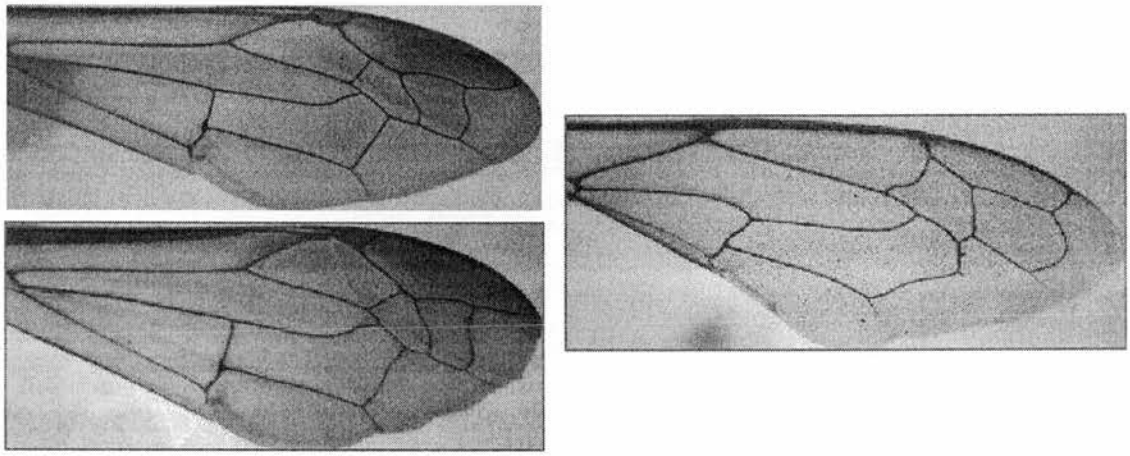


Fig. 3.16 – The small wing size of *Heriades* sp. 1 made it difficult to get good quality images of wing venation, especially when the wing remained attached to the insect.

Only three genera were represented more than once: *Amegilla* (2), *Delta* (2) and *Megachile* (4). In *Amegilla* and *Megachile* all mis-identifications were of species in the same genus. In *Delta* a few specimens (40% of method S mis-identifications and 25% of method S mis-identifications) were identified as *Bembix forcipata,* another wasp. They do not look very similar so this is hard to understand (Fig. 3.17). The general trend towards same-genera mis-identifications suggests that same-genera species are much closer in morphospace than different-genera species so species-level identification is more difficult for DAISY.

Fig. 3.17 – *Delta* was sometimes mis-identified as *Bembix forcipata* rather than the other *Delta* species. *Delta hottentotta* (bottom left) and *Delta lepeleterii* (top left) resemble each other much more closely than they resemble *Bembix forcipata* (right).



The number of classes was greatest at species level and declined at higher taxonomic levels (Table 3.4). With fewer classes to discriminate it makes sense for the family-level identification to be the most accurate.

Table 3.4 – number of possible taxa at different taxonomic levels.

Taxonomic level	Number of possible taxa
Family	7
Genus	19
Species	22

In every case family-level did give the most accurate identification, followed by genus-level then species-level (Table 3.5). The difference in identification accuracy between family and genus levels was smaller than that between genus and species. This is especially pronounced in method A where family-level and genus-level identification only differed slightly in accuracy (0.1% FTPT and 1.4% Coord3) but genus-level and species-level differed more substantially (5% FPTP and 13% Coord3).

Table 3.5 – Mean identification accuracy at different taxonomic levels.

Taxonomic level	Mean accuracy (%)					
	90% certainty			95% certainty		
	S	B	A	S	B	A
Family	97.5	96.9	95.8	94.4	86.0	89.0
Genus	95.3	94.3	95.7	87.1	82.0	87.6
Species	93.1	92.2	90.7	78.8	72.3	74.6

3.2.4 Discussion

3.2.4.1 Boxed method

The method choice of Boxed or Standard is a trade-off between user input time and identification accuracy, influenced by the certainty of identification required. Method B is less time demanding than S, as only two landmark points must be added to the image (at diagonally opposite box corners) rather than the 20+ required to draw a polygonal region of interest overlay. This substantial time saving corresponds with 1% or less reduction in mean FPTP identification accuracy (>90% certainty) and no significant accuracy difference from the standard method (Wilcoxon's signed ranks test, $P>0.05$). However, the difference in mean identification accuracy between B and S was substantially greater (5 to 8%) at Coord3 (>95% certainty) and gave significantly lower accuracy than the standard method when tested with Wilcoxon's signed ranks test (at family, genus or species level).

Different uses of the DAISY system, such as biodiversity monitoring of a pollinator community or identifying a critical pollinator, may evaluate this trade-off differently. In biodiversity monitoring at species-level it may be very important to process as many specimens as possible and 90% certainty may be enough. In this situation it may be better to process many specimens and get 92% of identifications correct (B) than to process fewer and get 93% correct (S) (Table 3.5). When identifying a crucial pollinator fewer specimens are considered and a high level of certainty (>95%) may be the most important factor. In this case 79% accuracy (S) may be acceptable but 72% accuracy (B) too poor to be worthwhile (table 3.5). If both methods are offered to the user then this decision can be made on a case-by-case basis.

3.2.4.2 Attached method

The major benefit of the Attached method is that additional specimens (e.g. from museum collections) can be imaged for training. This has been reflected slightly in the trial, where 30 additional specimens could be used for A that were not available for S or B. This resulted in method A achieving slightly

greater accuracy than method S at genus-level (by around 0.5%), both at 90% and 95% certainty. However, at species-level method A was 2.4% less accurate than method S with 90% certainty and 4% less accurate with 95% certainty. At family, genus or species level, A was never significantly less accurate than S (Wilcoxon's signed ranks test, $P > 0.05$).

The small difference in accuracy between S and A is encouraging for the use of the NMK and NHM museum collections in future DAISY training. If these large specimen sources were accessible for DAISY training then method A may well out-perform method S, even at species level. The success with attached wings also encourages the development of a DAISY wing clamp system for live insects.

3.2.4.3 Taxonomic level

All three methods identify best to family level, then to genus, and finally to species. The same can generally be said of non-automated taxonomy. People usually find it easier to identify to higher than to lower taxonomic levels. For example, most people can recognise a bee and many a carpenter bee but few can distinguish species of *Xylocopa*. This makes intuitive sense as a more precise taxonomic level gives more options to choose from, as seen in Table 3.4.

If the same assortment of images is used in each case, identification at lower taxonomic levels also reduces the average training set size of each of those lower levels. Previous DAISY studies on Hymenoptera (e.g. Weeks *et al.*, 1999b) have found that large pool size and small training set size have led to reduced performance.

It is interesting that accuracy improved more when taxonomic level was raised from species to genus than it did when genus was raised to family (Table 3.5), even though genera and species were more similar in their numbers of possible taxa than families and genera (Table 3.4). This is presumably because there are venation characteristics that vary between same-family genera but little within those genera. This idea is supported by the species-level mis-identifications, which were almost all species in the same genera.

If FPTP identification provides sufficient certainty (90%) then identification to species-level is only 6% less accurate (at most) than family-level identification or 5% less accurate than genus-level. Working with a relatively well studied group such as bees, many species have been monographed and species-level discrimination makes all this information available to the ecologist. As the trade-off in FPTP accuracy is small, species-level identification seems the best general option. In cases where certainty

must be as high as possible (95%) the difference in accuracy between genus and species-level accuracy is greater (around 10%) so it is best that this remains a choice for the individual user.

3.3 Normalised polar thumbnail size

3.3.1 Introduction

The creation of Normalised Polar Thumbnails was outlined in section 2.3. NPTs are the pixel grids that are compared to obtain an identification. An example is shown in Fig. 3.18. They can be 20-80 pixels across, and the smaller the grid the greater the degree of sub-sampling. The sub-sampling maximises the signal to noise ratio of the image, and the conversion from Cartesian to polar allows spatially irregular regions of interest to be analysed. Grid resolution may affect performance. The default grid size used so far in DAISY analyses is 32 x 32 pixels but it has not been determined whether this is optimal for bee and wasp wings. Investigation of the effect of NPT size may also provide better understanding of how DAISY image analysis differs from human vision.

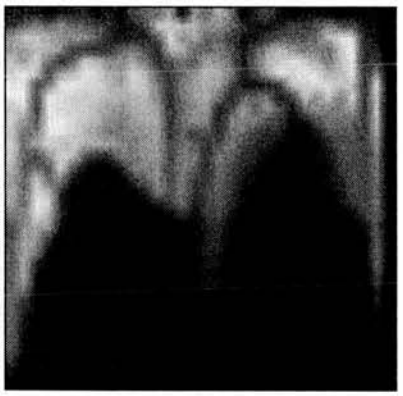


Fig. 3.18 – An example NPT, the wing venation pattern of *Megachile basalis* (a megachilid bee) transformed to a polar grid.

.

3.3.2 Method

The species-level image set was identified three times (by jack-knifing and calculation of mean values) with the NPT size set to 24 x 24, 32 x 32 and 48 x 48 pixels (shown as "24", "32" and "48" on result tables).

3.3.3 Results

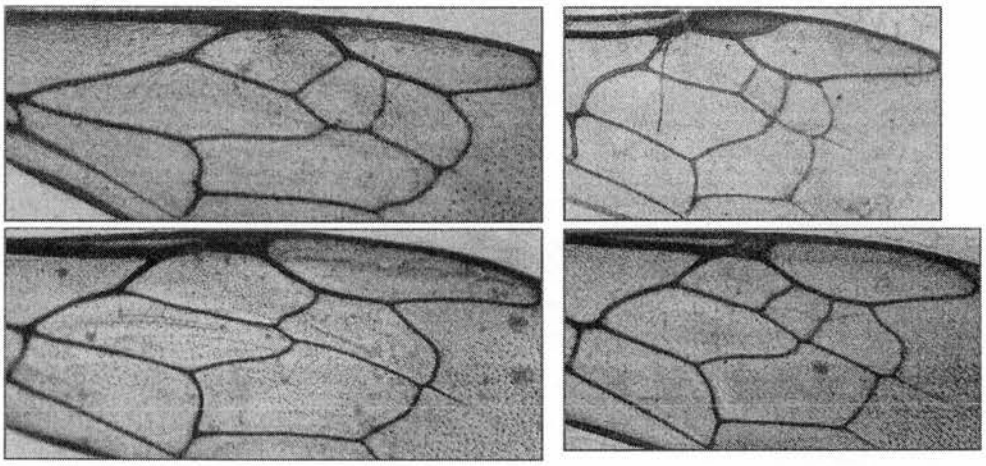
Changing NPT size only had a small impact on mean identification accuracy (<2% overall for 90% certainty and <5% for 95% certainty) (Table 3.6).

Table 3.6 – Full data table for impact of NPT size on identification accuracy. The Coord3 boxes are highlighted dark if they were unchanged and pale if they were changed by at least 75%.

Species	FPTP (%)									Coord3 (%)								
	Standard			Boxed			Attached			Standard			Boxed			Attached		
	24	32	48	24	32	48	24	32	48	24	32	48	24	32	48	24	32	48
Amegilla calens	100	100	80	100	80	80	80	80	80	20	100	0	20	0	0	20	20	20
Amegilla fallax	79	79	83	83	83	83	100	100	100	71	71	71	67	63	58	85	90	85
Apis mellifera	100	100	100	100	100	100	100	100	100	100	100	100	100	100	95	100	100	100
Bembix forcipata	100	100	100	100	100	100	100	100	100	100	100	100	92	92	92	100	100	94
Ceratina moerenhouti	83	83	100	83	83	83	100	100	100	83	83	83	50	33	17	100	100	100
Colletes sp.1	100	100	100	100	100	100	91	82	82	100	100	100	83	75	92	82	82	82
Delta hottentottum	40	40	40	60	60	60	50	57	57	0	0	0	0	40	40	14	14	14
Delta lepeleterii	71	71	71	57	71	57	72	64	73	57	57	43	29	43	29	27	9	9
Heriades sp.1	75	75	75	75	75	50	100	67	67	25	0	25	0	0	50	33	17	17
Lasioglossum sp.1	83	100	67	83	100	67	78	78	89	0	100	0	0	100	17	22	22	22
Macrogalea candida	95	95	95	95	95	95	95	95	95	86	86	86	95	95	95	95	95	90
Megachile gratiosa	100	100	100	80	80	100	50	50	50	20	100	20	40	80	40	0	0	0
Megachile sp.1	86	86	86	86	86	86	100	100	100	86	86	86	86	86	86	100	100	100
Megachile sp.4	88	88	88	75	75	75	100	100	100	88	88	88	13	13	13	86	57	71
Megachile sp.6	100	100	100	89	100	100	88	88	88	100	89	89	89	89	89	63	75	50
Meria sp.1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	93
Patellapis sp.1	83	92	92	83	83	83	100	100	100	50	67	67	58	58	58	100	100	100
Philanthus triangulum	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	94
Pseudapis sp.1	100	100	100	100	100	100	85	92	85	80	80	70	60	70	80	62	69	69
Tetralonia nigropilosa	100	100	100	100	100	100	100	100	100	100	100	75	75	75	75	75	75	75
Tetraloniella sp.2	100	100	100	100	100	100	100	100	100	100	75	100	50	25	25	25	100	50
Xylocopa somalica	100	100	100	100	100	100	94	94	94	100	100	100	100	100	100	94	94	94
Mean accuracy (%)	92.21	93.07	92.64	91.34	92.21	90.91	91.46	90.65	91.06	79.65	78.79	78.35	72.73	72.29	71.00	76.02	74.80	71.95

However, individual species reacted differently to NPT size. Many did not change with NPT size (blue on Table 3.6) but four bee species (*Amegilla calens, Lasioglossum* sp.1, *Megachile gratiosa* and *Tetraloniella* sp.2) changed by at least 75% in at least one method (yellow on Table 3.6). These species had venation patterns on the same spatial scale as the other species and no wing pigmentation to make them distinct (Fig. 3.19). The Boxed method had slightly more species affected by NPT size (12) than the other two methods (10).

Fig. 3.19 – The venation patterns of the four species sensitive to NPT size: *Amegilla calens* (top left), *Lasioglossum* sp.1 (top right), *Megachile gratiosa* (bottom left) and *Tetraloniella* sp.2 (bottom right).



With such great variation between species it is unwise to advocate a single NPT size that gives most accurate wing identifications overall. There was no consistent trend for the First Past the Post mean values but Coord3 gave 1 - 4% greater accuracy as NPTs became smaller in all three methods. This was especially pronounced for A, where NPT24 was 4% more successful than NPT48. However, the benefit of smaller NPTs did not continue as NPT20 gave FPTP = 91.06 and Coord3 = 73.58.

3.3.4 Discussion

In previous studies where DAISY was used to identify Hymenoptera, the default thumbnail size (32) was used without comment. The smaller NPT size of 24 gives better Coord3 results than 32 in all three methods but the improvement is only slight. As changing from 32 to 24 would also improve the FPTP in method A, and would only cause a small decrease in FPTP in the other methods, this suggests that 24 is the best NPT size of the three investigated. It is possible that 28 X 28 may be even better.

All the species had venation patterns on the same spatial scale and the sensitive species had no distinctive wing colouration so there is no obvious reason why these species were much more sensitive to NPT size than the another species.

It seems counterintuitive for the smaller NPT size to give the better overall result as the greater level of sub-sampling means that less information is encoded, and to human vision an image so 'blurred' would be more difficult to recognise. The benefit could come from a reduction in non-informative noise. The images of method A are the most likely to have pattern distortion from orientation perspective and wing curl and it is this method where the Coord3 results show the greatest improvement from decreasing NPT size.

3.4 Training set size

3.4.1 Introduction

The number of specimens used to construct the training pool has previously been shown to have a substantial effect on DAISY accuracy. As training set (TS) size increases, the number of additional correct identifications decreases until, at some critical size, adding more specimens would lead to no further improvement in identification accuracy (Weeks *et al.*, 1997; Weeks *et al.*, 1999b; Gauld *et al.*, 2000; Pajak, 2001; Watson *et al.*, 2004). For five ichneumonid wasp the most accurate identification was achieved by a training size of 20 (Weeks *et al.*, 1997, Fig. 3.20), for 49 biting midge species the asymptote was reached by around 15 (Gauld *et al.*, 2000; Fig. 3.21).

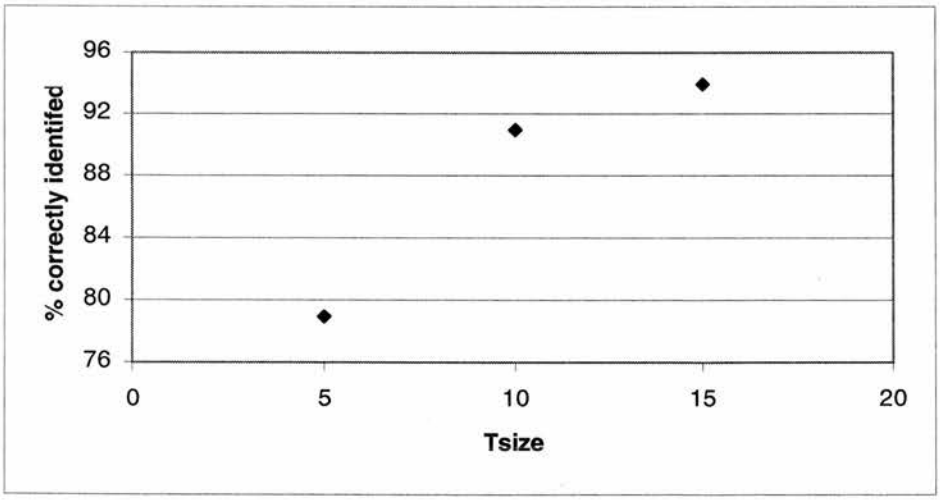


Fig. 3.20 – Weeks *et al.* (1997), the effect of TS size on identification accuracy of five ichneumonid wasp species.

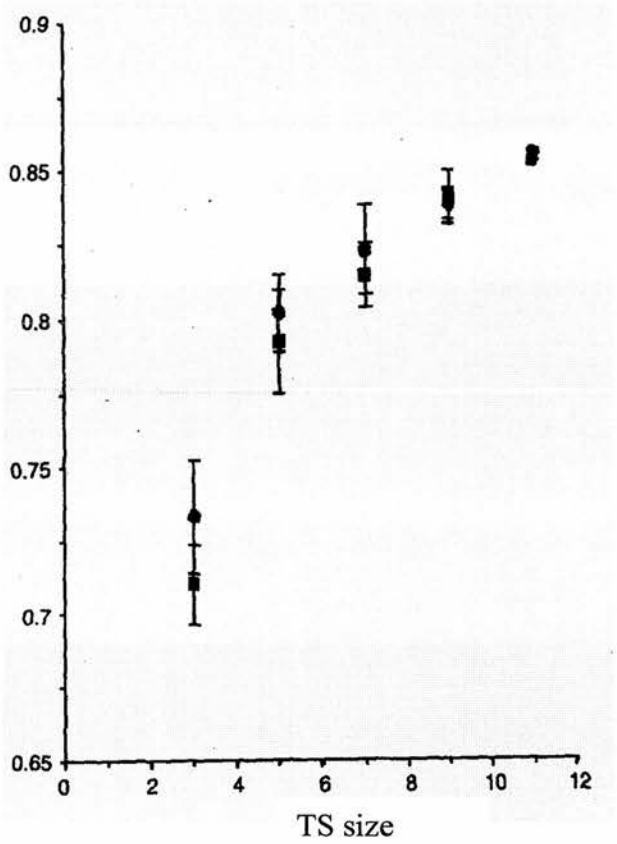


Fig. 3.21 - Gauld *et al.* (2000), DAISY identified correctly a greater proportion of specimens as the TS size was increased. Results based on 49 species of *Culicoides* and *Forcipomyia* biting midge.

Larger training sets are likely to span a broader range of intraspecific variability but they take more time to prepare and their use may not be feasible if a species is rare; small training sets are much quicker to prepare, but they may poorly represent intraspecific variation.

When producing a small training set from a larger pool of images, it is possible either to use a random assortment, or to select the best images for inclusion. If the best images are used then the small set may still span a broad range of intraspecific variation. However, it will not realistically reflect a situation in which few images are available. In pollination ecology studies it will be most time-efficient to catch and image the minimum number of specimens essential for DAISY training, leading to small training sets that are not optimal subsets of larger pools of images. For this reason it was decided to remove images at random when reducing an image set for a comparison run.

3.4.2 Method

The specimens had been imaged without conscious ordering and given consecutive file numbers so processing them in order of file number was considered to make them approximately random (in retrospect a more reliable method of randomisation should have been used). A training set of *n* images used the first *n* images when they were listed in order of file number. Deleting the last image in each list removed one image at random, so a start was made with TS10, working down to TS4.
The same methodology was used to analyse the effect of TS size in each of the imaging methods, first S, then B and finally A. This work was done before the analysis of NPT size (section 3.3) so the default polar thumbnail grid size of 32 was used throughout.

1) Images removed so only the first ten in the file list remained.
2) Set identified for a training set of 10 (TS10).
3) Final image for each species deleted and set re-identified.
4) Stage 3 repeated until the training set reached T4.

It should be noted that although an image set of ten images is referred to as T10 only nine of the images are used in training, as one is always functioning as the "unknown" image to be identified against the rest.

3.4.3 Results

The FPTP (90%+ certainty) results are discussed first, moving on to the Coord3 (95%+) results. The FPTP results for the three methods are shown in Table 3.7. Method S achieved 100% accuracy in all species at any of the TS sizes. B had reduced performance at smaller TS sizes in three species. Method A had reduced performance in three different species, spread throughout the full range of TS size. While the number of images correctly identified always increased with TS size this didn't always keep pace with the TS number so increasing TS size sometimes decreased the overall % accuracy. In *Xylocopa somalica,* a single specimen, present in all training sets, was consistently mis-identified. At TS4 this caused a substantial reduction in accuracy (25%) but by TS10 it had less impact (10%).

The mean FPTP identification accuracies using S, B and A are plotted in Fig. 3.22. Method S was 100% accurate at all TS sizes. Methods B and A showed greater sensitivity. This supports the initial suggestion made in section 3.2, when species with smaller training set sizes varied the most between methods in identification accuracy. Accuracy in B increased with TS size reaching 100% by TS8. A fluctuated around 97% accuracy, never reaching 100%. B and A both have large standard errors.

The Coord3 accuracies for S, B and A are shown in Table 3.8. In method S, two species dropped below 100% accuracy as TS size decreased. More identifications failed (as well as fewer succeeding) at smaller TS sizes. In method B, five species had accuracy below 100%. Three of these increased in accuracy with each increase in TS size but *Colletes* and *Pseudapis* sometimes decreased. In method A, four species changed in accuracy with TS size. *Xylocopa* increased in accuracy with each TS size increase, first dramatically then gently. The other three species fluctuated around 92%.

The mean Coord3 identification accuracies using methods S, B and A are plotted in Fig. 3.23. Accuracy in method S increased first steeply (as fewer identifications failed) then more gradually (as a constant single failure was joined by an increasing number of passes), reaching 100% by TS9. A similar pattern of accuracy increase was seen in method B but the highest accuracy reached was 97%. In method A there was an initial increase but above TS4 it fluctuated around 92%.

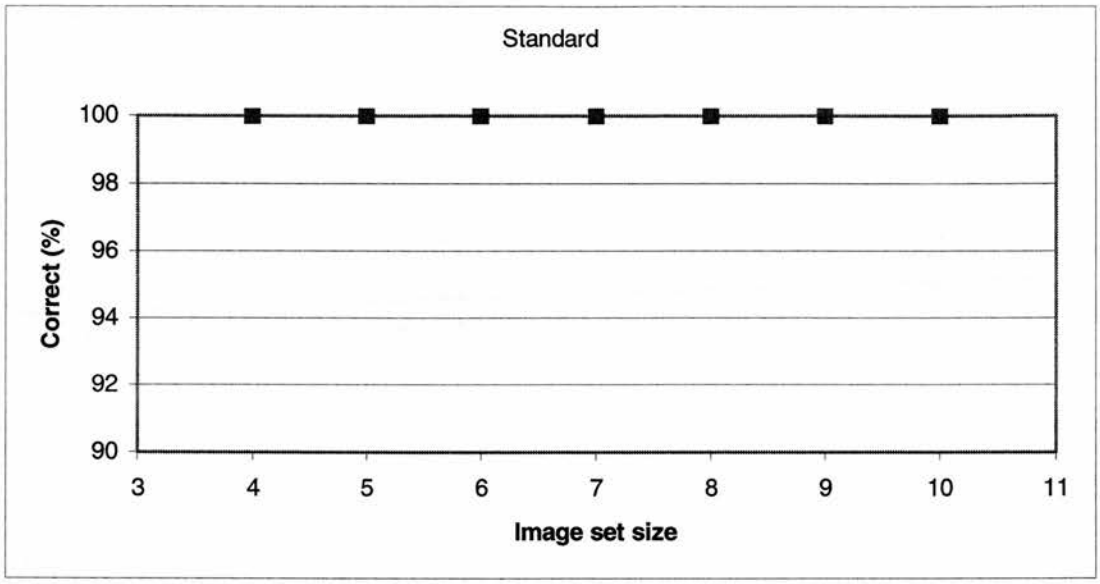
Table 3.7 – Number of specimens identified incorrectly FPTP for S (top), B (middle) and A (bottom) with TS 4 – 10. Cases in which mis-identifications were made, and associated species names, are highlighted.

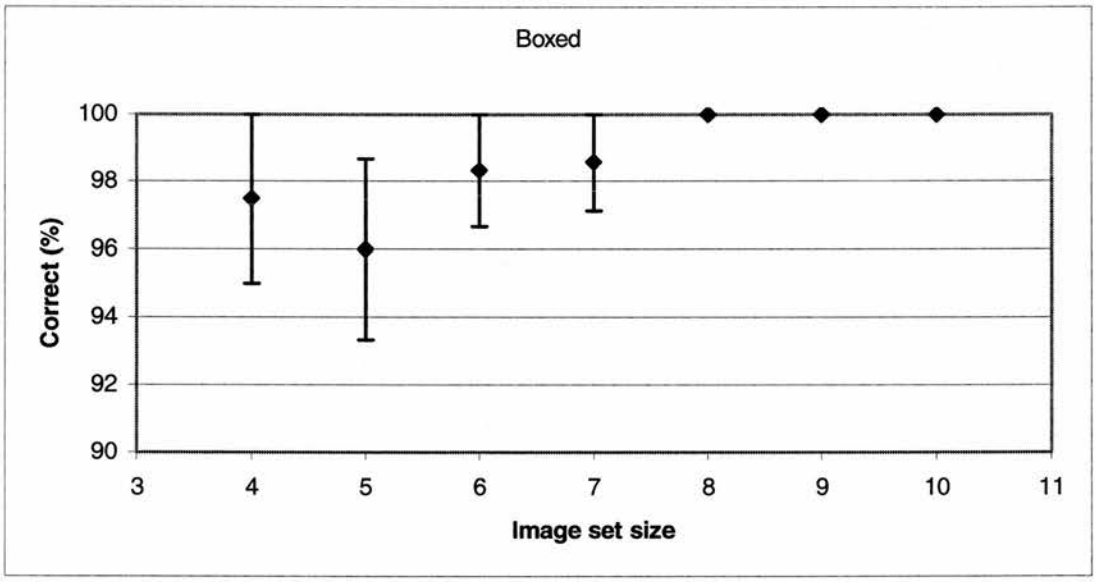
Standard method	Image set size						
Species	4	5	6	7	8	9	10
Amegilla fallax	0	0	0	0	0	0	0
Apis mellifera	0	0	0	0	0	0	0
Bembix forcipata	0	0	0	0	0	0	0
Colletes sp.1	0	0	0	0	0	0	0
Macrogalea candida	0	0	0	0	0	0	0
Meria sp.1	0	0	0	0	0	0	0
Patellapis sp.1	0	0	0	0	0	0	0
Philanthus triangulum	0	0	0	0	0	0	0
Pseudapis sp.1	0	0	0	0	0	0	0
Xylocopa somalica	0	0	0	0	0	0	0
Mean accuracy (%)	100	100	100	100	100	100	100
sample s.d.	0	0	0	0	0	0	0
s.e. of mean	0	0	0	0	0	0	0

Boxed method	Image set size						
Species	4	5	6	7	8	9	10
Amegilla fallax	0	0	0	0	0	0	0
Apis mellifera	0	0	0	0	0	0	0
Bembix forcipata	0	0	0	0	0	0	0
Colletes sp.1	0	0	0	0	0	0	0
Macrogalea candida	1	0	0	0	0	0	0
Meria sp.1	0	0	0	0	0	0	0
Patellapis sp.1	0	1	0	0	0	0	0
Philanthus triangulum	0	0	0	0	0	0	0
Pseudapis sp.1	0	1	1	1	0	0	0
Xylocopa somalica	0	0	0	0	0	0	0
Mean aacuracy (%)	97.5	96	98.33	98.57	100	100	100
sample s.d.	7.91	8.43	5.27	4.52	0	0	0
s.e. of mean	2.5	2.67	1.67	1.43	0	0	0

Attached method	Image set size						
Species	4	5	6	7	8	9	10
Amegilla fallax	0	0	0	0	0	0	0
Apis mellifera	0	0	0	0	0	0	0
Bembix forcipata	0	0	0	0	0	0	0
Colletes sp.1	0	0	0	1	1	2	2
Macrogalea candida	0	0	0	0	0	0	0
Meria sp.1	1	0	0	0	0	0	0
Patellapis sp.1	0	0	0	0	0	0	0
Philanthus triangulum	0	0	0	0	0	0	0
Pseudapis sp.1	0	0	0	0	0	0	0
Xylocopa somalica	1	1	1	1	1	1	1
Mean	95	98	98.33	97.14	96.25	96.67	97
sample s.d.	10.54	6.32	5.27	6.03	8.44	7.5	6.75

Fig. 3.22 – Mean FPTP identification accuracy plotted at TS 4 – 10. Bars show standard errors.





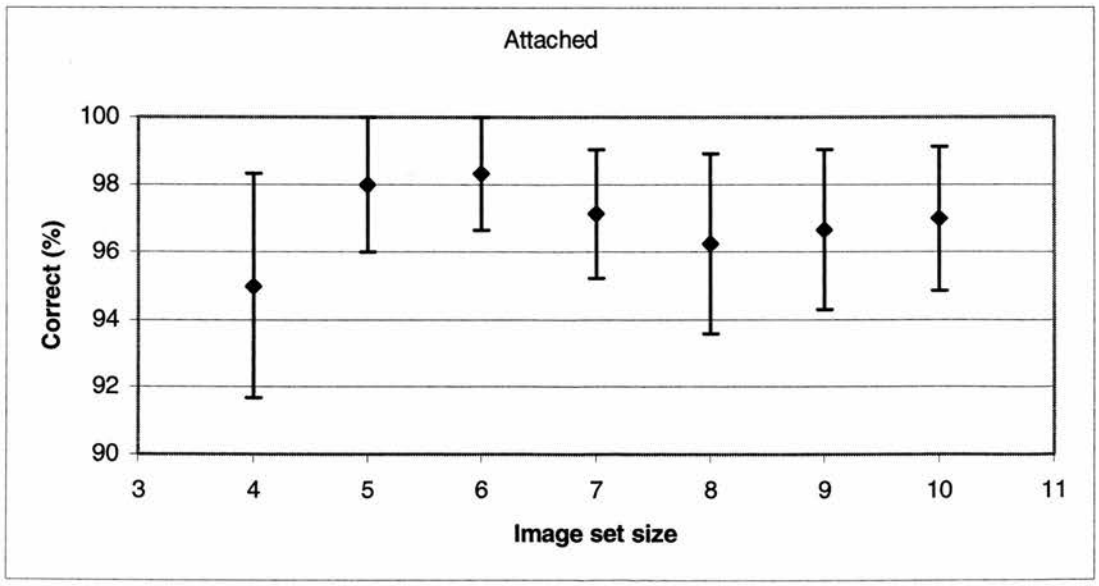


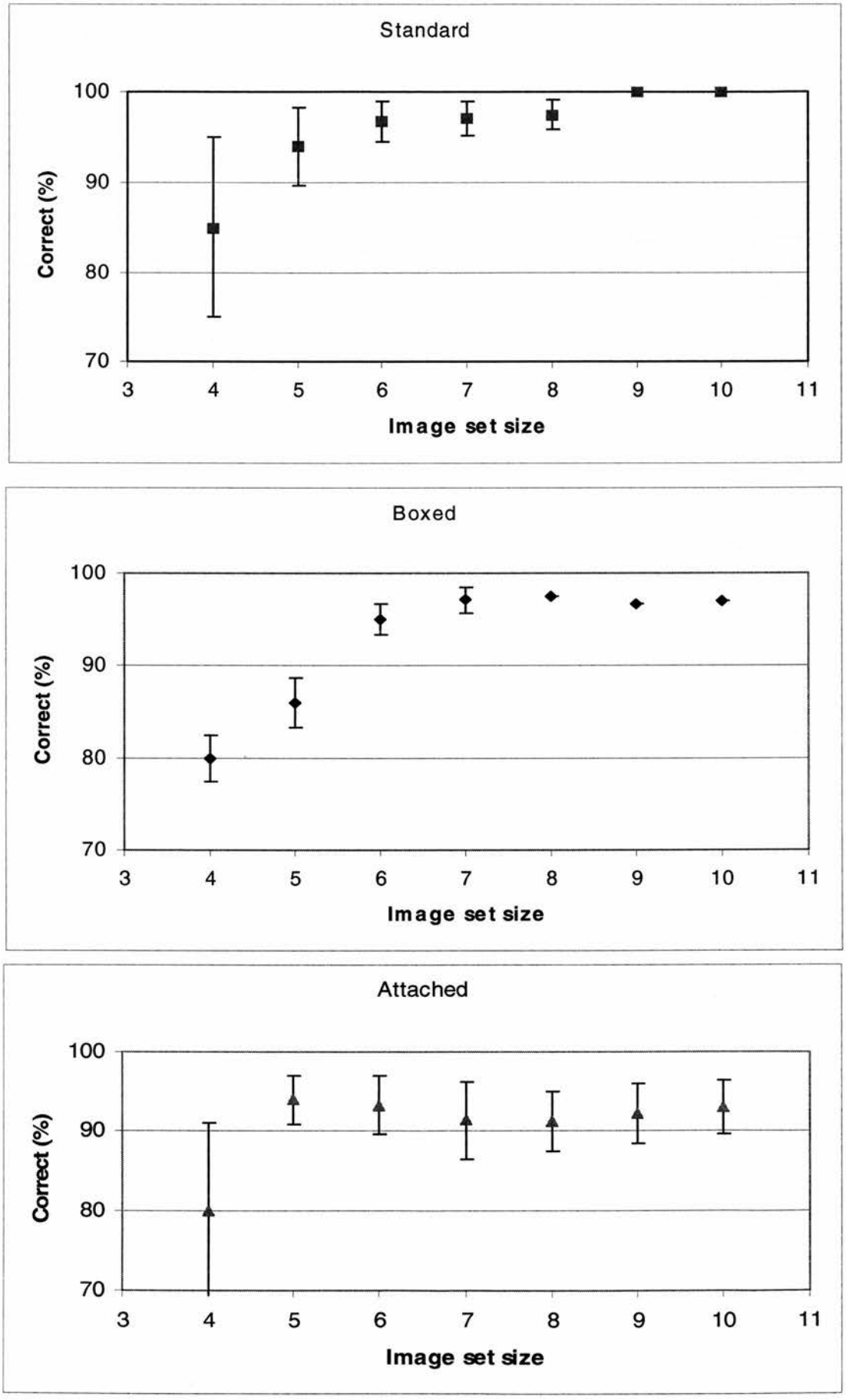
Table 3.8 – Coord3 results for S (top), B (middle) and A (bottom) with TS 4 – 10. Cases in which identifications failed to be made correctly with 95% certainty, and associated species names, are highlighted.

Standard method	Image set size						
Species	4	5	6	7	8	9	10
Amegilla fallax	0	0	0	0	0	0	0
Apis mellifera	0	0	0	0	0	0	0
Bembix forcipata	0	0	0	0	0	0	0
Colletes sp.1	0	0	0	0	0	0	0
Macrogalea candida	2	2	1	1	1	0	0
Meria sp.1	0	0	0	0	0	0	0
Patellapis sp.1	3	1	1	1	1	0	0
Philanthus triangulum	0	0	0	0	0	0	0
Pseudapis sp.1	0	0	0	0	0	0	0
Xylocopa somalica	0	0	0	0	0	0	0
Mean accuracy (%)	85	94	96.67	97.14	97.5	100	100
sample s.d.	31.6	13.5	7.03	6.03	5.27	0	0
s.e. of mean	10	4.27	2.22	1.91	1.67	0	0

Boxed method	Image set size						
Species	4	5	6	7	8	9	10
Amegilla fallax	0	0	0	0	0	0	0
Apis mellifera	0	0	0	0	0	0	0
Bembix forcipata	0	0	0	0	0	0	0
Colletes sp.1	0	1	0	0	0	0	0
Macrogalea candida	2	1	1	0	0	0	0
Meria sp.1	0	0	0	0	0	0	0
Patellapis sp.1	2	2	1	1	1	1	1
Philanthus triangulum	2	2	0	0	0	0	0
Pseudapis sp.1	2	2	1	1	1	2	2
Xylocopa somalica	0	0	0	0	0	0	0
Mean accuracy (%)	80	86	95	97.14	97.5	96.63	96.97
sample s.d.	25.82	16.47	8.05	6.03	5.27	7.5	6.75
s.e. of mean	8.16	5.21	2.55	1.91	1.67	2.37	2.13

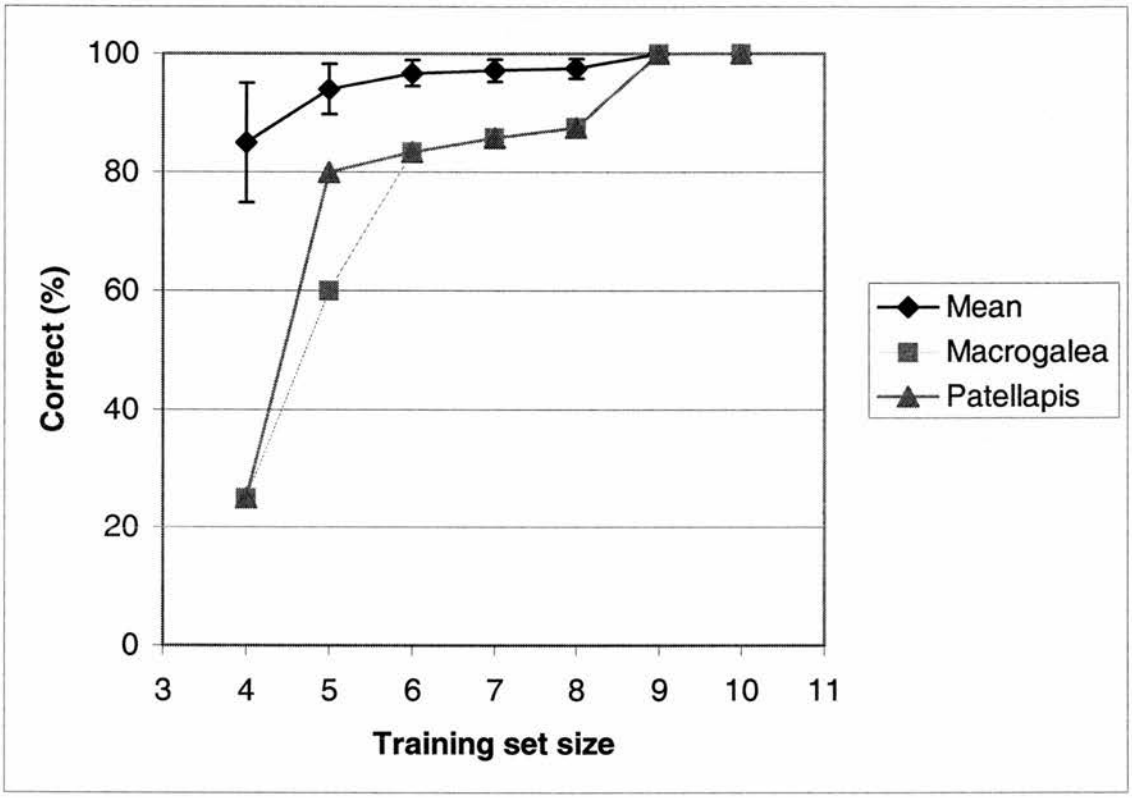
Attached method	Image set size						
Species	4	5	6	7	8	9	10
Amegilla fallax	0	0	0	0	0	0	0
Apis mellifera	0	0	0	0	0	0	0
Bembix forcipata	0	0	0	0	0	0	0
Colletes sp.1	2	1	1	2	2	2	2
Macrogalea candida	0	0	0	0	0	0	0
Meria sp.1	2	1	2	3	2	1	1
Patellapis sp.1	0	0	0	0	0	0	0
Philanthus triangulum	0	0	0	0	0	0	0
Pseudapis sp.1	0	0	0	0	2	2	3
Xylocopa somalica	4	1	1	1	1	1	1
Mean accuracy (%)	80	94	93.33	91.43	91.25	92.22	93
sample s.d.	35	9.66	11.65	15.36	11.86	11.77	10.59
s.e. of mean	11.1	3.06	3.69	4.86	3.75	3.72	3.35

Fig. 3.23 – Mean Coord3 identification accuracy for methods S, B and A, plotted at TS 4 – 10. Bars show standard errors.



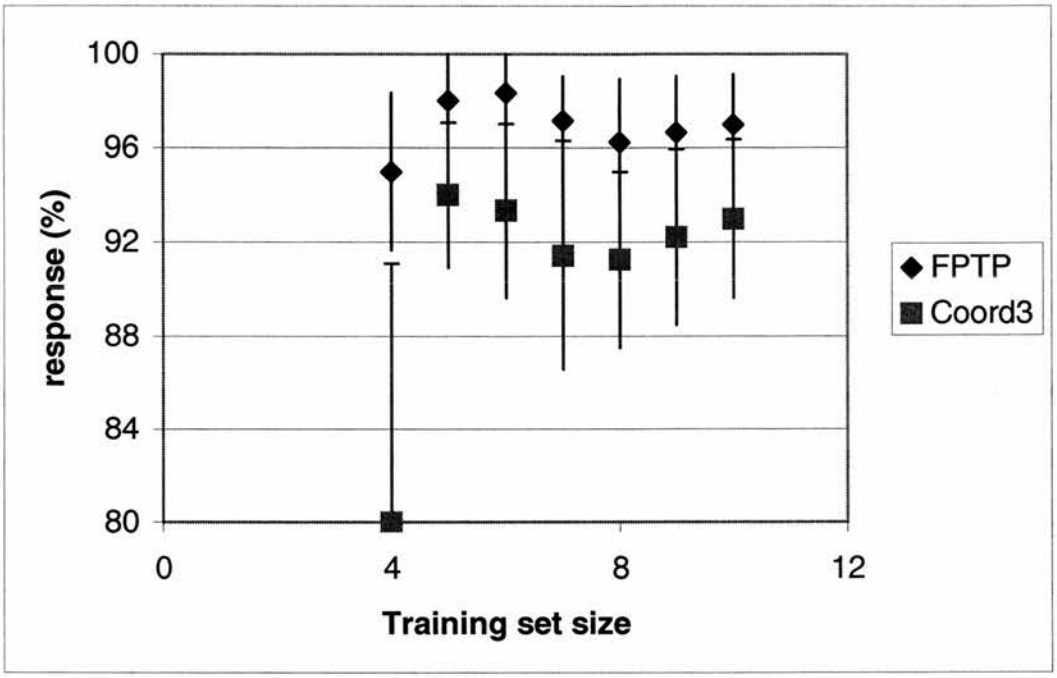
The large number of species that show 100% accuracy, irrespective of training set size dominate the mean values so they vary little with TS size. This is illustrated in Fig 3.24. *Macrogalea candida* and *Patellapis* sp.1 were the only species to have their Standard method identification reduced from 100%, even at the 95%+ certainty demanded by Coord3. When plotted independently of the other species they each display a more dramatic curve than that of the mean.

Fig. 3.24 – Standard method Coord3 accuracy with TS size for *Macrogalea candida, Patellapis* sp.1 and the mean. Error bars on mean show standard errors.



The patterns seen for the Attached method at FPTP and Coord3 (Fig. 3.25) do not form simple, asymptotic curves. When 90%+ accuracy was specified (FPTP) the lowest accuracy was 95% at TS4 but this was only 3% lower than the highest accuracy, 98% at TS6. Above TS6 accuracy reduced to 96% by TS8 then increased again, staying close to 97%. When 95%+ accuracy was specified (Coord3) TS4 gave 80% accuracy, 11% below any larger TS size. This low value was mainly due to a zero for *Xylocopa somalica* (discussed further in 3.4.4). The highest accuracy of 94% was obtained by increasing TS size by a single image to TS5. It is this large initial response that makes these data resemble those for the other two methods. Above TS5, accuracy dipped to TS8 then increased again, staying close to 92%. It is likely that both dips are not general patterns, merely reflections of the particular set of images used in this case. The standard errors were much larger for this method than the other two, suggesting that further data are required to confirm any patterns.

Fig. 3.25 – Method A, FPTP and Coord3 mean responses to TS size. Error bars show standard errors.



3.4.4 Discussion

Data from the Standard and Boxed methods agree with the findings of previous DAISY studies (Weeks *et al.*, 1997; Weeks *et al.,* 1999b; Gauld *et al.,* 2000; Pajak, 2001; Watson *et al.,* 2004), in that the addition of specimens to training sets produced a curve of diminishing returns (see Figs. 3.20 and 3.21).

Table 3.9 – Summary of the accuracy asymptotes in previous DAISY studies.

Study	Specimens identified	Accuracy asymptote
Weeks *et al.* (1997)	5 ichneumonid wasp spp.	15 – 20
Weeks *et al.* (1999b)	49 biting midge species	15
Pajak (2001)	4 bumblebee spp.	Not reached, 100% by 20
Watson *et al.* (2004)	35 live moth spp.	50

The values at which asymptotes of accuracy were reached in previous DAISY studies are shown in Table 3.9. In these studies, the asymptote was not reached before TS15. In the present study, it is achieved sooner, at around TS6 or TS7. The asymptote at low TS size may be because this image set brings together two favourable factors, found separately in previous studies but never before in combination. These are as follows:

1. Simplicity of pattern tones. The study organisms (bees and wasps) have a black and white wing pattern caused by venation (as was the case in Weeks *et al.,* 1997; Weeks *et al.,*1999b; and

Pajak, 2001). This 'analogue' pattern may be better suited to computer vision than the many pixel intensities of a lepidopteran scale pattern (Watson *et al.* 2004).

2. Taxonomic diversity of the study organisms. The organisms in this study came from 21 genera and seven families. This is much more diverse than in Weeks *et al.* (1997), Weeks *et al.* (1999b) and Pajak (2001), where the study organisms came from just one or two genera from a single family, and it has more in common with Watson *et al.* (2004), which used species from 30 genera.

Another important influence in this comparison is that DAISY has become more accurate over time. The current version of DAISY would probably produce more accurate identifications of the same data sets than those published.

The early asymptote in this study supports an idea put forward by Gauld *et al.* (2000), that for routine identifications, only relatively small training sets need to be constructed. Indeed, Weeks *et al.* (1999b) found that even with T3 more than 70% of identifications were correct. However, whilst it takes longer to create larger training sets there is no time penalty in the identification phase, so if more training images are available they should be used (Weeks *et al.*, 1999b).

The new methods (B and A) appear to be more sensitive to TS size than method S. They produced some FPTP mis-identifications, when S managed 100% accuracy even at small TS sizes (Fig 3.22), and caused twice as many species to drop below 100% Coord3 accuracy (five for B and four for A) as method S (two). This is not surprising as they have greater intraspecific variation due to orientation and larger training set sizes represent variation more fully.

An additional problem arises here in that the metric Coord3 is based on the identities of three nearest neighbours so it is poorly suited to image sets as small as four. In the case of T4 a single distorted image could lead to a Coord3 value of 0% as this image is essential as a nearest neighbour to the other three. This can be seen in the Attached method T4 result of zero for *Xylocopa somalica.* When the images are examined the venation pattern of the 4th one is badly distorted by wing tilt. It can be seen in comparison with the 3rd image in Fig 3.26.

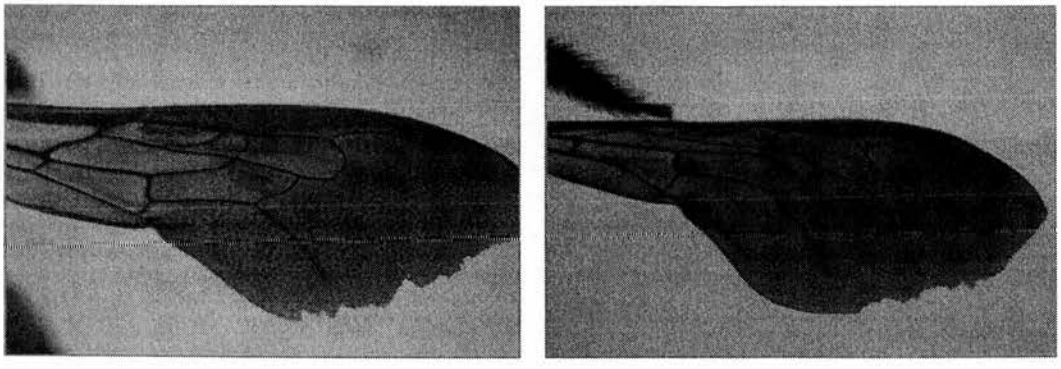


Fig. 3.26 – The 3rd (left) and 4th (right) images of *Xylocopa somalica* for method A. The venation pattern of the 4th image is distorted by wing tilt.

The misidentification of this image has reduced the FPTP performance to 75% but for Coord4 all the way to 0%. If the results had come from the average of many deletion sequences this problematic image would usually have been removed before TS4 and the average Coord3 result may well have been higher.

In the Attached method the data suggest an asymptote has been reached by T6, but the method A curves in Figs. 3.22 and 3.23 were not simple curves, as an increase in training set size sometimes led to a decrease in performance. When wings are imaged in a range of orientations, overall success can be greatly decreased by a small number of problematic images. This was found to be a problem with live moth images (Watson *et al.,* 2004) and is likely to be an issue with the Undamaged method images. If only problematic images are included as the training set is expanded then this may explain why enlarging a training set can make it perform more poorly. In method A, a reduction in performance with an increase in TS size can be seen in *Colletes, Meria* and *Pseudapis*. The sixth *Meria* image and the ninth *Pseudapis* image are compromised by pattern distortion but there is no obvious fault to the human eye in the seventh *Colletes,* seventh *Meria* or eighth *Pseudapis* images.

This entry order problem is akin to the "greedy algorithm" problem in phylogenetic analysis. Taxa can be added to a cladistical tree in several ways, the easiest of which is "AS IS", i.e. the order the taxa are entered into the data set. This can lead to a locally optimal solution because the best choice at a given point in tree construction may not be the best option overall (for the global optimum); but having made a decision the program cannot revise it later (Kitching *et al.,* 1998). Using image file name order to randomise image deletion was an AS IS approach, this method may not have been a reliable way to randomise images (as individuals caught at a single site or on a single day would be concurrent in file number) and this method may be producing sub-optimal identification accuracy. One way proposed to deal with greedy algorithms is to apply a majorization operator, to make a distribution "less unequal" (Menon, 2004), but the simplest solution is to repeat an ASIS analysis many times with the input order randomised between replicates (Kitching *et al.,* 1998). Beginning with a set of ten images and applying this approach to image deletion, there could be 151200 (10 X 9 X 8 X 7 X 6 X 5) deletion sequences. Once the identification accuracies have been calculated for all deletion sequences a mean accuracy could be taken for each TS size. These replications would take far too long done manually but it may be possible to code them into DAISY algorithms. There may be other lessons DAISY could learn from cladistic programs.

3.5 Summary and the way forward

The boxed and attached methods provided accuracy almost as high as the standard method, so it is worth bringing them into general use. The substantial time saving of box-cropping as opposed to drawing polyROIs corresponds with 1% or less reduction in accuracy when 90% certainty was enough. If 95% certainty was required then the difference was more substantial (up to 8%). Different users of the system must make an informed decision about the certainty they require and the time they are willing to invest to achieve that. Use of the attached method made some extra specimens available for training, this resulted in method A achieving 0.5% greater accuracy than method S at genus level and only 4% lesser accuracy (at most) at species-level. The small accuracy difference when wings remained attached is encouraging for the use of live and museum specimens.

Identification is more accurate at higher taxonomic levels but the difference between family-level and genus-level accuracy is generally small and probably insufficient to merit the poorer depth of information. If 90% certainty is enough then the difference between genus-level and species-level accuracy is only small (5% at most) so species-level discrimination is advised. If 95% certainty is necessary than the difference is greater (10% at most) and an accuracy / information value trade-off must be made my the user.

Changing the size of the Normalised Polar Thumbnails produced for analyses had a dramatic impact on the identification of some species but only a small impact on mean accuracy. Overall a 24 pixel thumbnail was identified more accurately than the 32 pixel thumbnail previously recommended.

There were insufficient specimens available for full investigation of training set size. Generally the addition of specimens to training sets produced a curve of diminishing returns. This supported the findings of previous DAISY studies (Weeks *et al.*, 1997; Weeks *et al.,* 1999b; Gauld *et al.,* 2000; Pajak, 2001; Watson *et al.,* 2004), however the asymptote was reached with fewer training specimens (just six or seven). This is encouraging data for practical application of DAISY as large training sets are very time consuming to produce and require specimens that may be unavailable to a field ecologist.

This trial included only bees and a few wasps. While these may be the most important insects for *Acacia* pollination they are only a small subset of the total *Acacia* visitors. What would be most useful is a generic training set including bees, wasps, flies, butterflies and beetles, sufficient specimens have been collected and identified to make this feasible (detailed in Appendix 1). For comparability between these differing taxa it would be necessary to select the entire wing (or the right half of the carapace in the case of beetles). However, this would be a huge undertaking and may only be used by a small

number of researchers. It may be a better investment of time to focus on the insect visitors to a cultivated plant, as this is likely to include fewer species, be of interest to many more researchers and have value as a marketable product.

Chapter 4 – DAISY identification of pollen

4.1 Introduction

4.1.1 Aims

The direct aim of this research was to assess the potential for DAISY to be a pollen identification fieldwork tool. For this reason

- Methods and equipment were kept as simple as possible.
- Pollen grain images were not standardised with regard to orientation to maximise performance, as a) this would be difficult and time consuming in the field, and
  b) pollen identity must be known to determine which is the "standard" orientation.

The indirect aim was to present challenges to DAISY identification that have not been faced in published trials. Wing venation was a two-dimensional, greyscale pattern (e.g. Weeks *et al.,* 1997; Gauld *et al.*, 2000); wing scale pattern involved colours and subtleties of tone but was still two-dimensional (Watson *et al.,* 2004); but pollen grains are three-dimensional, coloured and often complex in form.

Pollen grains of one species can look very different from alternative angles (Fig. 4.1). Views were taken at random and all were included in analysis.

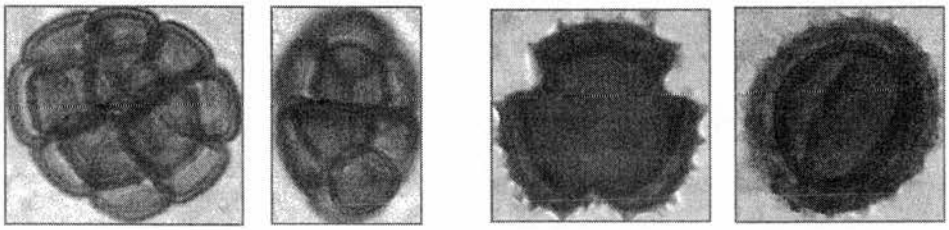
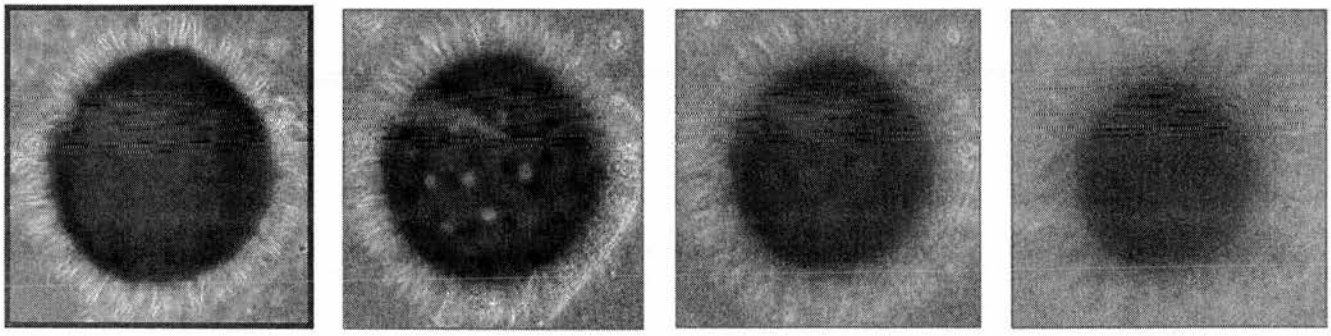


Fig. 4.1 – Two views of *Acacia brevispica* (left) and *Kleinia abyssinica* (right) pollen.

Small depth of field can be a problem when attempting to capture high-quality images of small, three-dimensional specimens, as grains can look very different at different focal planes (Fig. 4.2). Oliver *et al.* (2000) recognised this problem when imaging insects for virtual reference collections and suggested the use of photo montage software (e.g. 'Auto-Montage') to combine image slices. As simplicity is essential to DAISY identification this route has not been taken. Instead, pollens have been imaged in equatorial view, this was generally achieved by imaging a grain at maximum diameter.

Fig. 4.2 – *Abutilon grandiflorum* grain at four focal planes. The image on the far left was taken in equatorial view and could be included in DAISY identification.



The most successful automated identification of pollens to date has used surface features (see Section 4.1.3). The equatorial view was instead chosen as this approach is applicable to pollens of all sizes. Many of the smaller pollens had no distinct surface features when viewed under light microscope but all had a shape in equatorial view.

4.1.2 Non-automated pollen identification

Pollen taxonomy is important when investigating the pollens carried by insects (Dafni, 1992). Weber (1998) gives a useful review of pollen identification and Faegri (1992) is a well-respected pollen analysis textbook. Sawyer (1981) and Hodges (1984) focus on honeybee pollen loads. Kesseler & Harley (2004) include a section on pollen imaging.

Grains may occur singly (*monads*), or may be grouped as *tetrads* or *polyads*. Pollen size (maximal diameter) can vary from 2 microns to 250 microns. Size can also vary intraspecifically, and this variation can be taken as a feature in its own right. It is most common for pollen grains to be spherical or elliptical (Fig. 4.3) (Weber, 1998). Grains must be viewed from many dimensions and using different planes of focus if their shape is to be fully understood.

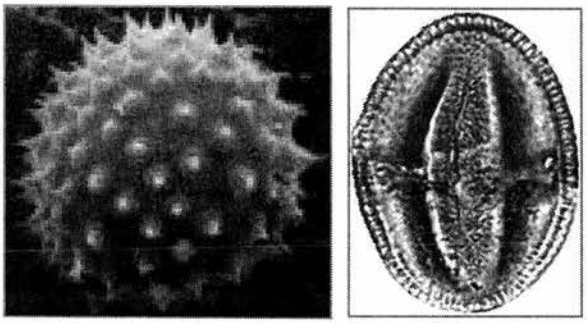


Fig. 4.3 – *Abutilon* (left) and *Grewia* (right) pollen grains are spherical and elliptical. Images from the African Pollen Database website hosted by medias.obs-mip.fr

The degree of elongation of elliptical grains can be used in identification. Kapp (1969) suggested using the ratio of polar to equatorial diameters, the P/E index, and suggested descriptive phrases to refer to

certain ratios (Table 4.1) (Fig. 4.4).

Table 4.1 – P/E index of Kapp (1969) to describe elliptical pollen.

Phrase	Description	P/E
Perprolate	Very elongated	> 2.00
Prolate	Slightly elongated	1.30 – 2.00
Subspheroidal		0.75 – 1.30
Oblate	Slightly flattened	0.50 – 0.75
Peroblate	Very flattened	< 0.50

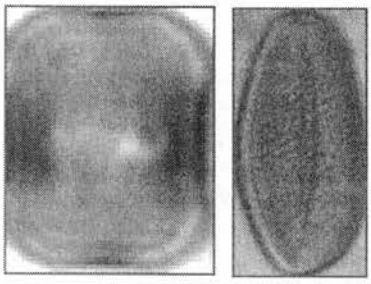
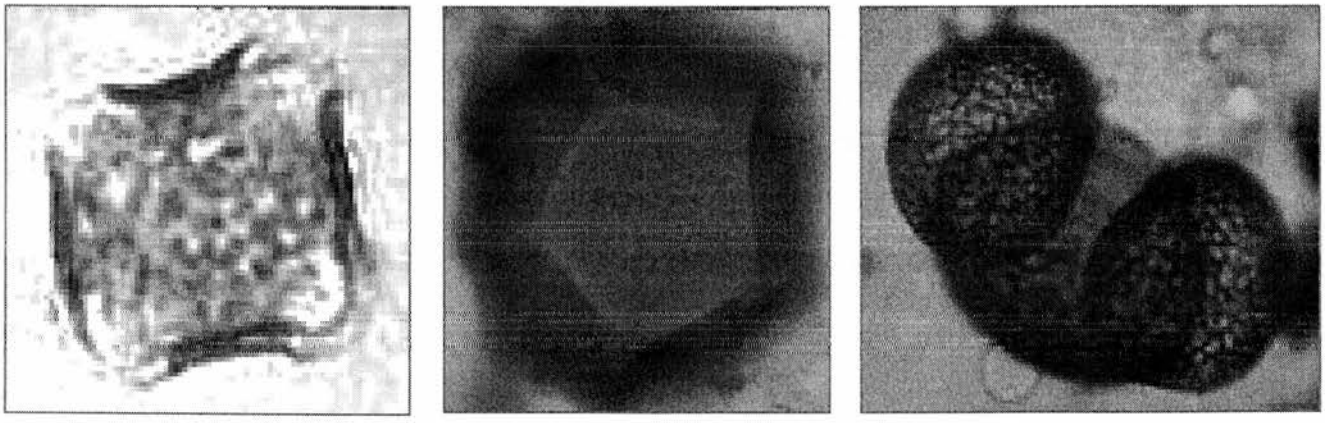


Fig. 4.4 – Using Kapp's (1969) P/E index, *Echiochilon* (left) would be described as prolate and *Gloriosa* (right) as *perprolate*.

The outer casing of a pollen grain (*exine*) has micropores and larger apertures: pores or furrows. The apertures are areas of thinned exine, exit points for a pollen tube in germination. The exine may angle at these thinner points to make the grain triangular, square or polygonal. Air bladders may extrude to affect the shape, and these bladders may be clear or darkly pigmented (Weber, 1998) (Fig. 4.5).

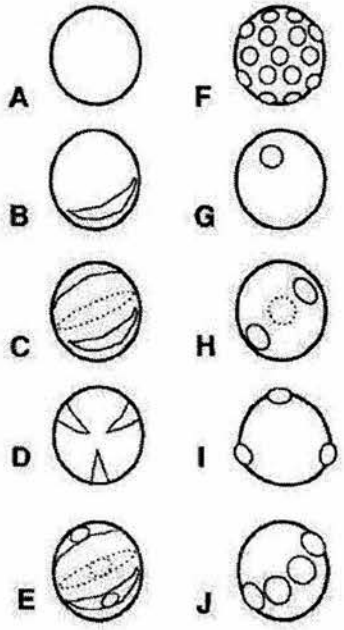
Fig. 4.5 – *Fraxinus* (left), *Portulaca* (middle) and *Pinus* (right) pollen illustrate that pollen can be square, polygonal or have its shape modified by air bladders. *Fraxinus* and *Pinus* from Weber (1998).



The numbers and positioning of pores and furrows are important to identification. Pores are generally circular, sometimes elliptical and may be single or multiple. When describing pollen a prefix indicates

the number of pores: *mono-, di-, tri-* or *peri-porate* (with one, two, three or many pores). Members of the Graminaceae are generally *monoporate*. *Triporate* grains are also common. Grains with at least four pores along the equator are known as *stephanoporate*.

Furrows are slits or boat-shaped areas. Grains with furrows are described as *colpate*. The same prefixes are used as with pores to express the number and orientation of furrows, e.g. *tricolpate*. Some pollens have both furrows and pores, and these are called *colporate*. In these cases the pores are within the furrows. Some types of pollen apertures are illustrated in Fig, 4.6 (Weber, 1998).



Fig, 4.6 – Ten possible arrangements of pores and furrows (structures in dotted lines are on the far side of pollen grains): A. *inaperturate*; B. *monocolpate*; C. *tricolpate*, equatorial plane; D. *tricolpate*, polar view; E. *tricolporate*; F. *periporate*; G. *monoporate*; H. *triporate*, equatorial plane; I. *triporate*, polar view; and J. *stephanoporate*. From Weber (1998).

Surface sculpturing into ridges and depressions may distinguish one species from another, although this is hard to see with normal light microscopy. A pollen with a smooth surface is termed *psilate*. Patterns may be *reticulate* (net-like), *striate* (parallel ridges), or *rugulate* (irregular). The roof of the exine may have distinctive projections. The different shapes of projection add to the description of a grain, e.g. *baculate* (rod-like), *verrucate* (lumpy), *echinate* (spiny) or *clavate* (club-like). These are illustrated in Figs. 4.7 and 4.8.

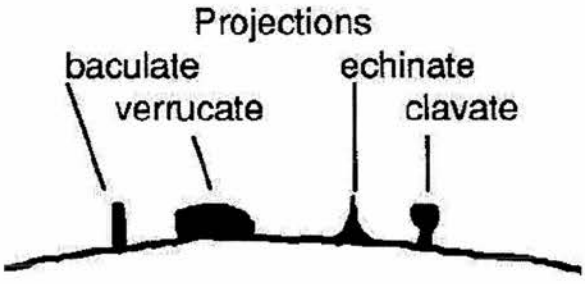


Fig. 4.7 - Different shapes of projection from the roof of the pollen exine. Taken from Weber (1998).

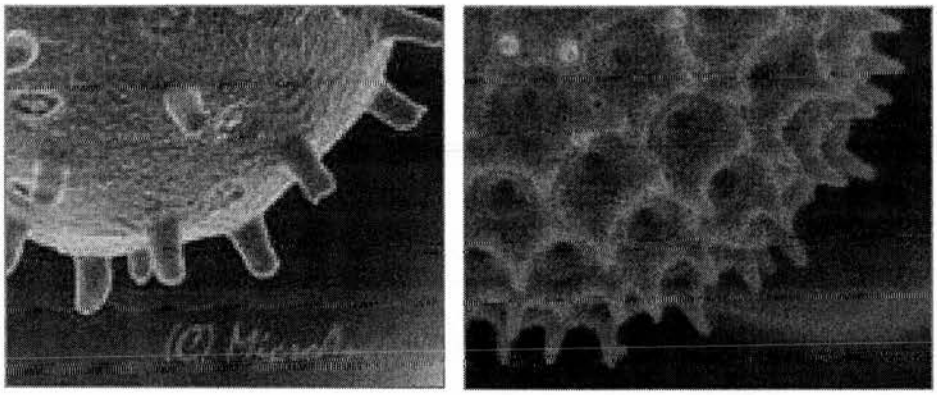


Fig. 4.8 – *Hibiscus* (left) has baculate projections and *Ipomoea* (right) has clavate projections. Images taken from www.pbrc.hawaii.edu and www.cas.muohio.edu

Guinet (1986) noted that exine structure and aperture type appear to be the most stable identification characters for *Acacia*. Arce and Banks (2001) examined *Acacia* pollen wall structure by transmission electron microscopy (TEM) of ultrathin sections, combining six wall characters with eight macromorphological characters for cladistic analysis of seven species in the *Acacia* subgenus *Aculeiferum (A.crinita, A.coulteri, A.erubescens, A.gregii, A.emilioana, A.picachensis* and *A.velutina*).

4.1.3 Previous automated pollen identification

Flenley (1968) proposed the automation of pollen identification but this has as yet been difficult to achieve. The variability in pollen size and shape was too great for the holography of Mirkin & Bagdasaryan (1972) to be successful. van Hout & Katz (2004) more recently managed to automatically measure the density and basic shape of pollen using digital holography. They measured the velocities of pollens as they settled in two different fluids and reconstructed their basic shapes from the distinctive diffraction patterns formed as they passed through a laser beam. (Fig. 4.9). Unfortunately, this system was not precise enough for the sort of pollen identification needed in pollination studies.

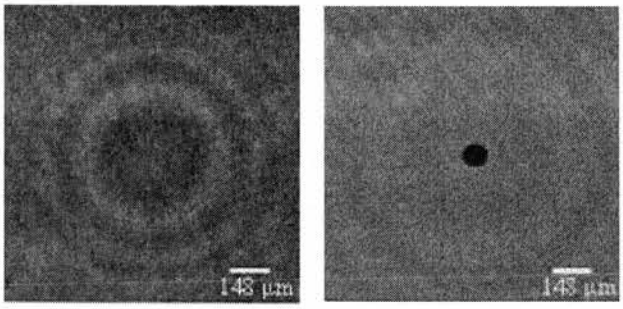


Fig. 4.9 – Hologram (left) and in-focus reconstruction (right) of a pollen grain using the digital holography system of van Hout & Katz (2004).

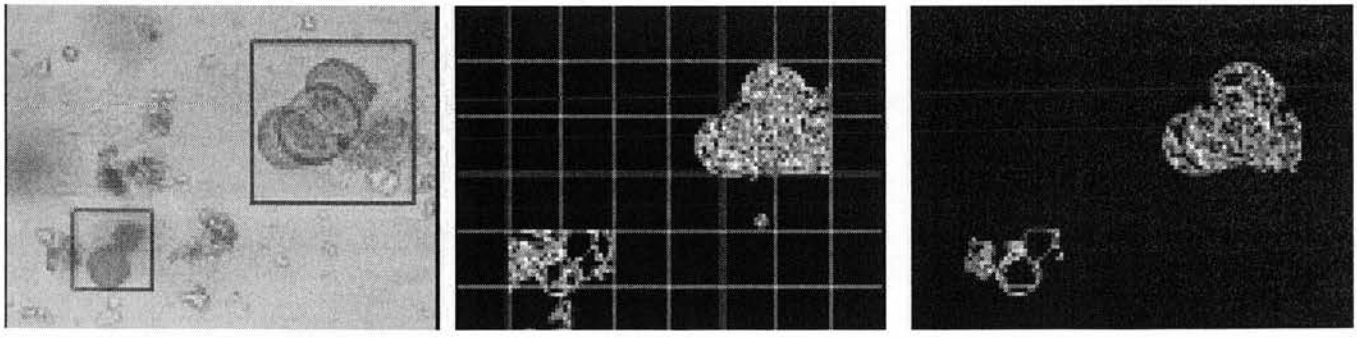
Langford *et al.* (1990) achieved much greater success using SEM images of the pollen grain surface. Over 94% of grains from six modern pollen taxa were correctly identified using just two measures of surface texture. However, a human operator was required to select suitable parts of each grain for

analysis. Treloar (1994) managed 90% success (separating six pollen taxa), but these images were from an optical microscope rather than SEM. Stillman & Flenley (1996) discussed how this could be improved. In 2004, Treloar *et al.* used geometric variables from SEM images of pollen surface structure. They identified 12 species of fresh pollen and their best set of variables had an overall classification rate averaging at 95%. De Sa-Otero *et al.* (2004) concentrated on Urticaceae pollen and achieved 86% correct to genus from optical microscope images. Li *et al.* (2004) achieved 100% success, applying a neural network classifier (see section 2.1) to surface texture data from light microscope images.

France *et al.* (2000) described a fully automated system which aimed to identify light microscope images of whole pollen grains. They applied a hierarchical approach, with each successive stage requiring higher resolution. The first two stages used variance and size to identify objects from background and pollen from debris. The latter two stages used neural networks to separate the images into classes of similar shapes then to classify to taxonomic group. They considered their results to be encouraging.

The modular system of Bonton *et al.* (2001) aimed to extract fuchsin-stained pollen grains from digital images, then identify whole grains in three dimensions (generated by combining 100 focal plane slices). For pollen grains to be extracted from an image it first had to be in focus, this was achieved using an automated focusing algorithm. A localisation algorithm, based on a split and merge scheme, then located the grain (Fig. 4.10). The localisation rate was estimated to be over 90%.

Fig. 4.10 – The pollen extraction of Bonton *et al.* (2001): RBG image (left), splitting result (middle) and merging result (right).



Bonton *et al.* (2001) identification also had two stages. First, they took global measures of the centre of the grain such as mean colour, size and concavity. These measures were matched with a database of 350 images (representing 30 pollen types) to select provisional pollen matches. The list of possible pollen types is then used to select the characteristics the system will search to prove or invalidate the

hypothesis. These domain-dependant characteristics are based on the diagnostic features used by palynologists, such as cytoplasm or pores. If only global measures were used the system was 67% accurate, if domain-dependant characteristics were also used accuracy rose to 73%.

Flenley (2003) argues that the effect of automation on palaeontology may be considerable, allowing larger counts, faster results, more objective and finer determinations.

4.1.4 Study taxa

The Laikipia Plateau has high floral diversity. Prof. Truman Young's plant list for Mpala Research Centre and Conservancy (available on the Mpala website http://www.mpala.org/researchctr/environment/pdf/mrc_vegetation_t_young.pdf) contains over 500 angiosperm species. Most of these are potential pollen or nectar sources to insects. Before working with pollen, the source flowers had first to be identified using 'The Collins Photo Guide to Flowering Plants of East Africa' (Blundell, 1992) and by drawing on the botanical experience of fellow researchers at Mpala Research Centre. Provisional identifications were confirmed by reference to the Mpala herbarium (largely the work of Prof. Truman Young). Flowering plants were digitally photographed as they were identified to produce a digital herbarium (on CD-ROM as Appendix 3a). Four example images are shown in Fig. 4.11. Fifty-six of the most common flower genera, from 29 families, were sampled (Table 4.2). Many families were represented only once; the dominant families were Acanthaceae (4), Asteraceae (9), Lamiaceae (5), Malvaceae (4) and Verbenaceae (5).

Fig. 4.11 – Mpala flowers from the digital herbarium: *Carissa edulis* (top left), *Commelina benghalensis* (top right), *Plectranthus* sp. (bottom left) and *Crossandra nilotica* (bottom right).

Genus	Family	Unclean	Clean
Abutilon	Malvaceae	y	y
Acacia	Mimosaceae	y	y
Achyranthes	Amaranthaceae		y
Aloe	Aloeaceae		y
Anthericum	Liliaceae	y	
Athroisma	Asteraceae		y
Barleria	Acanthaceae	y	y
Becium	Lamiaceae	y	y
Carissa	Apocyanaceae	y	y
Chascanum	Verbenaceae		y
Clerodendrum	Verbenaceae		y
Commelina	Commelinaceae	y	y
Commicarpus	Nyctaginaceae	y	y
Unidentified Composite	Asteraceae	y	y
Craterostigma	Scrophulariaceae	y	
Crossandra	Acanthaceae	y	y
Crotolaria	Fabaceae		y
Croton	Euphorbiaceae	y	
Cynodon	Poaceae		y
Echiochilon	Boraginaceae	y	y
Erigeron	Asteraceae		y
Erogrostis	Poaceae		y
Gloriosa	Liliaceae		y
Grewia	Tiliaceae	y	y
Gutenbergia	Asteraceae	y	y
Gynandropsis	Capparaceae	y	y
Helichrysum	Asteraceae		y
Heliotropium	Boraginaceae	y	y
Hibiscus	Malvaceae	y	y
Hypoestes	Acanthaceae		y
Indigofera	Fabaceae	y	y
Ipomoea	Convolvulaceae	y	y
Jasminum	Oleaceae		y
Justicia	Acanthaceae		y
Kalanchoe	Crassulaceae	y	y
Kleinia	Asteraceae		y
Lantana	Verbenaceae		y
Leonotis	Lamiaceae		y
Leucas	Lamiaceae	y	y
Lippia	Verbenaceae	y	y
Lycium	Solanaceae	y	
Melhania	Sterculiaceae	y	y
Ocimum	Lamiaceae	y	y
Opuntia	Cactaceae		y
Pavonia	Malvaceae	y	y
Pelargonium	Geraniaceae		y
Pentanisia	Rubiaceae	y	y
Plectranthus	Lamiaceae	y	y
Portulaca	Portulaceae	y	y
Priva	Verbenaceae	y	
Psiadia	Asteraceae		y
Sida	Malvaceae		y
Solanum	Solanaceae	y	y
Sphaeranthus	Asteraceae		y
Tagetes	Asteraceae		y
Tribulis	Zygophyllaceae		y

Table 4.2 – Pollen genera identification using unclean (section 4.2) and clean (section 4.3) pollen. [y indicates that a genus was included in the analysis].

4.2 Fuchsin gel: the simplest approach

4.2.1 Introduction

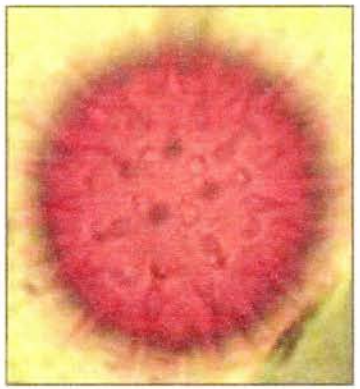
The simplest way to examine pollen in the field is to collect it onto a piece of Fuchsin gel (Beattie, 1971), which is then melted onto a microscope slide and enclosed with a coverslip. The Fuchsin powder ($C_{19}H_{17}N_3.NCl$) in the gel dyes the pollen purple, making structures easier to see. This is a cheap and easy technique, recommended for pollination ecology in Dafni (1992) and Kearnes & Inouye (1993). It was used for on-site pollen sampling in Costa Rican forest by Baldock *et al.* (2004) so was likely to be practical for on-site sampling in Kenyan savannah. Fuchsin stained pollens have already been used for automated identification by Bonton *et al.* (2001).

4.2.2 Methods

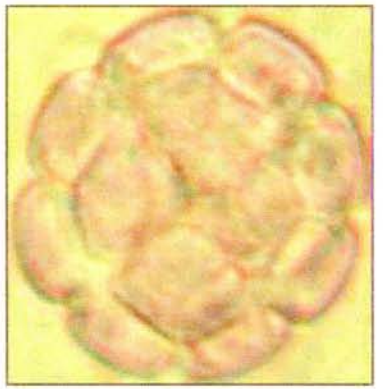
Fuchsin gel was made according to the protocols given by Dafni (1992). In the field, a small piece (2mm by 2mm) was cut from the storage tube and wiped over the anthers of the plant being sampled (having chosen a flower that was clearly dehiscent). In plants, such as *Solanum*, with poricidal anthers (enclosing the pollen until it is released by buzz pollination, as described in Shelly & Villalobos, 2000) the anthers were broken open prior to sampling. The piece of Fuchsin was then placed on a microscope slide, held over a candle flame to melt it and a cover slip placed on top. The slide was wiped clean of candle soot and labelled with permanent marker. Later in the lab, the slide was examined under a binocular optical microscope and the grains located. The right eyepiece was removed and a Nikon Coolpix 4500 digital camera mounted in its place. The microscope illumination was set as high as possible, the X40 objective lens used and images taken on macro mode without camera zoom.

Pollen images were rotated using Jasc Paint Shop Pro 8 so that the longest dimension ran vertically. No other attempts were made to standardise specimen orientation. Each grain was cropped from the larger image, brightness increased by 25 units and contrast increased by 25, and saved as a tiff file, with no compression. Up to a hundred grains of each plant species were imaged. All images are included on CD-ROM (Appendix 3b) and seven examples shown in Fig. 4.12. In the first analysis 33 genera were identified using DAISY. All available images were included, so the training sets were unbalanced with respect to numbers of images per plant species (Table 4.3). In the second analysis training sets were balanced (at TS size of 10) and only 27 genera were included (to allow comparison with acetolysed pollens in section 4.3) (Table 4.4). Pollen grain images were identified by DAISY using the jack-knife approach (see section 3.2.1.3), performing none of the DFE image modifications and setting the normalised polar thumbnail size to 32 X 32 pixels.

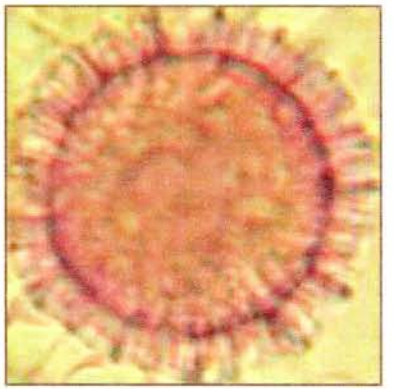
Fig. 4.12 – Seven pollen genera imaged on Fuchsin slides.



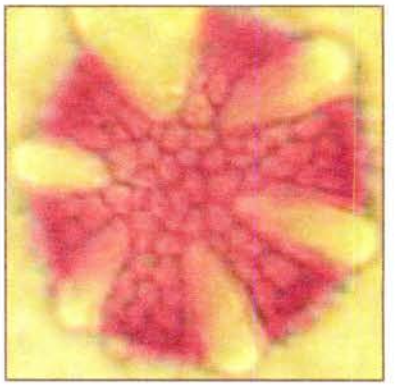
Abutilon



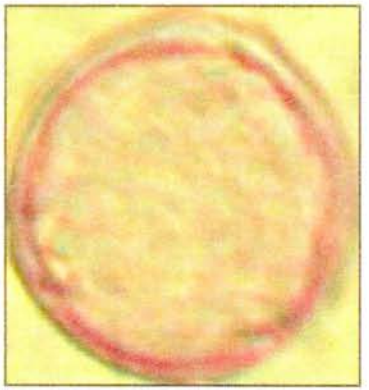
Acacia



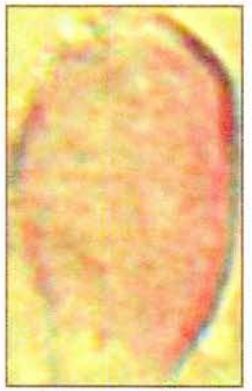
Barleria



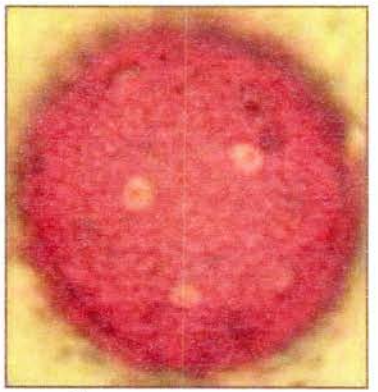
Becium



Carissa



Commelina



Commicarpus

4.2.3 Results

When all grains were included overall success was poor, with an average of just 34% of species correctly identified with first-past-the-post and a mere 10% when the greater certainty of Coord3 was demanded (Table 4.3).

Genus	Grain #	FPTP (%)	Coord3 (%)
Abutilon	17	82.35	23.53
Acacia	33	24.24	3.03
Anthericum	19	0.00	0.00
Barleria	23	13.04	0.00
Becium	23	30.43	17.39
Carissa	20	10.00	0.00
Commelina	40	32.50	15.00
Commicarpus	11	81.82	18.18
Composite	28	14.29	0.00
Craterostigma	24	45.83	12.50
Crossandra	20	95.00	90.00
Croton	32	50.00	12.50
Echiochilon	20	10.00	0.00
Grewia	24	25.00	4.17
Gutenbergia	25	12.00	0.00
Gynandropsis	33	36.36	12.12
Heliotropium	21	14.29	4.76
Hibiscus	64	42.19	4.60
Indigofera	32	18.75	3.12
Ipomoea	67	56.72	11.94
Justicia	45	64.44	46.67
Kalanchoe	25	56.00	0.00
Leucas	30	43.33	10.00
Lippia	23	13.04	0.00
Lycium	28	7.14	0.00
Melhania	8	12.50	0.00
Ocimum	28	21.43	10.71
Pavonia	12	25.00	0.00
Pentanisia	27	11.11	0.00
Plectranthus	40	37.14	5.71
Portulaca	44	63.64	15.91
Priva	26	26.92	7.69
Solanum	22	45.45	4.55
MEAN	**28**	**34.00**	**10.12**
s.d.	**13**	**23.42**	**17.22**
s.e.	**2**	**2.90**	**2.14**

Table 4.3 – Training set number, first-past-the-post performance and Coord3 performance for 33 pollen genera, imaged on Fuchsin gel slides. The best identified genera (FPTP > 80%) are highlighted pale and the worst (FPTP < 11%) darker.

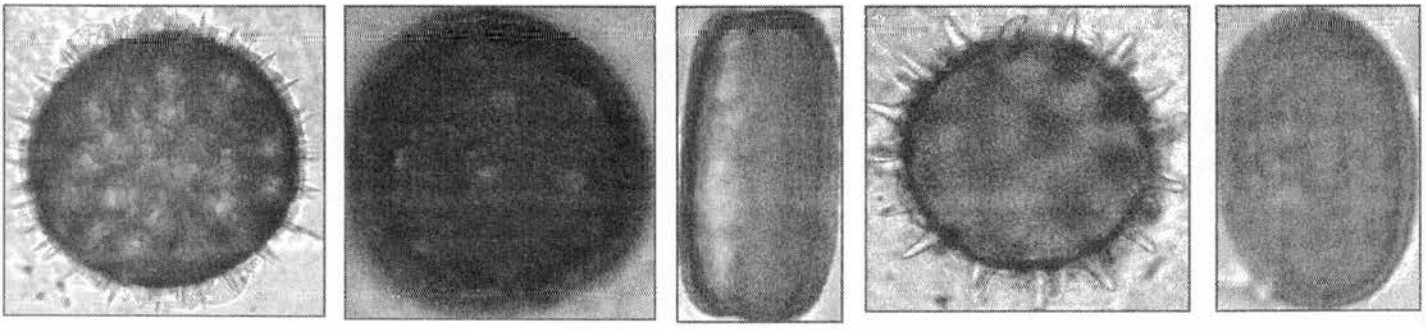
When the training sets are evenly balanced at 10 images and six genera removed the identification of grains was even less accurate. Only 33% of grains had the correct genus as their nearest neighbour (FPTP) and in only 2% of cases the three nearest neighbours agreed on the correct identification (Coord3) (Table 4.4).

Genus	FPTP (%)	Coord3 (%)
Abutilon	50	0
Acacia	20	0
Barleria	40	10
Becium	20	0
Carissa	20	0
Commelina	30	0
Commicarpus	90	30
Composite	10	0
Echiochilon	40	10
Grewia	0	0
Gutenbergia	20	0
Gynandropsis	20	10
Heliotropium	10	0
Hibiscus	40	0
Indigofera	0	0
Ipomoea	80	0
Justicia	50	30
Kalanchoe	50	0
Leucas	20	0
Lippia	0	0
Melhania	20	0
Ocimum	40	0
Pavonia	50	0
Pentanisia	30	0
Plectranthus	50	20
Portulaca	40	0
Solanum	40	10
ALL	32.59	2.22
s.d	22.12	8.92
s.e.	4.26	1.72

Table 4.4 - First-past-the-post and Coord3 identification accuracy for 27 acetolysed pollen genera. The best identified genera (FPTP > 80%) are highlighted pale and the worst (FPTP < 11%) darker.

It is interesting to see which genera were identified well and which poorly. In the unbalanced analysis, *Abutilon* (82%), *Commicarpus* (82%) and *Crossandra* (95%) were identified well (> 80%) with first-past-the-post (pale highlight in Table 4.3) (Fig. 4.13). Only *Crossandra* (90%) and *Justicia* (47%) had any real success (> 30%) with Coord3 (shown in Fig. 4.13).

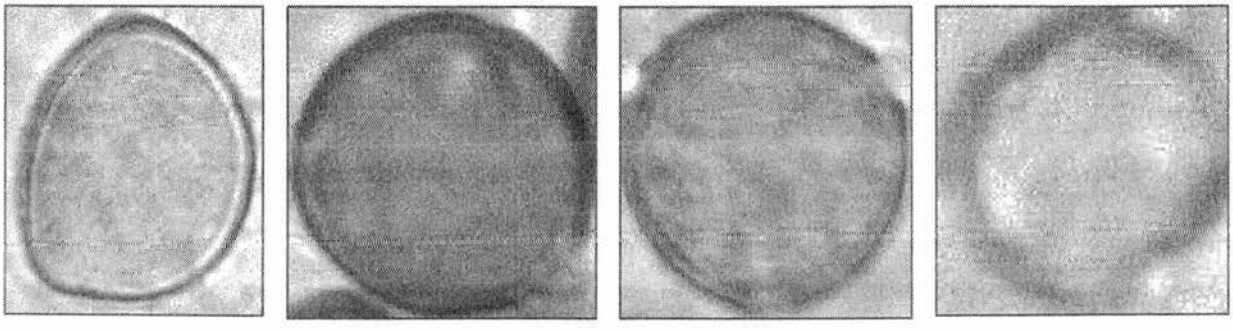
Fig. 4.13 – The better identified genera. L – R: *Abutilon, Commicarpus, Crossandra, Ipomoea* and *Justicia.*



In the balanced analysis only two genera were identified with at least 80% accuracy FPTP; *Commicarpus* re-occurs (90%) and *Ipomoea* (80%) replaces *Abutilon* (50%) (pale highlight in Table 4.4). *Crossandra* was missing from this image set. The Coord3 accuracy was low for all genera, the genera identified most accurately with Coord3 being *Commicarpus* (30%) and *Justicia* (30%) (shown in Fig. 4.13).

Some genera were identified very poorly (< 10% FPTP). In the unbalanced analysis *Anthericum* (0%), *Lycium* (7%), *Carissa* (10%) and *Echiochilon* (10%) were the worst identified with FPTP (green in Table 4.3) and had no grains correctly identified with Coord3 (along with several other genera). These genera are shown in Fig.4.14

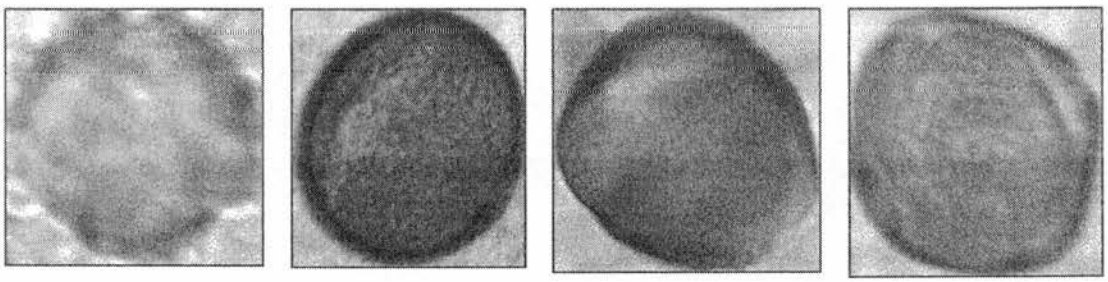
Fig. 4.14 – The least accurately identified genera in the unbalanced analysis. L – R: *Anthericum, Lycium, Carissa* and *Echiochilon.*



The genera that were the least accurately identified FPTP (< 10%) in the balanced analysis were completely different: Composite (10%), *Grewia* (0%), *Indigofera* (0%) and *Lippia* (0%) (highlighted

dark in Table 4.4). Again they had no Coord3 success. They are shown in Fig. 4.15. *Anthericum* and *Lycium* were missing from this analysis. *Carissa* (20%) and *Echiochilon* (40%) had moderate identification accuracy.

Fig. 4.15 – The least accurately identified genera in the unbalanced analysis. L – R: Composite, *Grewia, Indigofera* and *Lippia*..



The genera with fewer training images in the unbalanced analysis did marginally better than those with many. The genera identified most accurately had a mean training size of 16 and those identified the least accurately had more training, averaging at 22 images. *Ipomoea,* the genus with the largest training set in the unbalanced analysis (67), was identified with 23% more accuracy once all training sets were balanced at 10.

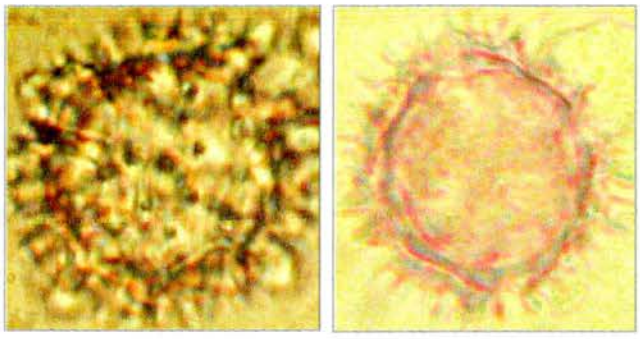
4.2.4 Discussion

An identification success rate of one in three (or one in ten with confidence) is too small for DAISY identification of Fuchsin-stained pollen grains to be a useful and practical field tool. The problems inherent in recognising three-dimensional objects were acknowledged from the earliest applications of DAISY. For example, Weeks and Gaston (1997: p. 267) wrote:

"The difficulties involved in consistently acquiring high-quality, in-focus images and objective feature measurements in poorly understood character spaces may restrict the application of these techniques."

It is difficult to be consistent in the use of fuchsin gel. Less than 0.1g of fuchsin powder is needed when fuchsin gel is made. This small amount of dye is interpreted differently by different researchers and it is difficult to make two batches of gel that have the same dye strength. Additional reference pollens were mounted in a weaker batch of fuchsin gel but they looked too different to include in the analysis (Fig. 4.16).

Fig. 4.16 – *Gutenbergia* pollen dyed with fuchsin gel from different batches, the grain on the right is from the stronger (more pink) batch used for the trial. The grain on the left has been imaged at lower illumination to make features as pronounced as possible.



Several factors may have determined which genera were identified successfully, such as distinctiveness of shape, size and surface complexity. It is unsurprising that *Crossandra* was identified well, as its perprolate form (see 4.1.2) makes it distinctive. The other three genera that were identified well FPTP (*Abutilon, Commicarpus* and *Ipomoea*) are large grained, and of complex form.

The least accurately identified genera were mainly those small grained, approximately spherical pollens with few discernable surface features. If pollen-cleaning techniques could be used that could uncover surface features on these genera, identification success might increase significantly. Such an approach is evaluated in the next section.

These data might appear to suggest that smaller training sets give more accurate identification. This goes against the findings of previous DAISY work, in which the general trend was for performance to increase asymptotically with the size of a training set (TS) (Weeks *et al.*, 1997; Weeks *et al.*, 1999b; Gauld *et al.*, 2000; Pajak, 2001; Watson *et al.*, 2004). However, these data are a weak foundation for discussion of TS size. While the genera differed in TS size (8 - 67) many other factors were also at play, confounding any analysis. The effect of training experience is investigated more thoroughly in section 4.5.

4.3 Acetolysis to clean away surface residues

4.3.1 Introduction

In section 4.2, pollen was collected directly from anthers onto Fuchsin gel. Although this is the quickest way to make slides it leaves pollen uncleaned. Bernhardt (1989) describes a caked and greasy texture of pollen loads on some *Acacia*-foraging bees, due to a mixture of pollen grains from co-flowering genera that bear copious deposits of lipid on their exines (e.g. *Hibbertia* (Dilleniaceae), Berhardt, 1986). Palynologists routinely clean all pollens using a process known as acetolysis (Faegri, 1992). Acetolysis dissolves cellulose, hemicellulose and chitin, the main components of the unwanted organic residue that collects on the surfaces of pollen grains and obscures surface features. The acid solution used reacts violently with water so glacial acetic acid washes are necessary to replace water with acetic acid. The acetolysis methodology is given in Appendix 2.

There are both advantages and disadvantages to acetolysis in the context of fieldwork.
Advantages include:

- Clarity of structure - surface features can be seen more clearly.
- Darkening - acetolysis darkens grain tissues so it is not necessary to stain acetolysed pollen. The extent of darkening does not differ significantly between batches (a weakness of fuchsin gel).
- Dispersal - clumps of pollen are broken up so the grains spread evenly on a slide.
- Comparability - palynologists making slide collections of pollen from large herbaria (such as those at the National Museums of Kenya and the Natural History Museum) acetolyse their pollen samples. Thus for such reference collections to be easily comparable, ecological studies should adopt this procedure.

Disadvantages include:

- An extra level of processing to pollen identification that requires:
    - Basic laboratory skills;
    - Substantial time investment (one day for 20 samples);
    - Access to a fully equipped laboratory;
    - Replaceable supplies of consumables, such as glacial acetic acid, acetic anhydride and sulphuric acid.
- A single sample would still take half a day so it is more efficient to process samples in batches.
- Acetolysed pollens would be incomparable with field specimens taken using the simpler Fuchsin approach

4.3.2 Methods

Pollen was collected from donor plants into 0.5 ml plastic centrifuge tubes of 70% ethanol. The method used depended on the size and structure of the flower. Large anthers were stroked with a mounted needle to collect the polled dislodged. Small anthers were cut from the plant with fine scissors and placed into the tube. Poricidal anthers (such as those of *Solanum*) were broken open then placed in the tube. The tube lids were coded with permanent marker, as this was liable to be washed off with ethanol-wet fingers they were also placed in a box with storage spaces arranged in a labelled grid. Both tube codes and box grid spaces were listed (in a field notebook) against the corresponding plant name to ensure that pollen identity remained clear.

Of these samples, 52 pollen genera were acetolysed and slide-mounted (both methods as Appendix 2) in the pollen laboratory at the Natural History Museum, London. Each slide was examined under an optical microscope, and grains were imaged and images modified using the same methods described above for the Fuchsin pollens (section 4.2). The acetolysed pollen images are included on CD-ROM as Appendix 3b. Twenty-seven genera were identified using DAISY, each genus being represented by ten grains (chosen at random) from a single species, to produce balanced training comparable to the balanced Fuchsin results. These were identified using the same protocols as before.

4.3.3 Results

When 27 genera were identified with training sets of ten grains, either unclean or clean, the clean grains were on average better identified. If grains were cleaned an extra 11% had their nearest neighbour in the correct genus (FPTP values). If the three nearest neighbours had to agree on the genus (Coord3 values) the improvement with cleaning was more marked at 16%. However, not all pollens were identified more accurately with cleaning (the more accurate method for each genus is highlighted yellow on Table 4.5). Sixteen genera were identified more accurately by FPTP if pollen was cleaned, only six were identified less accurately. With Coord3, 14 genera were identified more accurately with cleaning and four less accurately.

Table 4.5 – Twenty-seven pollen genera identified with or without cleaning. The method with greater accuracy for each genus is highlighted.

Genus	FPTP (%)		Coord3 (%)	
	unclean	clean	unclean	clean
Abutilon	50	80	0	40
Acacia	20	50	0	30
Barleria	40	40	10	0
Becium	20	30	0	20
Carissa	20	50	0	10
Commelina	30	0	0	0
Commicarpus	90	100	30	90
Composite	10	40	0	20
Echiochilon	40	60	10	30
Grewia	0	80	0	60
Gutenbergia	20	100	0	0
Gynandropsis	20	0	10	0
Heliotropium	10	80	0	50
Hibiscus	40	90	0	90
Indigofera	0	10	0	0
Ipomoea	80	20	0	0
Justicia	50	100	30	70
Kalanchoe	50	40	0	0
Leucas	20	0	0	0
Lippia	0	20	0	0
Melhania	20	50	0	30
Ocimum	40	40	0	10
Pavonia	50	70	0	10
Pentanisia	30	30	0	0
Plectranthus	50	40	20	10
Portulaca	40	40	0	0
Solanum	40	40	10	0
ALL	**32.59**	**44.81**	**2.22**	**18.15**
standard dev.	22.12	30.76	8.92	28.19
standard error	4.26	5.92	1.72	5.43

The mean results are shown graphically in Fig. 4.17. Clean pollen (C) was distinctly more accurately identified overall than unclean pollen (UC), with no overlap in standard error. This difference was statistically significant (paired t-test, P = 0.02) for FPTP and highly significant (paired t-test, P = 0.002) for Coord3.

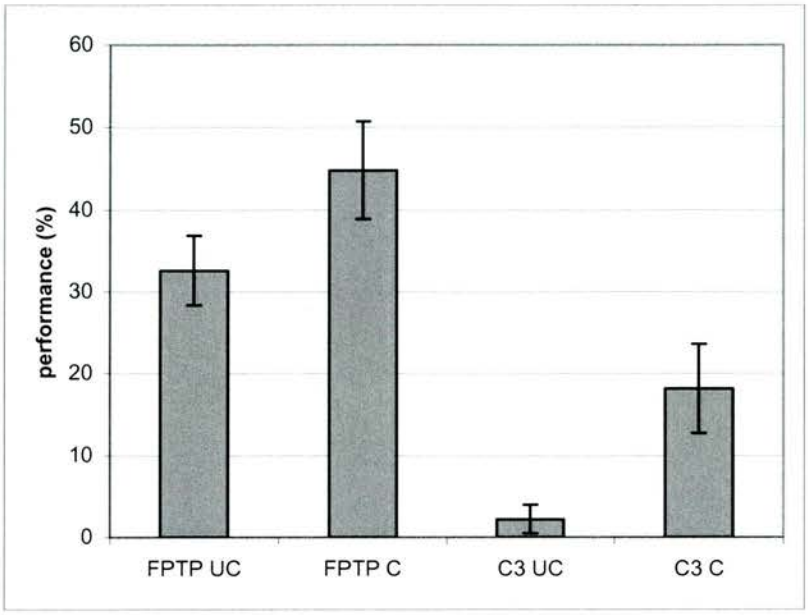


Fig. 4.17 – Clean pollen grains (C) were more accurately identified than unclean grains (UC) both in the first-past-the-post results (FPTP) and when the greater confidence of Coord3 was demanded (C3). Error bars are the standard error of the mean in each case.

The pollen genera that were identified with at least 70% greater accuracy when cleaned were *Grewia* (+80%), *Gutenbergia* (+80%) and *Heliotropium* (+70%) (Fig 4.18).

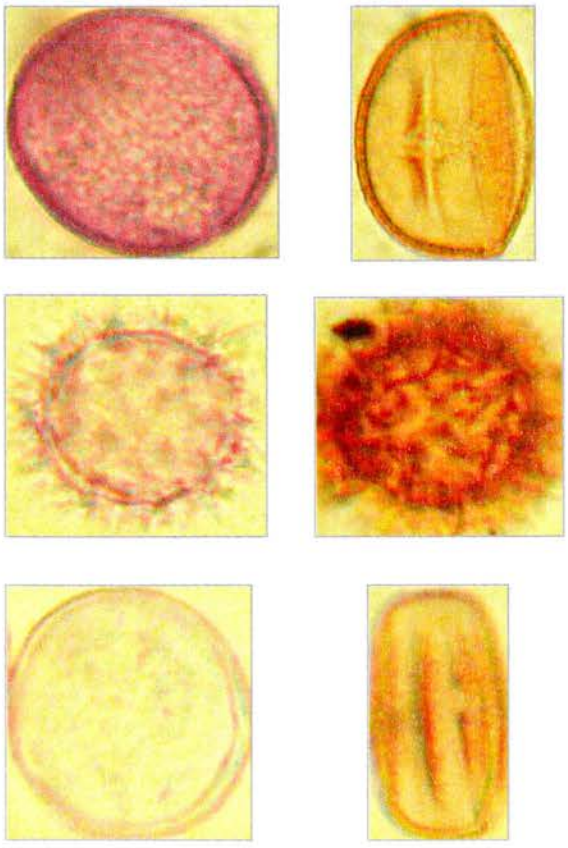


Fig. 4.18 – The pollen genera in which accuracy increased the most with pollen cleaning. Without cleaning (left) and with cleaning (right): *Grewia* (top), *Gutenbergia* (middle) and *Heliotropium* (bottom).

The grains of *Grewia* and *Heliotropium* look very different before and after cleaning, more so than acetolysis could really account for. It was only possible to make confident identifications at genus-level so these pollens may have been sampled from different species with very different pollen shape. Thus, only *Gutenbergia* can be considered a direct comparison. In this genus the darkening caused by acetolysis made structures more obvious than did the dyeing effect of Fuchsin gel.

The pollen genera that were identified with at least 20% poorer accuracy when pollens were acetolysed were *Ipomoea* (-60%), *Commelina* (-30%) and *Gynandropsis* (-20%). These are shown in Fig. 4.19.

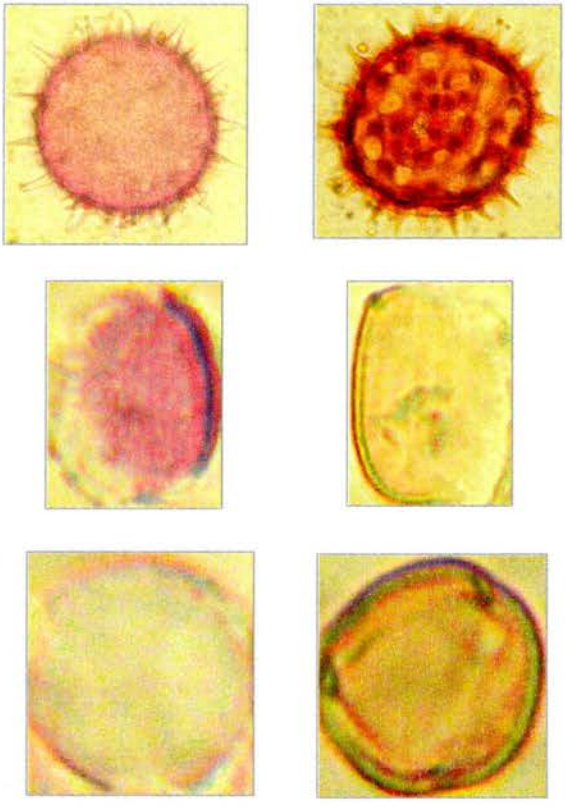


Fig. 4.19 – The pollen genera in which accuracy decreased the most with pollen cleaning. Without cleaning (left) and with cleaning (right): *Ipomoea* (top), *Commelina* (middle) and *Gynandropsis* (bottom).

These three genera share the same basic structure in unclean and clean images so are likely to have come from the same species. Clean pollen seems to show pollen shape more distinctly than unclean pollen in *Ipomoea* and *Gynandropsis* but cleaning makes little difference in *Commelina.*

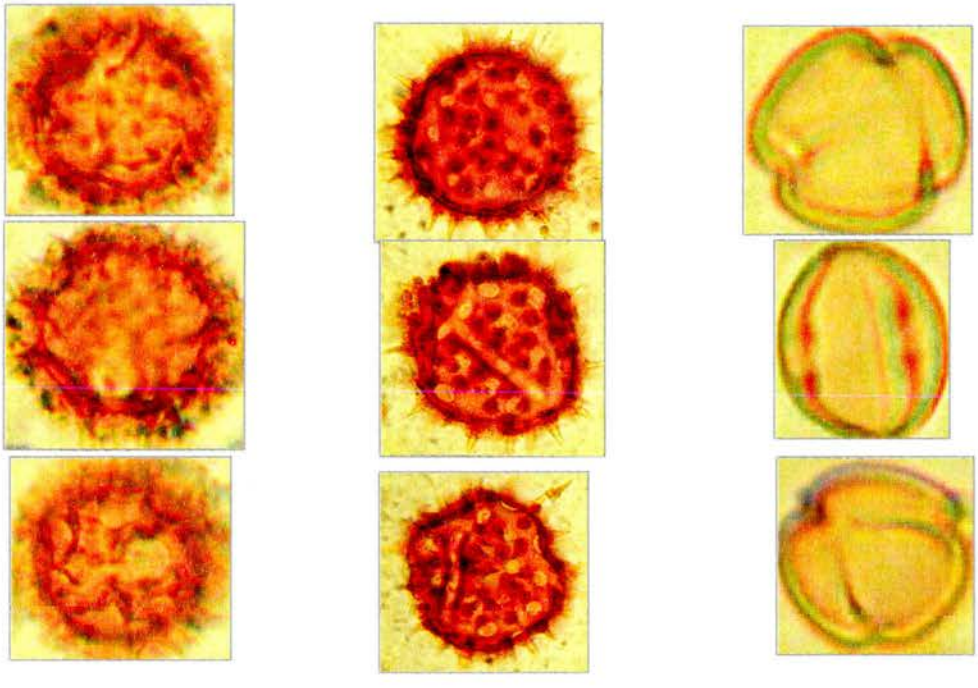
4.3.4 Discussion

Acetolysed pollen was significantly more accurately identified than uncleaned pollen. However, the 11% increase in FPTP accuracy still only produces an overall average value of 45% correct identification. Thus, the majority of specimens are still being wrongly identified. Acetolysis alone is not sufficient to make DAISY a useful pollen identification tool but it could contribute to a successful methodology.

Although 16 genera were identified more accurately with pollen cleaning six were identified less accurately. Pollen cleaning generally made surface detail more pronounced, increasing differences; this

can be an advantage or a disadvantage. This change could have increased intergeneric variability but it could also have emphasised intrageneric variability caused by orientation differences. *Gutenbergia* (+80% FPTP when cleaned) has very fine-scaled surface detail that looks similar from all angles, so intrageneric variability was unlikely to be a problem for this genus. *Ipomoea* (with its pale spots, -60%) and *Gynandropsis* (-20%), however, have shapes that will exhibit much more intrageneric variability with orientation (Fig. 4.20).

Fig. 4.20 – *Gutenbergia* (left) was identified 80% more accurately FPTP with acetolysis, having little intrageneric variation in surface features. *Ipomoea* (middle) and *Gynandropsis* (right) were identified less accurately after acetolysis, they will have greater intraspecific variation with orientation.



The time demands for acetolysis are substantial; it takes most of a day for each round of acetolysis and slide mounting. It is most time-efficient to acetolyse 10 samples concurrently. This could be a major limitation of this method as time in the field can be a precious commodity and a substantial delay may be unavoidable before DAISY identifications are available.

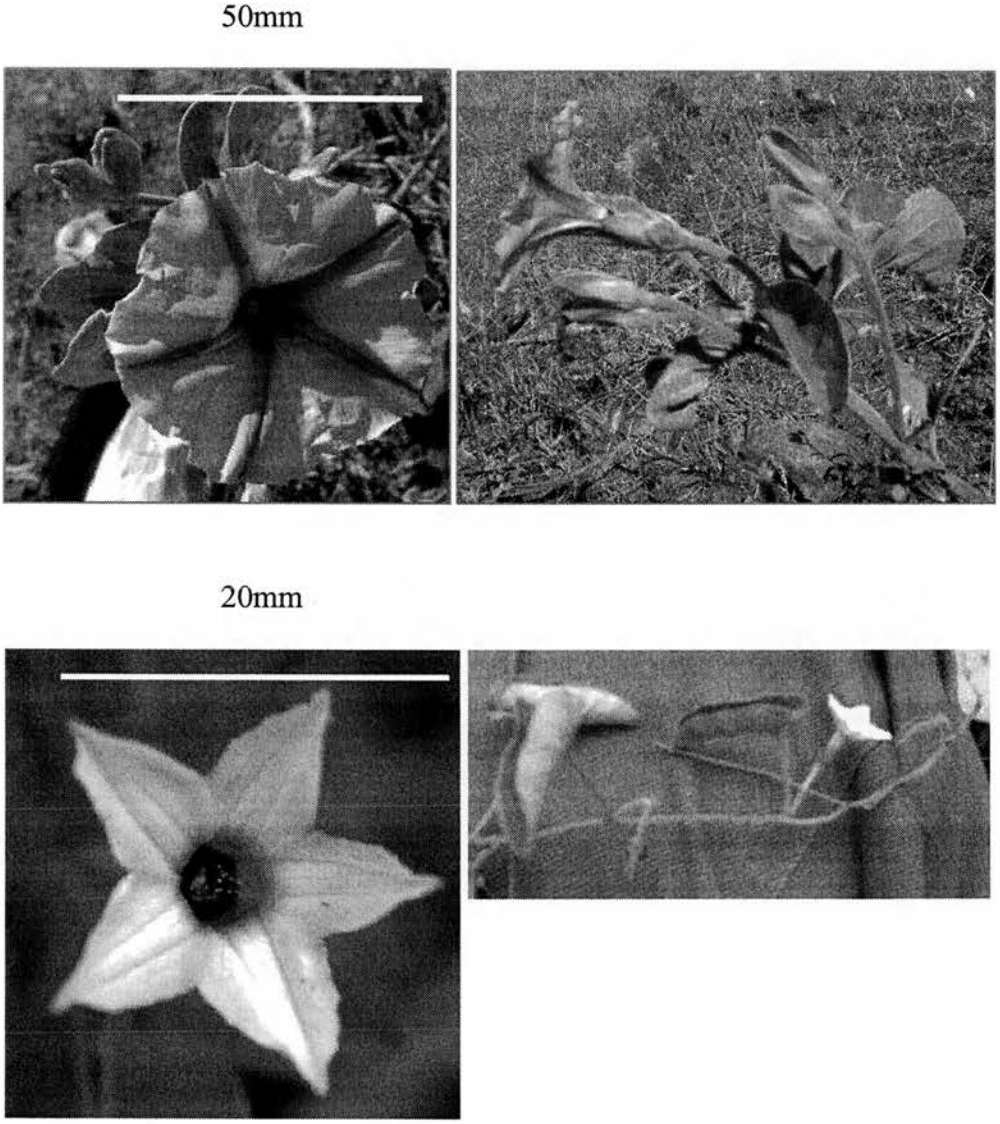
The acetolysis for this study was done in a laboratory designed for work with pollen and diatoms (the pollens lab of The Natural History Museum, London). However, most of the equipment used (e.g. centrifuge and fume cupboard) would be available in any laboratory. It may be possible to do such work at field site research stations. In the case of Mpala Research Centre (MRC) the chemicals are easily available in Nairobi. The MRC laboratory has all the essential equipment but it is already used by other projects so the monopolisation necessary for acetolysis would be difficult. Local field assistants in less developed countries are unlikely to have the required laboratory skills. Field assistants

could be trained in these simple procedures but would need extensive supervision at first to ensure safe practice in the handling of acid and in vapour control.

If different researchers present pollen in Fuchsin gel or acetolysed, then the acetolysed pollens are more likely to be comparable between projects (the problem of Fuchsin dye strength has already been discussed in section 4.2.4). If a future study were to use acetolysis it may be possible to do DAISY training using pollen slides already present in herbarium collections (such as those at NMK and NHM). Floral inventories (one of the most common surveys at every field station, e.g. Truman Young's plant list for MRC) could advise the taxa for inclusion. Herbarium training has many potential advantages:

- Large training sets could be constructed quickly.
- Training could include plant species that flower at times of the year when no pollination researcher would have sampled them.
- This work could be done in the UK producing a fully trained identification tool before the first field trip.
- Museum reference pollens would have been identified to species (or even sub-species) level by professional palynologists. This greater taxonomic distinction could allow pollination to be discussed at species-level rather than genus-level, adding a great deal to ecological discussion as different species within a plant genera can differ greatly (Fig. 4.21).

Fig. 4.21 – *Hibiscus cicatricosa* (top) and the species referred to as 'little *Hibiscus*' (bottom) are very different flowers within a single genus. *Hibiscus cicatricosa* has large, pink flowers on upright stems. Little *Hibiscus* is a climber with much smaller flowers. They may well differ substantially in their pollination but they are being treated as a single unit.



4.4 Dark-field microscopy

4.4.1 Introduction

Pollen cleaning by acetolysis made surface features more visible, improving overall pollen identification. The images discussed in sections 4.2 and 4.3 were taken with light-field (LF) microscopy but dark-field (DF) microscopy may make features even more pronounced. Thomas *et al.* (2004) found that reproductive details were clearly visible on a range of specimens using incident-light DF microscopy and suggested that it had enormous potential for palaeobotany. Micscape, the monthly magazine of the Microscopy-UK website (www.microscopy-uk.org.uk/mag.html) has several articles relating to DF microscopy, among which Walker (1999) used DF images in his discussion of pollen microscopy, and Overney (2004) argued that DF is a powerful method to study the fine structure of diatoms.

Some optical microscopes can be equipped with DF capability (e.g. the Leitz Dialux 20 used for this imaging), making DF microscopy as easy as LF. If this fitting is not available then DF can be implemented inexpensively by putting an appropriately sized opaque disk into the light path located close to the condenser's diaphragm. This approach is advocated on numerous websites, including Overney (2004).

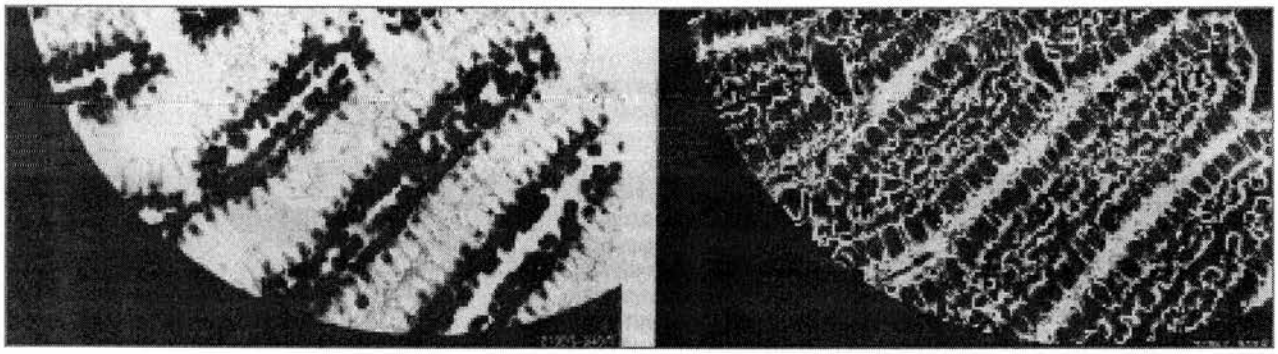
In LF microscopy the specimen is viewed against a bright background and its contrast is largely the result of absorption by the stains in the specimen or by the scattering power of the specimen. If the background is too bright there may be little contrast between the background and the specimen so that few features will be observable (Lacey, 1999) (see left image of Fig 4.22).

DF microscopy uses the light scattered from obliquely illuminated objects, which then behave as self-luminous objects (Weiss *et al.,* 1999) (see right image of Fig. 4.22 and Fig. 4.23). DF condensers deflect the light from the illuminator so that it intercepts the object plane as a shallow, hollow cone, but does not enter the objective. The direct light is excluded from image formation. If no object is in the object path, the light bypasses the objective and the field of view is dark. If an object is placed in the optical path it will deflect part of the light from the illuminator and some of this deflected light enters the objective. This is illustrated in Fig. 4.24.

DF is best for objects whose structures are made visible by changes of the refractive index. For example, at an edge the refractive index may change abruptly and light is scattered. Part of this

scattered light will reach the objective. A brightly illuminated edge is thus seen against a dark background. In LF this edge would hardly be visible (Determann & Lepusch, 1979) (Fig. 4.22).

Fig. 4.22 – Example object imaged with light-field (left) and dark-field (right). The edges of the object light up brilliantly in the DF. Taken from Determann & Lepusch, 1979.



In this set of images the distinction is clearly seen when *Commelina* pollen is imaged (Fig. 4.23).

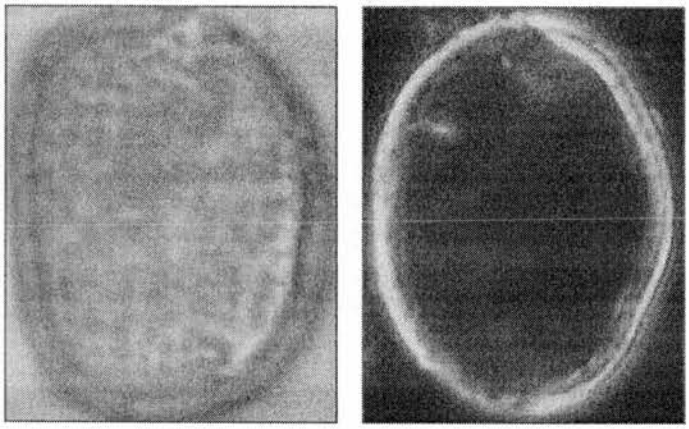
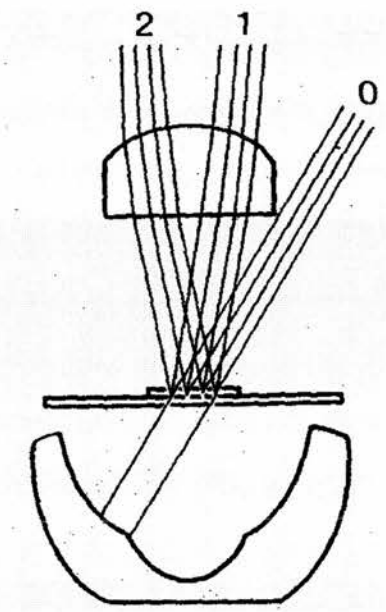


Fig. 4.23 – *Commelina* pollen imaged with bright-field (left) and dark-field (right). The edge is very clear in the dark-field image.



Fig, 4.24 – Theoretical diagram of DF microscopy, taken from Determann & Lepusch, 1979. The direct (0^{th} order) light no longer reaches the objective. Only the diffracted (1^{st} and 2^{nd} order) light is used for image formation.

There are both advantages and disadvantages of DF microscopy (compared to LF)

Advantages:

- Increased contrast for features that appear pale with LF.
- May make images more colourful, the colours being a diffraction phenomenon (Sterrenburg, 1997) (Fig. 4.25).
- Dark background means it is 'eye-friendly' for long working sessions (Sterrenburg, 1997).

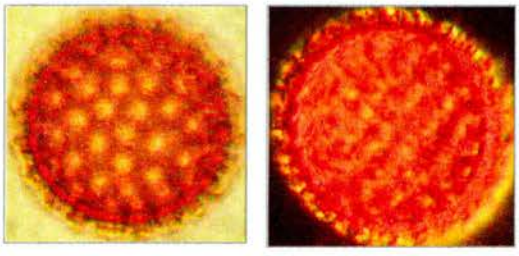


Fig. 4.25 – *Tribulis* pollen imaged with LF (left) and DF (right). The DF image has brighter colours.

Disadvantages:

- Stained objects are usually unsuitable (Determann & Lepusch, 1979).
- The specimen must not be too thick or dense, otherwise the contrast in dark-field will be reduced and the object lightened excessively (Determann & Lepusch, 1979).
- Microscope slides, coverslip, condenser and the front lens of the objective must be free from scratches and absolutely clean as any impurity would lighten the background (Determann & Lepusch, 1979).
- It is not possible to obtain a faithful representation of the whole object, for which direct light is essential (Determann & Lepusch, 1979).

4.4.2 Methods

Fifteen pollen genera were selected at random from the pool of cleaned pollen images. One hundred grains of each genus were imaged by both LF and DF microscopy and prepared for identification using the same protocols as described previously, then identified using NPTs of size 32.

4.4.3 Results

The DF images were identified more accurately overall than the LF ones, with 13% more correctly identified with FPTP and 17% more when Coord3 demanded greater certainly (Table 4.6 and Fig. 4.26). Pollen genera vary greatly in structure so some will be better suited to DF microscopy than others. The more accurate method for each genus is highlighted yellow on Table 4.6. Nine genera (60%) were identified more accurately with FPTP if pollen were cleaned, only four (27%) were identified less accurately. Eleven genera (73%) were identified more accurately with cleaning and three (20%) less accurately.

Table 4.6 – Identification success of 15 pollen genera using LF and DF microscopy. The better identification success (LF or DF) is highlighted.

Genus	FPTP (%)		Coord3 (%)	
	LF	DF	LF	DF
Abutilon	99	99	95	98
Acacia	61	43	34	12
Achyranthes	33	64	6	31
Barleria	100	100	94	94
Commicarpus	100	94	89	91
Erigeron	27	43	1	2
Erogrostis	26	35	1	3
Grewia	75	70	51	40
Gutenbergia	19	76	0	63
Hypoestes	100	95	98	90
Jasminum	78	92	46	84
Lantana	56	91	20	80
Sida	40	97	15	92
Sphaeranthus	86	100	74	100
Tribulis	99	100	94	98
Mean	66.62	79.93	47.9	65.2
s.e.	8.01	6.04	10.2	9.58

The mean results are shown graphically in Fig. 4.26. Dark-field (DF) pollen had higher mean accuracy than light-field pollen (LF) but there was some overlap in standard error. This difference was statistically significant for FPTP (paired t-test, P = 0.039) and Coord3 (paired t-test, P = 0.041).

Fig. 4.26 – Mean FPTP and Coord3 (C3) identification accuracy with LF and DF; the error bars show the standard error of each mean

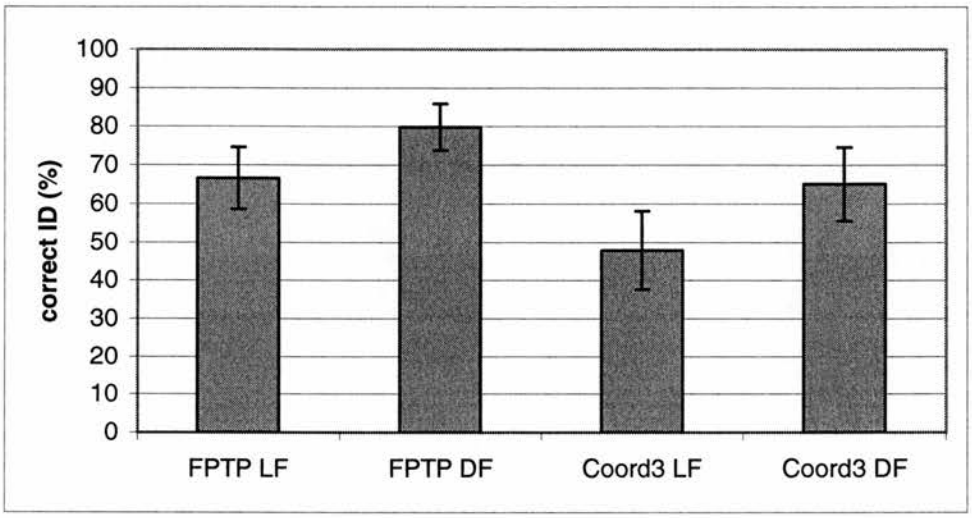


Table 4.7 evaluates LF / DF accuracy in greater detail, giving the number of objects with correct and incorrect positive identifications with FPTP and Coord3 and the identification certainties. The FPTP data give strong support to DF imaging. DF had more objects correctly identified with FPTP (9 of 15 genera, all +200) and fewer incorrectly identified with FPTP (5 of 15 genera, all –200) than LF. DF also had greater certainty of FPTP identifications (all +13%) than LF. Similar patterns emerge from the Coord3 data. DF had more objects correctly identified with Coord3 (11 of 15 genera, all +259) and greater certainty (all +6%). DF increased the number of incorrect identifications in more genera (6) than it decreased them (5) but the overall values still suggest that DF decreased the number of incorrect identifications (all – 40).

Table 4.7 – The number of objects correctly and incorrectly identified and the certainty of identification with light-field and dark-field imaging. The highest value for each LF / DF pairing is highlighted.

Genus	FPTP						Coord3					
	Correct +ve ID		Wrong +ve ID		Certainty (%)		Correct +ve ID		Wrong +ve ID		Certainty (%)	
	LF	DF	LF	DF	LF	DF	LF	DF	LF	DF	LF	DF
Abutilon	99	98	1	1	99	99	96	97	0	1	100	99
Acacia	61	43	39	57	61	43	34	12	13	13	72	48
Achyranthes	33	64	67	36	33	64	6	31	9	11	40	74
Barleria	100	100	0	0	40	97	94	94	0	0	65	100
Commicarpus	100	94	0	6	100	100	89	91	0	2	100	100
Erigeron	27	43	73	57	100	94	1	2	14	16	0	98
Erogrostis	26	35	74	65	27	43	1	3	9	11	7	11
Grewia	75	70	25	30	26	35	51	40	4	8	10	21
Gutenbergia	19	76	81	24	75	70	0	63	31	7	93	83
Hypoestes	100	95	0	5	19	76	98	90	0	0	0	90
Jasminum	78	92	22	8	100	95	46	84	6	0	100	100
Lantana	56	91	44	9	78	92	20	80	13	3	88	100
Sida	40	98	60	3	56	91	15	93	8	0	61	96
Sphaeranthus	86	100	14	0	86	100	74	100	6	0	93	100
Tribulis	99	100	1	0	99	100	94	98	0	0	100	100
ALL	**999**	**1199**	**501**	**301**	**67**	**80**	**719**	**978**	**109**	**69**	**87**	**93**

The genera that were identified substantially more accurately using DF microscopy were: *Achyranthes* (+ 31% FPTP), *Gutenbergia* (+ 57%), *Lantana* (+ 35%) and *Sida* (+ 57%) (Table 4.7, Fig. 4.27). Changes to the number of genera identified correctly and incorrectly and certainty for these four genera are shown in Table 4.8. *Lantana* and *Sida* had many more genera correctly identified with DF, fewer incorrect identifications and greater certainty with both FPTP and Coord3. *Achyranthes* had slightly more incorrect identifications with Coord3, and *Gutenbergia* had less certainty with DF.

Table 4.8 – The four pollens that were identified substantially more accurately when DF was used in place of LF. Changes in the number of genera correctly and incorrectly identified and the identification certainty.

Genus	FPTP			Coord3		
	Correct	Incorrect	Certainty (%)	Correct	Incorrect	Certainty (%)
Achyranthes	31	-31	31	25	2	34
Gutenbergia	57	-57	-5	63	-24	-10
Lantana	35	-35	14	60	-10	12
Sida	58	-58	35	78	-8	35

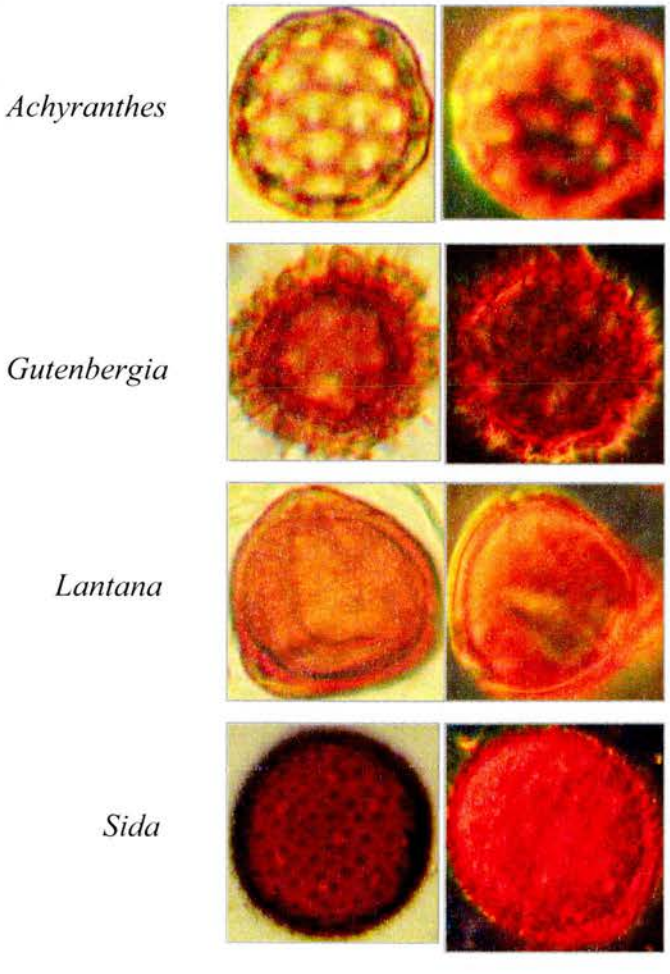


Fig. 4.27 – Four pollen genera which were identified much more accurately using DF. LF grains are on the left and DF on the right.

In the genera that were identified less accurately using DF microscopy (Fig. 4.28) the FPTP difference in performance was less pronounced. *Acacia* was the only genus to be unquestionably less distinctive with DF (Fig. 4.28); it was identified 18% less accurately with FPTP (Table 4.9) and had 18% less FPTP certainty (Table 4.9). This was also seen in the Coord3 results, where 22 fewer genera were correctly identified with 24% less certainty and there was no difference in the number of incorrect positive identifications (Table 4.9).

Table 4.9 – Four pollens that were identified less accurately when DF was used in place of LF. Change in the number of genera correctly and incorrectly identified and certainty.

Genus	FPTP			Coord3		
	Correct	Incorrect	Certainty (%)	Correct	Incorrect	Certainty (%)
Acacia	-18	18	-18	-22	0	-24
Commicarpus	-6	6	0	2	2	0
Grewia	-5	5	9	-11	4	11
Hypoestes	-5	5	57	-8	0	90

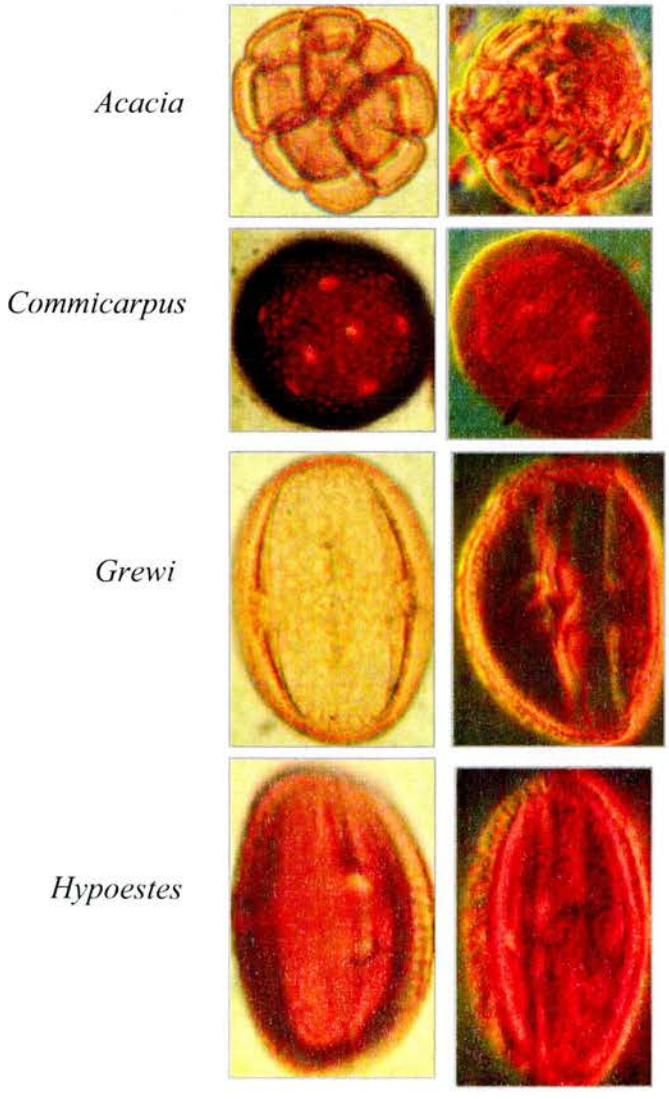
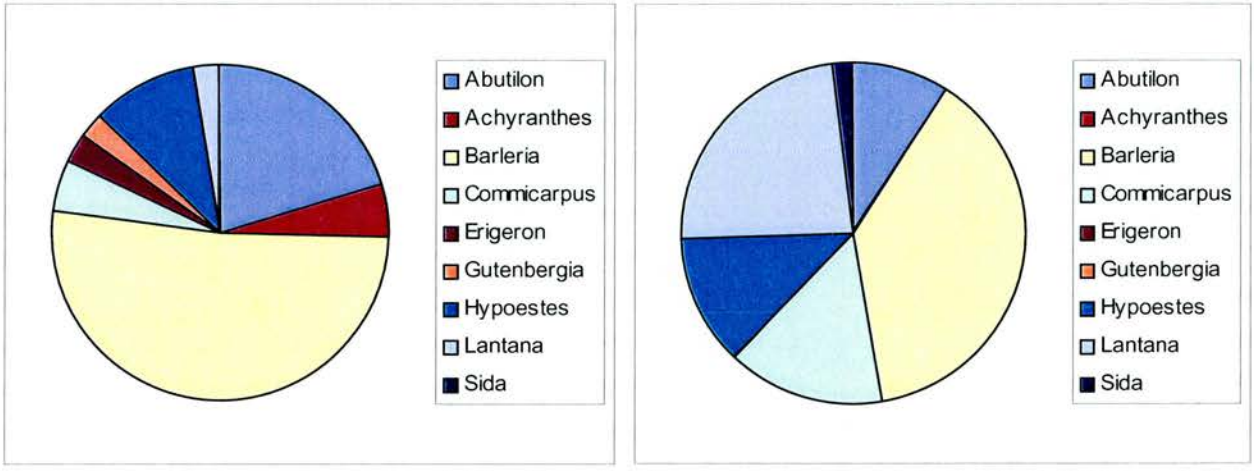


Fig. 4.28 – Four pollen genera which were identified less accurately using DF. LF grains are on the left and LF on the right.

The first-past-the-post misidentifications (i.e. the genus that was the nearest neighbour when this neighbour was of the wrong genus) made with LF and DF differed. Fig. 4.29 illustrates these misidentifications for *Acacia*.

Fig. 4.29 – LF (left) and DF (right) FPTP mis-identifications for *Acacia*.



Barleria was the most common misidentification for *Acacia* both with light-field (51%) and dark-field (37%) but the relative importance of other genera differed. *Abutilon* (21%) and *Hypoestes* (10%) made the next largest contributions in LF, whereas it was *Lantana* (23%), *Commicarpus* (14%) and *Hypoestes* (12%) in DF.

4.4.4 Discussion

Pollen grains imaged with dark-field microscopy are better identified overall than pollen grains imaged with light-field microscopy. However, as with acetolysis, this method brings limitations and does not increase the percentage of correct identifications dramatically (+13% FPTP and +17% Coord3) (Fig. 4.26). The mean values are much higher than in the fuchsin / acetolysis comparison (section 4.3) but this does not mean that DF is making a big difference. This improvement in accuracy must be largely due to the fewer genera considered (15 as opposed to 27) and the larger training sets (100 as opposed to 10); the LF images have been imaged in the same way as the acetolysed images of the fuchsin / acetolysis comparison yet they are twice as well identified in this trial (67% FPTP compared with 33%).

DF seems to make edge features stand out bright orange but hides surface detail, making most of the grain black. In *Achyranthes,* where surface details are strong but will vary greatly intraspecifically (mainly due to differences in orientation) this loss of surface detail may have been advantageous. The genera that were identified better with DF tended to be those that were indistinctive in their LF colouration, so little information was lost when DF made them appear largely black. Those which were

identified more accurately with LF had more distinctive LF colouration, e.g. very dark in *Commicarpus* and pale in *Grewia.*

It is unfortunate in the context of the present project that *Acacia* is the genus that responds most poorly to DF imaging, as distinguishing *Acacia* from non-*Acacia* pollen is the most basic requirement of *Acacia* pollination studies. When *Acacia* was imaged with DF the grains making up the polyad could no longer be discriminated, so the distinctive shape of *Acacia* polyads was lost and the DF image looked very different to the LF one (Fig. 4.28). This is reflected in the different assortment of failed matches (Fig. 4.29). It is common in automated identification for different approaches to produce different error patterns. Gaston & O'Neill (2004) suggested that they could form "voting ensembles" in which the identity of a specimen is deemed to be the taxon that receives the most votes. An ensemble of LF and DF may provide more reliable identification than either approach alone but this must be balanced against the increase in user input required to achieve a DAISY identification.

In theory the background of DF images is black but in reality small marks and particles on the slide lead to a lightening of this background. This background colour may be quite consistent for grains imaged from a single slide, and may have biased DAISY in favour of identification success. When the same genus is processed from a different sample the background may be very different and this may then lead to a poorer performance. This problem could be addressed by ensuring that DAISY training pollen comes from more than one processing batch (so a range of backgrounds are represented), or by selecting the grain with a more precise region of interest, for example, a polyROI or a circle as opposed to a box-crop, so most of the background is excluded.

These analyses were undertaken using the acetolysed pollens, as stained objects are usually unsuitable for dark-field examination (Determann & Lepusch, 1979). Thus, one would have to decide at the outset of a project whether dark-field microscopy was to be used and either mount uncleaned pollens without dye or use acetolysis. Fuchsin gel slides prepared for other purposes or projects would not be of use if DF were accepted as the standard for DAISY identification.

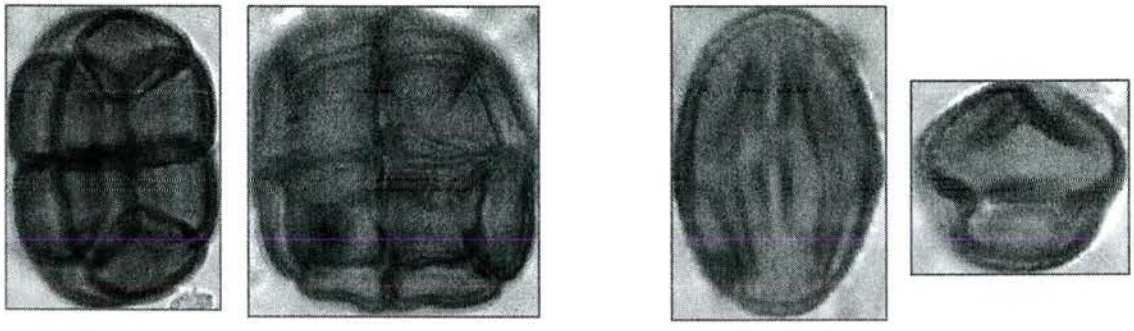
DF microscopy is not used widely in research biology. When a search was made for DF-equipped microscopes in the Biology Department of St Andrews University and at Mpala Research Centre neither had one available. Other DAISY users would likely face the same lack of equipment. Although a makeshift method has been advocated to make a LF microscope produce DF images (Overney, 2004), DF equipment achieved by such means is likely to vary substantially in the quality of the images produced. This could mean that training images taken with one microscope would look very different from test images from another microscope, leading to misidentifications.

4.5 Training set (TS) size

4.5.1 Introduction

A large training set gives DAISY experience of a wider range of intraspecific variation but takes more effort to produce. Training set size has already been discussed in relation to wing identification in section 3.4; the same considerations apply when DAISY is trained to recognise pollen. Pollen grains are varied in colour and tone, three-dimensional and often complex in structure, so images representing a single species will vary greatly (Fig. 4.30), much more so than was the case for the wing images (in section 3.4).

Fig. 4.30 – *Acacia* (left pair) and *Crotolaria* (right pair) pollen viewed from two different angles.



This high level of intraspecific variation suggests that a much larger TS size will be necessary before variation is represented adequately and the increase in performance levels off.

4.5.2 Methods

The same 15 genera and 100 image sets used in the LF/DF comparison were used here. NPT size was set at 20 throughout. For both LF and DF, the TS100 set was jack-knifed (see section 3.2.1.3). The TS was then reduced in steps to TS50, TS20 and TS10, jack-knifing at each TS size. The images removed at each step were those with the highest file numbers; as pollen had been imaged in a random order this removed a random subset.

4.5.3 Results

Although light-field images have huge differences in identification success from genus to genus (leading to large standard errors), most genera are correctly identified more frequently with each increase in TS size (Table 4.10). Eleven of the 15 genera had their greatest FPTP accuracy (highlighted on Table 4.10) with the largest TS size, and for Coord3 this rose to 13.

Table 4.10 – TS sizes of 10, 20, 50 and 100 for 15 pollen genera, imaged with light-field microscopy. Best FPTP identification accuracy is highlighted.

Genus	Training set size							
	FPTP (%)				Coord3 (%)			
	10	**20**	**50**	**100**	**10**	**20**	**50**	**100**
Abutilon	100	90	96	96	70	85	92	92
Acacia	30	45	58	68	10	15	26	34
Achyranthes	0	10	28	27	0	0	4	8
Barleria	90	90	98	100	60	70	88	92
Commicarpus	90	90	90	100	30	60	74	91
Erigeron	0	10	16	22	0	0	0	1
Erogrostis	0	0	14	23	0	0	0	2
Grewia	50	55	74	78	40	40	40	49
Gutenbergia	10	15	20	22	0	5	2	1
Hypoestes	100	100	98	100	100	100	98	97
Jasminum	50	45	72	76	0	15	40	44
Lantana	10	25	38	46	0	5	8	15
Sida	40	25	38	47	0	0	8	15
Sphaeranthus	70	90	72	79	50	50	56	61
Tribulis	90	80	100	99	40	70	88	95
Mean	**48.67**	**51.33**	**60.80**	**65.53**	**26.00**	**34.33**	**41.60**	**46.47**
s.e.	**10.10**	**9.26**	**8.40**	**8.13**	**6.61**	**6.61**	**6.63**	**6.65**

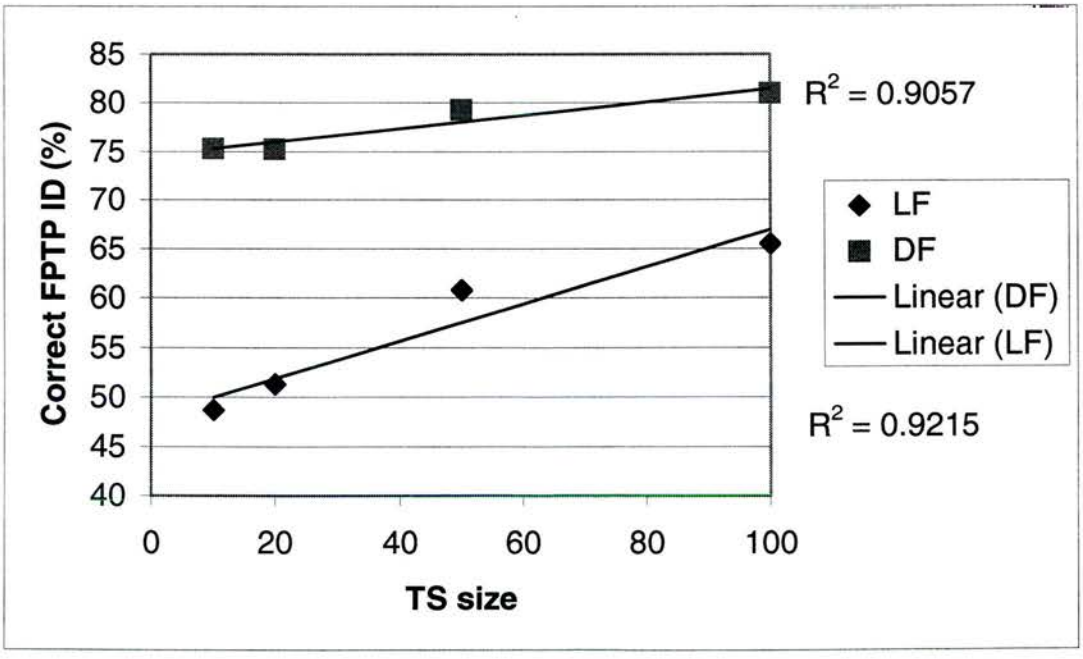
Large differences in dark-field accuracy from genus to genus are also observed but few genera showed a steady improvement in identification as TS size increased (Table 4.11 and Fig. 4.31) and only 5 / 15 (FPTP) and 6 / 15 (Coord3) (highlighted on Table 4.11) had their greatest accuracy at TS100 .

Table 4.11 – TS sizes of 10, 20, 50 and 100 for 15 pollen genera, imaged with dark-field microscopy. Greatest identification accuracy highlighted.

Genus	FPTP (%)				Coord3 (%)			
	10	**20**	**50**	**100**	**10**	**20**	**50**	**100**
Abutilon	100	100	100	98	100	100	100	96
Acacia	40	20	38	44	0	5	10	11
Achyranthes	40	30	66	69	20	25	36	40
Barleria	100	100	100	100	90	95	100	98
Commicarpus	90	95	90	93	80	70	78	89
Erigeron	40	30	30	48	0	5	6	6
Erogrostis	20	25	38	41	10	15	2	6
Grewia	40	70	64	66	10	25	24	41
Gutenbergia	90	80	88	81	50	55	74	69
Hypoestes	80	95	96	98	50	75	82	92
Jasminum	100	100	94	94	80	95	80	77
Lantana	90	85	86	86	50	65	72	77
Sida	100	100	100	97	80	100	90	88
Sphaeranthus	100	100	100	100	100	100	100	100
Tribulis	100	100	100	100	100	100	98	99
Mean	**75.33**	**75.25**	**79.31**	**81.00**	**52.87**	**59.20**	**63.42**	**65.93**
s.e.	**7.68**	**8.24**	**6.59**	**5.64**	**9.90**	**9.68**	**9.55**	**9.16**

When LF and DF mean values are plotted together it can be seen that TS size has a greater effect on LF than on DF performance. If linear trendlines are fitted to the FPTP data set the gradient for LF is 2.7 times that for DF (Fig. 4.31). Linear trendlines describe the data well, as reflected in R^2 values greater than 0.9.

Fig. 4.31 – FPTP accuracy for LF and DF as TS size increases. The linear gradient of the LF trendline is 2.7 times that of the DF trendline.



However, logarithmic trendlines fit these data more closely (Figs. 4.32 and 4.33), the small deviations from a logarithmic trend being trivial compared to the large standard errors within each TS size set (errors were omitted from Figs. 4.32 and 4.33 as substantial error overlap would complicate the graph). The curves approach an asymptote by TS100.

Fig. 4.32 – FPTP accuracy as TS Size increases, both for LF and DF with logarithmic trendlines.

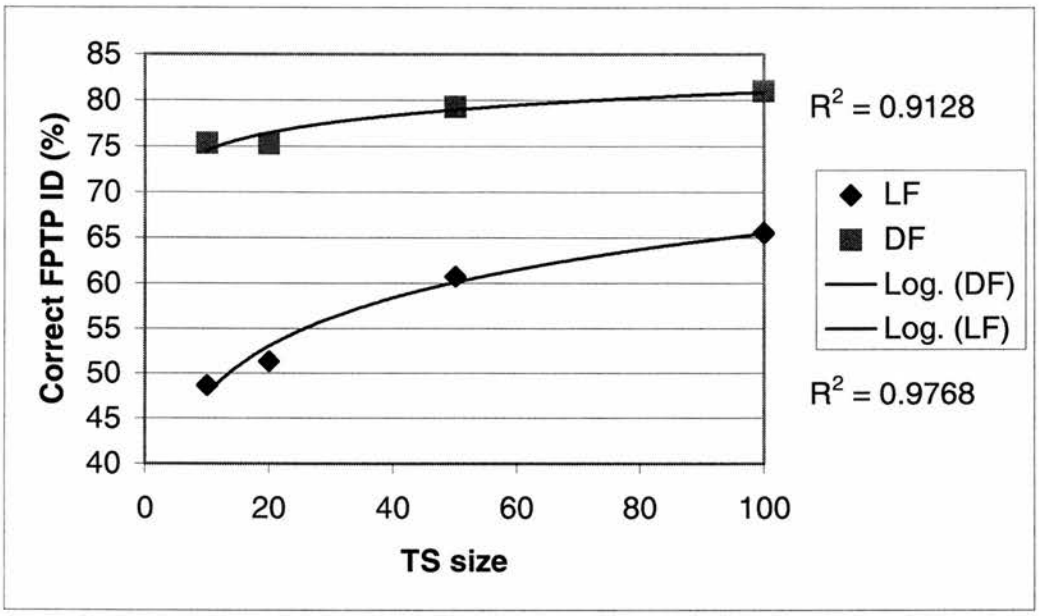
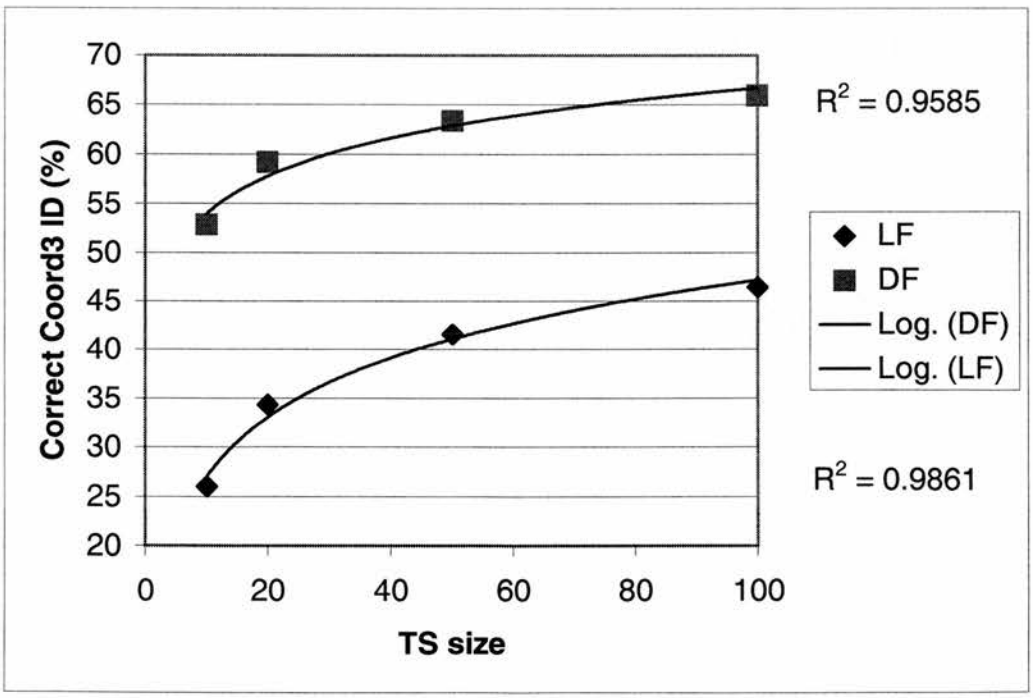


Fig. 4.33 – Coord3 ID success as TS size increases, both for LF and DF.



4.5.4 Discussion

Increasing TS size (above TS10 used in section 4.3) does improve identification success substantially. Most studies of automated identification systems have employed small training sets of five to ten representatives per taxon, whilst commonly observing that ideally they should be larger. The general trend is for performance to increase asymptotically with the TS size, with larger set increases necessary to distinguish between species that are narrowly separated in morphological space (Gaston & O'Neill, 2004). This is the trend seen in DAISY studies to date (Weeks *et al.*, 1997; Weeks *et al.,* 1999b; Gauld *et al.,* 2000; Pajak, 2001; Watson *et al.,* 2004) so it is unsurprising to observe a similar shape of curve in pollen identification.

In the ichneumonid, bee and midge studies, the asymptote was approached by TS20. The present pollen results are more similar to those reported by Watson *et al.* (2004) for moths, in which the curve was starting to level off by T50. These two projects differ from the wing venation studies in two ways. First, the pattern analysed is one of changing colour tones, as opposed to a monochrome venation pattern. Second, the images include a departure from two-dimensionality. Watson *et al.* (2004) imaged live moths in a range of resting postures, in none of which was the wing lamina horizontal. This introduced errors due to perspective and led to some pattern distortion. The three-dimensionality in the present pollen study is even more pronounced, so it is unsurprising that pollens appear to need even larger TS sizes than moth wings before the curve asymptotes.

Fortunately, it is generally easy to image large numbers of pollen grains, as hundreds may be present on a single sample slide. The other studies that have used large TS sizes have investigated phytoplankton identification by flow cytometry (e.g. Boddy *et al.,* 2001). This system used a focused stream of particles and thus was able to create huge data sets very quickly by processing in the order of 10^3 cells per second (Boddy & Morris, 1997).

The improvement with increased TS size is more pronounced for light-field imaging than for dark-field. However, LF starts out with poorer performance (49% at TS10 as opposed to 75% for DF, FPTP) so there is much more potential for improvement. LF images seem to have more surface detail visible (Fig. 4.34). If this leads to greater intrageneric variation that would also help explain the greater training benefit in LF.

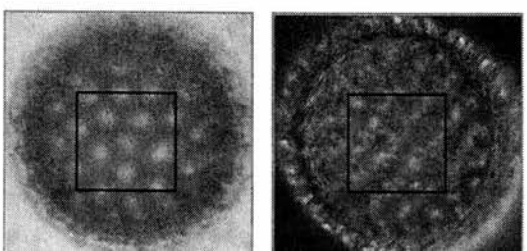


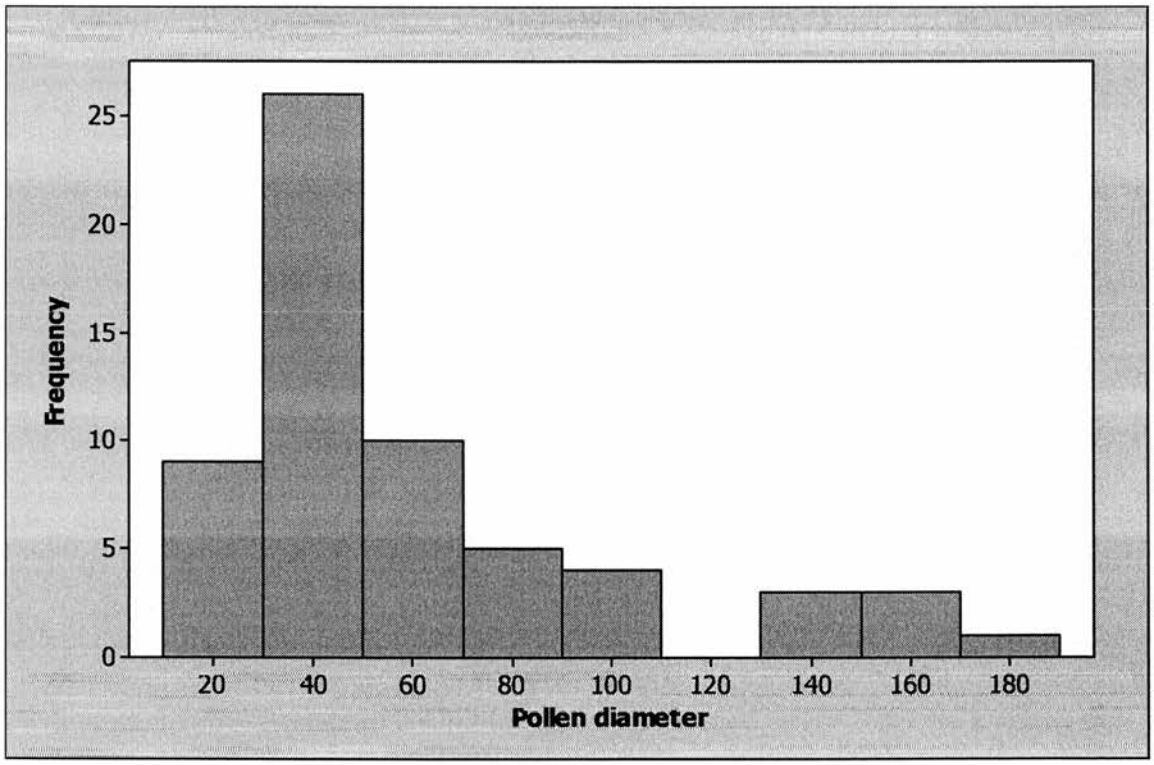
Fig. 4.34 – *Tribulis* pollen LF (left) and DF (right). The LF image shows surface detail (boxed) more clearly.

4.6 The effect of pollen size

4.6.1 Introduction

In addition to surface sculpturing, pollen grains also vary in size. The pollen sampled ranged from 15μm in *Echiochilon* to 185 μm in *Abutilon*. The modal size class is 30 - 50 μm containing 26 genera, frequency is less than half of this in either of the neighbouring classes. There are no grains smaller than 10 μm but the histogram extends into a long tail for the largest diameter classes (Fig. 4.35). Fig. 4.36 shows the full range of pollen diameters, calculated as the means of ten grains per pollen.

Fig. 4.35 – Mean frequencies of maximum pollen diameter for 52 genera of Kenyan pollens, calculated as the average of 10 randomly-selected grains per genus.



Pollen size is a critical factor in non-automated pollen identification but grain size is standardised when sub-sampling to create a NPT so this important information is lost to DAISY. If all grains are imaged at the same magnification, then cropping each grain and pasting it within a black frame of standard diameter can maintain their relative sizes. This results in some genera that previously could have been confused looking very different (Fig. 4.37).

Fig. 4.36 - Mean diameter of Kenyan pollen genera (n = 10). Error bars represent maximum and minimum diameter values.

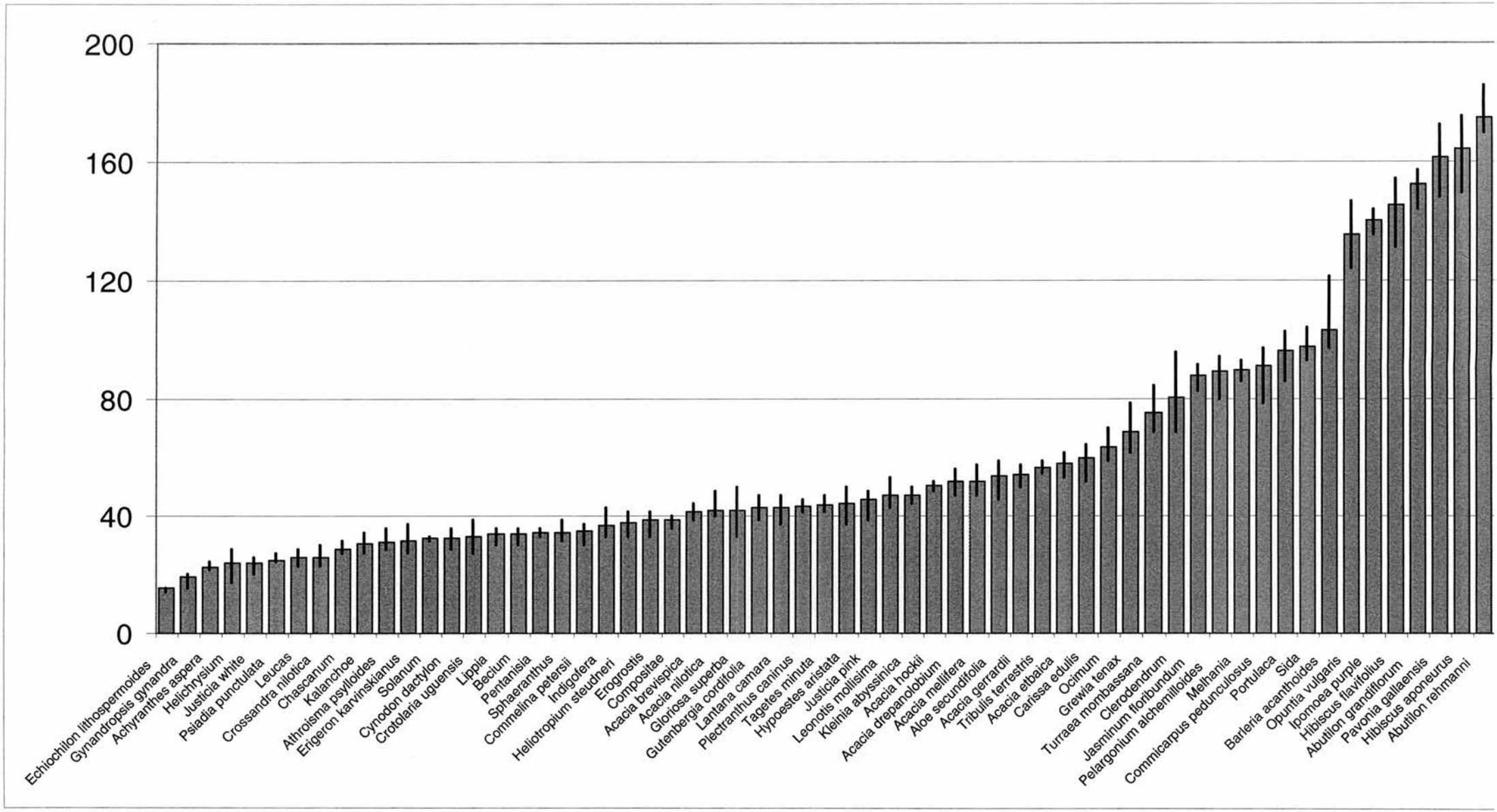
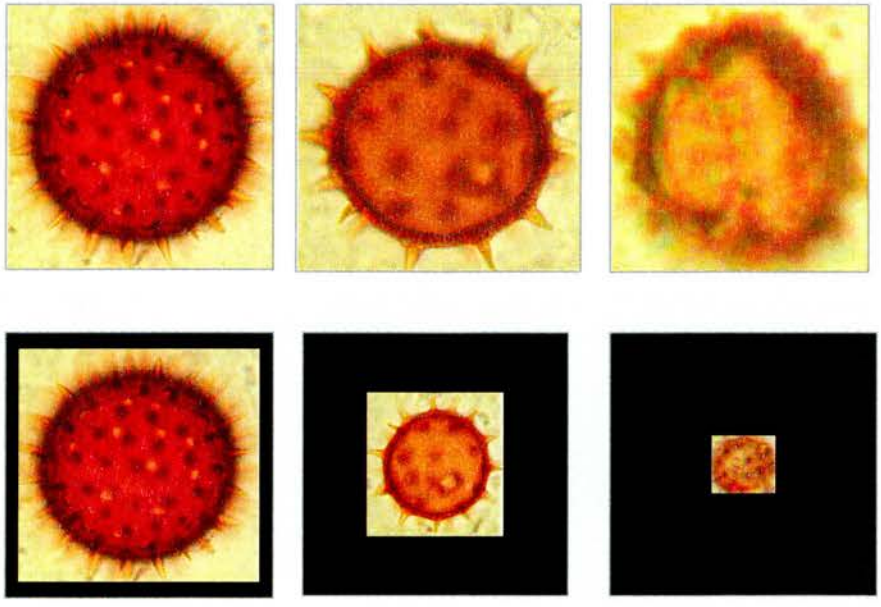


Fig. 4.37 – Top tow: Single pollen grains of *Abutilon, Melhania* and *Helichrysum* with their sizes standardised. Bottom row: The same grains imaged at a uniform magnification.



Some genera may benefit more from this than others. Genera with small pollen will become a clearly distinguished group but will also be very coarsely sub-sampled for the NPTs, and so may easily be confused among themselves. Medium sized pollen is most common, so including information on their size may render them less distinctive. Large pollen should gain the most from this approach, as there is substantial size variation within this class and they will not lose detail in the sub-sampling.

4.6.2 Methods

The full set of pollens acetolysed in section 4.3 (52 genera) were used for analysis. Ten grains from each genus were digitally imaged with minimal camera zoom, this being the only wide angle that would capture all the largest grains, thus ensuring standardization of their sizes. The images were cropped and rotated so that the longest dimension ran vertically (as described in section 4.2). They were then each pasted into the centre of a 600 by 600 pixel black square and the resulting file saved as their 'true size' image. These images were then jack-knifed with the NPT set at 32. The same image sets were also identified using the methods described in section 4.3 (i.e. with size standardised) and the results compared.

It is interesting to compare any improvement in identification accuracy from maintaining size with the improvement from cleaning pollen. To allow this comparison some genera were deleted, such that only the 27 genera used for the uncleaned / cleaned comparison (in section 4.3.3) remained. This true size set (TS10) of 27 genera was then identified using the same methods.

4.6.3 Results

I will begin by discussing the 52 genera data set. Pollen identification was substantially more accurate when pollen size was maintained. The sized images were identified with almost twice the accuracy of those with size standardised, both for FPTP and Coord3 (Table 4.12). Thirty-six genera were identified more accurately by FPTP when size was maintained, while six were unchanged and nine were identified less accurately. In the Coord3 results 29 were identified more accurately, 14 had no success with either method and eight were identified less accurately. The certainty of FPTP and Coord3 identifications also increased when size was maintained, + 27% for FPTP and + 16% for Coord3.

Acacia was identified more accurately with size maintained (FPTP + 60%, Coord3 + 30%). As *Acacia* pollen identification is essential to *Acacia* pollination studies this is a major benefit of maintaining size. The nine genera identified less accurately were *Sida* (-40%), *Hypoestes* (-20%), *Ocimum* (-20), *Barleria* (-20%), *Hibiscus* (-20), *Athroisma* (-10%), *Solanum* (-10%), *Plectranthus* (-10%) and *Jasminum* (-10%). *Sida, Barleria* and *Hibiscus* are three of the genera with the largest pollen (diameter > 90μm). This group included none of the genera with the smallest pollen (diameter < 30μm) but otherwise covered the full size range.

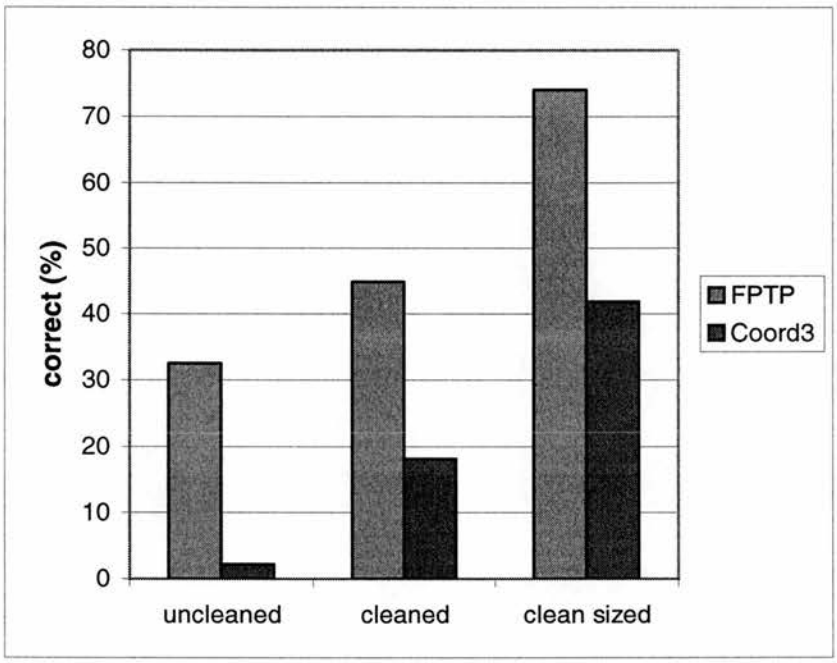
There is no significant correlation between pollen diameter (from Fig. 4.38) and true size FPTP accuracy ($R^2 = 0.0507$) or between pollen diameter and the change in FPTP identification accuracy when size is introduced ($R^2 = 0.0552$).

Table 4.12 – Fifty-two pollen genera, with ten grains per genus, both with size standardised and size maintained. The true size images were identified almost twice as accurately. The more accurate identifications are highlighted.

Genus	FPTP (%)		Coord3 (%)	
	standardised	maintained	standardised	maintained
Abutilon	80	100	40	60
Acacia	30	90	0	30
Achyranthes	10	60	0	0
Aloe	20	100	10	60
Athroisma	30	20	0	0
Barleria	90	70	70	60
Becium	10	50	0	0
Carissa	10	80	0	20
Chascanum	10	30	0	0
Clerodendrum	20	40	10	0
Commelina	20	40	0	0
Commicarpus	90	90	70	40
Composite	20	100	0	70
Crossandra	10	20	0	0
Crotolaria	10	50	0	10
Cynodon	10	50	0	0
Echiochilon	40	90	0	70
Erigeron	20	20	0	0
Erogrostis	0	30	0	0
Gloriosa	50	90	20	70
Grewia	60	80	30	50
Gutenbergia	0	60	0	30
Gynandropsis	0	100	0	40
Helichrysum	0	20	0	0
Heliotropium	70	90	40	60
Hibiscus	100	80	100	80
Hypoestes	60	40	30	10
Indigofera	0	20	0	0
Ipomoea	20	60	0	50
Jasminum	80	70	20	30
Justicia	60	90	30	60
Kalanchoe	40	40	0	0
Kleinia	10	50	0	30
Lantana	30	60	0	20
Leonotis	40	80	0	40
Leucas	40	60	0	10
Lippia	40	60	0	10
Melhania	20	100	0	100
Ocimum	20	0	0	0
Opuntia	30	50	0	10
Pavonia	40	40	0	10
Pelargonium	10	80	0	30
Pentanisia	20	60	0	0
Plectranthus	70	60	0	10
Portulaca	40	80	0	30
Psiadia	30	50	0	10
Sida	100	60	80	0
Solanum	20	10	10	0
Sphaeranthus	60	60	10	20
Tagetes	10	80	0	40
Tribulis	100	100	90	70
Turraea	missing	90	missing	10
MEAN	**35**	**62**	**13**	**25**
Standard error	4.1	3.75	3.59	3.78
Certainty (%)	34.81	61.54	71.43	87.84

I will now move on to the 27 genera data set. When the identification accuracies of these images are compared to those from the unsized pollens, both uncleaned and cleaned, it can be seen that the accuracy improvement made by sizing is greater than that made by cleaning. The FPTP accuracy rose by 12% with pollen cleaning, then a further 29% when the cleaned pollen genera were also sized. The Coord3 accuracy shows a similar trend, as it rose 16% with pollen cleaning, then a further 24% when cleaned pollens were sized.

Fig. 4.38 – Mean identification accuracy of uncleaned pollen (no size), cleaned pollen (no size) and clean sized pollen.



4.6.4 Discussion

Maintaining the relative sizes of pollen does seem to improve accuracy substantially, almost doubling overall success levels and raising the *Acacia* FPTP accuracy from 30% to 90%. This is encouraging as this approach is quick and requires no special equipment. Here it has only been tested with light-field imaging and low TS size. Dark-field images with TS100 should perform even better.

It has been shown that some genera respond better than others to true size analysis. However, this does not appear to be governed by pollen size. There is no evidence that the smallest pollens are unsuited to the true size approach, as all of the smallest pollens performed the same as or better than they did when size was standardised. Neither is there evidence that the largest pollens benefit disproportionately from true size. This refutes two of my original hypotheses.

Nevertheless, if different genera respond better to different approaches (size standardised or size maintained) then maintaining size is complementary to the standardised size approach and they would form a good voting ensemble. It would be preferable if size data could be included in DAISY analysis without the need to paste grains into a larger box (as this is an extra manual task and may mean that important details of small grains are lost to sub-sampling). This functionality is being developed for DAISY.

The decision to use true size must be made early on, to ensure that all grains are imaged at a standard magnification. When the pollens were first imaged this was not considered and many of the smaller grains were imaged with greater magnification. Therefore, many of the smaller grains had to be re-imaged. It would have been preferable to have had an uncleaned + sized treatment for the 27 genera analysis but this was not possible as some of the slides were not available to be re-imaged, having been taken back to Kenya as a fieldwork reference collection.

4.7 Pool size

4.7.1 Introduction

The range of flowering plants within a small geographical area, such as Mpala Research Centre, can be huge (Prof. Truman Young's plant inventory listed over 200 angiosperm species within Mpala). Therefore, to be a useful tool for pollen identification DAISY must be able to consider many taxa. The number of taxa used in DAISY training is known as the 'pool size'.

The response of DAISY to large pool sizes is unclear. Weeks *et al.* (1999b) and Gauld *et al.* (2000) assessed the importance of pool size in the identification of biting midges. There was a 12% decline in FPTP identification accuracy as pool size was increased from 2 to 49. If this decline had been linear then less than half the fly specimens would have been identified correctly by a pool size around 200, so DAISY would be of little use for surveying a diverse group. Instead, the decline was a concave curve Poisson "tail-off", the initially steep decline in accuracy became more gentle. An increase from five to 15 species-classifiers resulted in a 5% decrease in FPTP accuracy, but a similar increase from 35 to 45 species-classifiers caused only a 2% drop. The data suggested that accuracy may level off at around 80% for large groups of midge species (Fig. 4.39).

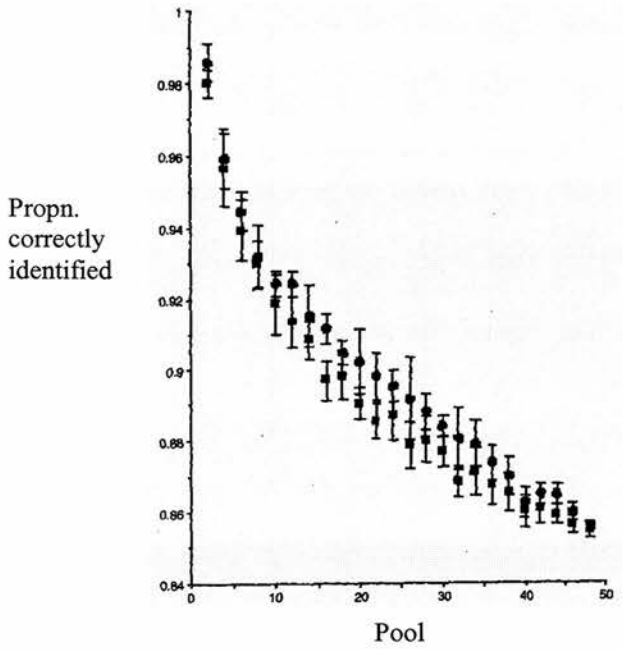


Fig. 4.39 – Accuracy decline with increased pool size, from two to 49 biting midge species. Gauld *et al.* (2000).

Gauld *et al.* (2000) considered this decline to be a potential weakness of the system, but O'Neill later suggested that it was not representative as the study specimens were atypically similar, coming from species clouds much closer in morphological space than usual.

The largest automated identification implementation using DAISY known to Gaston & O'Neill (2004) utilized about 200 species (previously unpublished data), which achieved around 90% FPTP accuracy in identification. However, it must be noted that this was a composite image set containing wings of flies, bees, wasps, butterflies and moths so had much wider spacing in morphological space than 200 species of a single insect order.

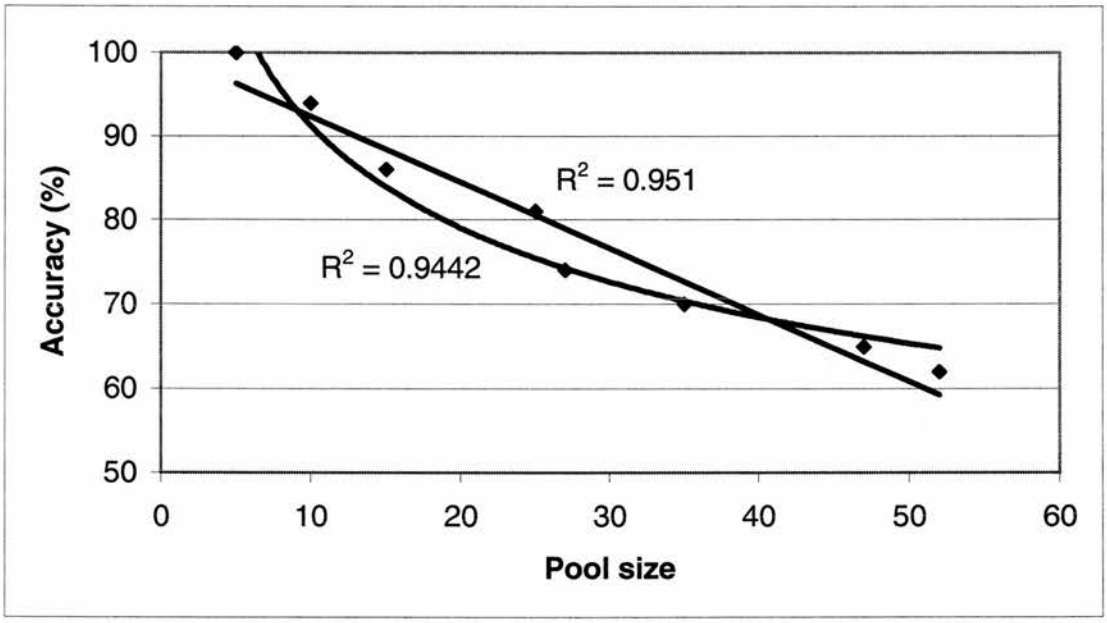
4.7.2 Methods

The cleaned, true size pollen grain images were chosen for this analysis as they were the most accurate methodological combination so far, thus TS size was limited to ten. Image sets with 52 and 27 genera had already been processed (for section 4.6) so these were supplemented with sets of pool size 47, 35, 25, 15, 10 and 5. The #47 set was selected by deleting five genera at random (the 10^{th}, 20^{th}, 30^{th}, 40^{th} and 50^{th} when genera were in alphabetical order) from the #52 set and the smaller sets produced by further steps of random deletions (equally spaced within alphabetical order). Each set was jack-knifed (as described in section 3.2.1).

4.7.3 Results

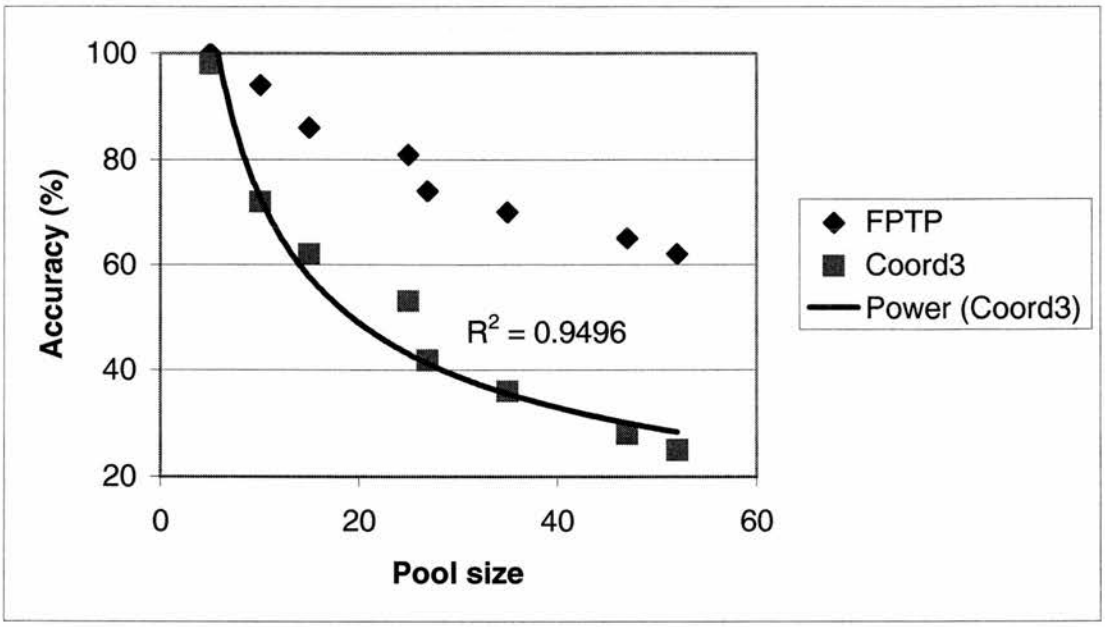
Increasing the pool size decreased the FPTP accuracy more dramatically than in Gauld *et al.* (2000). Accuracy declined from 100% with five genus-classifiers to 62% with 52 genus-classifiers (Fig. 4.40), almost three times the decline reported by Gauld *et al.* (2000) (Fig. 4.39). The Gauld *et al.* (2000) analysis produced a concave curve, but these data fit either a linear trend line ($R^2 = 0.951$) or a power function ($R^2 = 0.9442$) with high confidence. The accuracy at small pool size was very good, achieving 100% correct identifications with five genus-classifiers (Fig. 4.40), whereas the Gauld *et al.* (2000) analysis had just 96% accuracy with this pool size (Fig. 4.39).

Fig. 4.40 – The decline in FPTP accuracy as the number of genera is increased from 5 to 52.



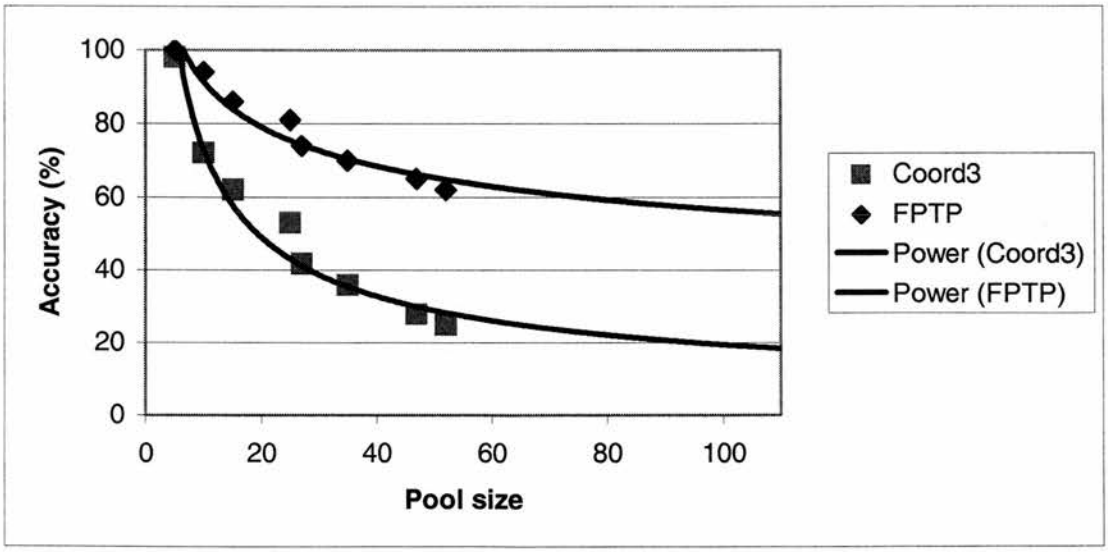
In this case I also have data for Coord3 accuracy. Coord3 demanded greater certainty of each identification, e.g. when there were 10 genus-classifiers the FPTP certainty was 94% while the Coord3 certainty was 97.3%. The decline in Coord3 accuracy is much greater than the decline in FPTP accuracy, falling from 98% with five genus-classifiers to 25% with 52 genus-classifiers (Fig. 4.41). This decline is similar in shape to that of Gauld *et al.* (2000), fitting a power function ($R^2 = 0.9496$) much better than a linear trend line ($R^2 = 0.8833$).

Fig. 4.41 – The decline in FPTP and Coord3 accuracy as the number of genera is increased from 5 to 52.



If a power function is fitted to both FPTP and Coord3 and the curves extrapolated forward to 110 genus-classifiers the curves both almost asymptote by 100 (Fig. 4.42). However, these asymptotes gives much lower accuracy than the 80% extrapolation of Gauld *et al.* (2000), around 55% FPTP and less than 20% Coord3.

Fig. 4.42 – The declines in FPTP and Coord3 accuracy, extrapolated forward to 110 genus-classifiers.



4.7.4 Discussion

Increasing pool size, from 5 to 52, decreased accuracy by three times more than was reported by Gauld *et al.* (2000). This could be because these pollen images appear to have much larger levels of intraspecific variation (due to differences in size and orientation) than wing venation images, producing

more dispersed genus clouds in morphological space. This dispersion could mean that there is more likely to be overlap in morphological space, leading to misidentifications, so adding extra genus-classifiers leads to a greater decrease in accuracy. The sizing of images (pasting them onto a black background) may also have pushed the many genera of moderate size into more similar regions of morphological space.

The FPTP data were represented well by either a linear trend line or a shallow concave curve. The fall in accuracy with a five genus increase in pool size at small pool size was not substantially larger than that from the same pool size increase at larger pool size. This also differs from Gauld *et al.* (2000). This could be because the pollens varied greatly in size, shape and surface features, so they would cover a large area of morphological space. This contrasts to the ichneumonids, which are known for their similarity. In such a large morphological space it is unlikely that the first few genera to be introduced will overlap sufficiently in feature space that the nearest neighbour is of an incorrect genus. This greater morphological spread between genera could also explain why this pollen analysis achieved 100% accuracy of FPTP identification with five genus-classifiers, when the Gauld *et al.* (2000) analysis gave only 96% accuracy with five species-classifiers.

It is unsurprising that Coord3 was more affected by pool size than FPTP as Coord3 demanded greater certainty of each identification making overlap in morphological space greater. At this higher certainty the trend is a concave curve not a straight line, resembling the curve in Gauld *et al.* (2000) and suggesting that a shallow curve may be the more appropriate trend for the FPTP data.

When the FPTP and Coord3 curves were projected forwards to a pool size of 110 asymptotes were almost reached by 100. While these asymptotes suggest that some identification accuracy would remain at very large pool sizes they come at relatively low identification accuracies, much lower than the 80% of Gauld *et al.* (2000). As FPTP identification is always over 50% accurate, even at very large pool sizes this may still be good enough for DAISY to be used for some tasks, such as screening pollen loads to identify their dominant pollen genera. Coord3 asymptotes below 20% and this is too low to be of practical application.

It may be possible to circumvent the accuracy reduction with increased pool size by combining pollen load sampling with a floral survey and then restrict the analysis to those plants observed. However, bees can forage over a wide area, so it may be difficult to ensure that all potential pollen sources have been included. When Osborne *et al.* (1999) fitted bumblebees with lightweight radar transponders the foraging range varied from 61m to 631m, often to forage destinations beyond the nearest available forage. Gathmann & Tscharntke (2002) compared foraging distance in 16 species of solitary bee,

finding that the maximum distance from nesting site to food patch varies from 150m to 600m and that foraging distance was correlated positively with body length. Honeybees forage even further. Steffan-Dewenter & Kuhn (2003) found that honeybees foraged further in simple (1743m) than in complex habitats (1543m). Beekman *et al.* (2004) found that large and small honeybee colonies both foraged for about 650m from the hive when forage was abundant but when forage was scarce small colonies foraged as far as 1430m and large as far as 2850m.

Even if there is no floral survey on the day of catching a long-term plant list for an area could be combined with data on flowering times to inform the DAISY pool. This would be much easier in countries such as the UK, where flowering seasons are distinct and the flora well studied.

If large pool sizes are unavoidable then accuracy may be improved using a hierarchical approach. When identifying bees to species, ABIS (Hajdaoud *et al.,* 2005) first identifies a list of possible genera. These are then passed to species identifiers for final identification. Algorithms already exist for DAISY to generate optimal genera training sets, which contain those images that are most typical of the genus, lying towards the centre of the species clusters in morphological space (Gaston & O'Neill, 2004).

As more taxa are included in a pool of similar looking taxa there will be increasing overlap in morphological space. This problem may be lessened by the use of manifold reduction (Gaston & O'Neill, 2004; e.g. Penev & Atick, 1996) to amplify the weightings of particular morphological features. This process has been explained further in sections 2.1.7 and 2.3.2.3. Although manifold reduction has been developed for DAISY it has yet to be tested fully and is not part of the version of DAISY used in the present study.

4.8 The effect of normalised polar thumbnail (NPT) size

4.8.1 Introduction

I introduced NPT size and reasons to investigate this parameter in sections 2.3 and 3.3. For wing venation patterns (section 3.3) an NPT size of 24 (i.e. 24 x 24 pixels) gave a slightly better performance overall than did an NPT of 32. The wing patterns of some species were more distinct with small thumbnails, but others were identified more accurately with large thumbnails. However, pollen images are very different to images of wing venation, they are colour-rich and represent three-dimensional objects, so the effect of NPT size may be very different.

It is no simple matter to judge whether large or small NPTs are better suited to pollen. If certain features of a particular pollen make it distinctive in non-automated identification then a larger NPT may accord these features greater information value and improve identification performance. However, this argument may have two possible flaws. The first flaw concerns orientation (Fig. 4.43). Pollen images in the present study have not been standardised in orientation for two reasons. First, they are fixed in place on the microscope slide and can only be viewed clearly from one angle. Second, to know how to standardise the orientation of a particular grain, you need to know what taxon it is and, of course, if this knowledge is already available then clearly there is no need for undertaking an identification.

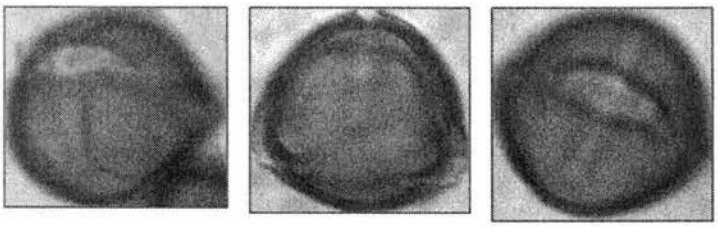


Fig. 4.43 – *Solanum* grains have a simple, distinctive shape but it looks different from different angles.

The second problem concerns the geometry of surface features. The pollen of a particular genus may appear distinctive in its general surface appearance; for example, *Commicarpus* is a periporate pollen (refer to section 4.1.2 for terms in pollen taxonomy) making it look 'spotty'. Unfortunately, such grains, even when imaged with standard orientation, can differ in the size, number and location of those pores (Fig. 4.44). So although they look very similar to the human eye in general terms, they are quite different when the images are divided into pixels and compared pixel by pixel. If a higher level of sub-sampling makes this irrelevant then it will reduce intraspecific variability and such NPTs may be processed more accurately.

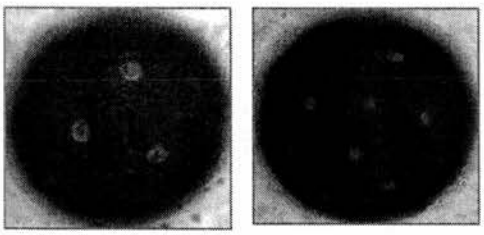


Fig. 4.44 – *Commicarpus* grains differ in the locations and sizes of their surface pores.

4.8.2 Method

Two sets of analyses were undertaken to study the effects of NPT size. The first set looked at a wide range of possible NPT sizes and the effect on each genera. The second set investigated the generality of those results, testing just the smaller NPT sizes and obtaining overall means for three further pollen processing regimes.

The first set of analyses used the 'raw' (i.e. uncleaned, unbalanced TS sizes, light-field and with standardised size) image set of 33 pollen genera. This image set has already been described in section 4.2 and tabulated in Table 4.4. The same set of pollen images was identified repeatedly by jack-knifing (see 3.2.1) changing only the NPT size from replicate to replicate. The NPT sizes tested were 20, 24, 32, 48, 64 and 80.

The second set of analyses compared the raw image data to those obtained from three processed image sets:

- **Cleaned**, TS10, 33 genera, light-field and standardised size (see section 4.3).
- Cleaned, TS10, 15 genera, **dark-field** and standardised size (see section 4.4).
- Cleaned, TS10, 47 genera, light-field and **relative size** (see section 4.6).

The same sets of images were identified repeatedly with NPT size 20, 24 and 32.

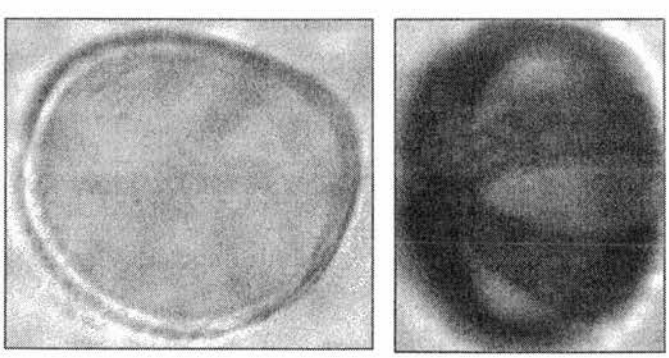
4.8.3 Results

In the raw image set, different genera had different NPT optima, as seen in the scattered optima highlights of Table 4.13. In a few cases, such as *Anthericum* and *Becium* (Fig. 4.45), NPT size had very little effect (5% or less) on FPTP accuracy. *Anthericum* had minimal success at all NPT sizes (0 – 5%) and *Becium* was consistently moderate (26 – 30%). In others, such as *Justicia* (72%), *Commelina* (55%) and *Echiochilon* (45%) (Fig. 4.46) the response to NPT size was great. These genera went down to 0 – 4% accuracy at the largest thumbnail size.

Table 4.13 – Different raw pollen genera had different NPT size optima (highlighted).

GENUS	TS	FPTP (%)						Coord3 (%)					
		20	24	32	48	64	80	20	24	32	48	64	80
Abutilon	17	59	65	82	82	82	88	18	12	24	35	35	35
Acacia	33	27	21	24	12	6	0	0	0	3	0	0	0
Anthericum	19	5	5	0	0	0	0	0	0	0	0	0	0
Barleria	23	30	26	13	4	4	9	9	4	0	0	0	0
Becium	23	30	30	30	30	30	26	9	9	17	13	13	13
Carissa	20	15	0	10	15	15	5	0	0	0	0	0	0
Commelina	40	45	55	33	20	13	0	18	20	15	15	5	0
Commicarpus	11	55	55	82	64	64	73	0	9	18	27	27	27
Composite	28	11	18	14	18	11	14	0	4	0	0	4	4
Craterostigma	24	33	46	46	29	17	8	4	13	13	4	0	4
Crossandra	20	95	100	95	75	70	70	85	90	90	70	70	70
Croton	32	59	50	50	59	50	44	19	13	13	25	28	28
Echiochilon	20	45	25	10	0	0	0	0	5	0	0	0	0
Grewia	24	38	29	25	42	38	42	4	0	4	4	4	4
Gutenbergia	25	12	12	12	12	12	12	0	0	0	0	0	0
Gynandropsis	33	36	39	36	24	3	0	9	6	12	3	0	0
Heliotropium	21	19	14	14	10	0	0	0	0	5	0	0	0
Hibiscus	64	48	44	42	44	36	39	23	22	5	20	16	17
Indigofera	32	25	25	19	16	16	13	0	3	3	6	6	6
Ipomoea	67	63	60	57	45	43	39	18	18	12	10	12	9
Justicia	45	76	73	64	36	18	4	44	47	47	22	11	2
Kalanchoe	25	60	52	56	48	48	48	0	0	0	0	0	0
Leucas	30	43	40	43	57	77	70	10	17	10	27	27	33
Lippia	23	13	9	13	9	9	9	0	0	0	0	9	0
Lycium	28	18	18	7	11	7	7	0	0	0	0	0	4
Melhania	8	13	13	13	0	0	0	0	0	0	0	0	0
Ocimum	28	25	32	21	7	7	7	0	7	11	7	7	7
Pavonia	12	33	42	25	25	25	25	0	0	0	8	0	8
Pentanisia	27	15	11	11	15	19	26	0	0	0	0	0	0
Plectranthus	40	23	28	37	20	23	15	3	8	6	8	5	5
Portulaca	44	61	73	64	57	41	25	23	20	16	7	0	2
Priva	26	50	46	27	27	12	15	12	15	8	4	0	0
Solanum	22	41	50	45	64	55	41	0	5	5	14	14	14
ALL	**28.3**	**40**	**39**	**36**	**31**	**26**	**23**	**11**	**12**	**12**	**10**	**8**	**8**
Opt. genera #		17	13	5	6	3	3	5	9	9	7	8	10

Fig. 4.45 - *Anthericum* (left) and *Becium* (right) were the genera least affected by NPT size.



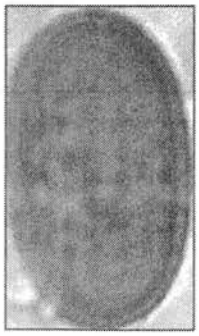
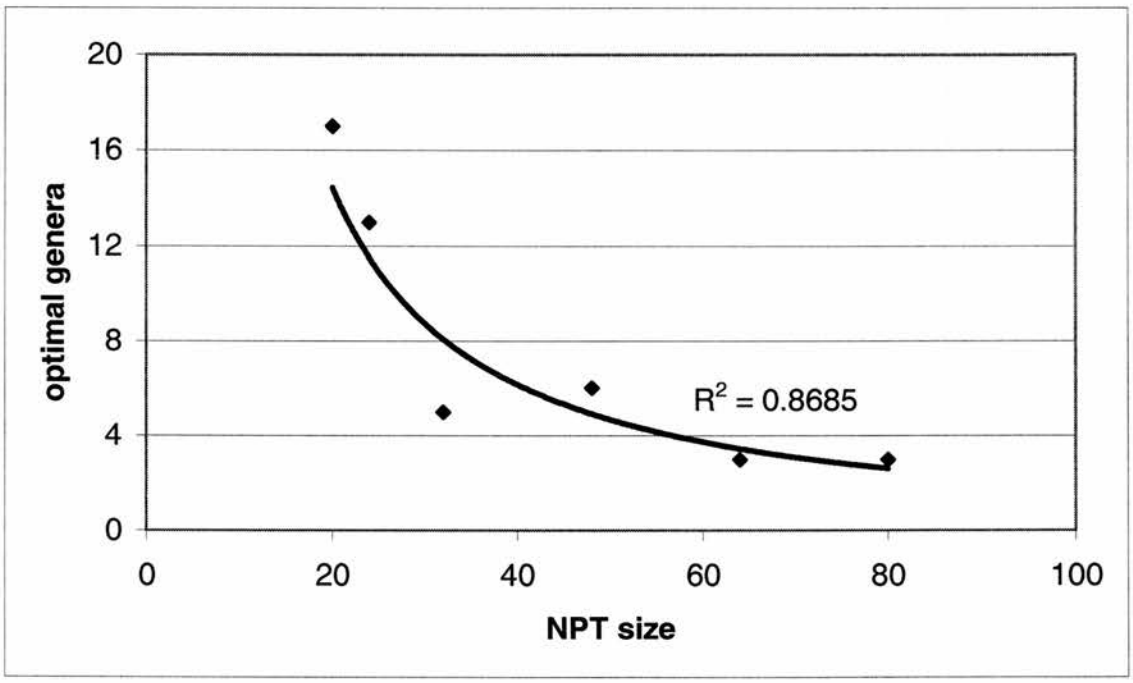


Fig. 4.46 – *Justicia* was most affected by NPT size and best identified with low NPT size.

In the FPTP results about half (17 of 33) of genera had optimal accuracy with NPT20. The number of genera for which an NPT size was optimal declined as NPT size increased (Fig. 4.47). This decline was a concave curve, fitting a power function with R^2 of 0.8685. When the Coord3 results are examined in the same way the genera optima are more evenly spread, with no overall trend apparent.

Fig. 4.47 – The number of genera with best accuracy at each NPT size (FPTP).

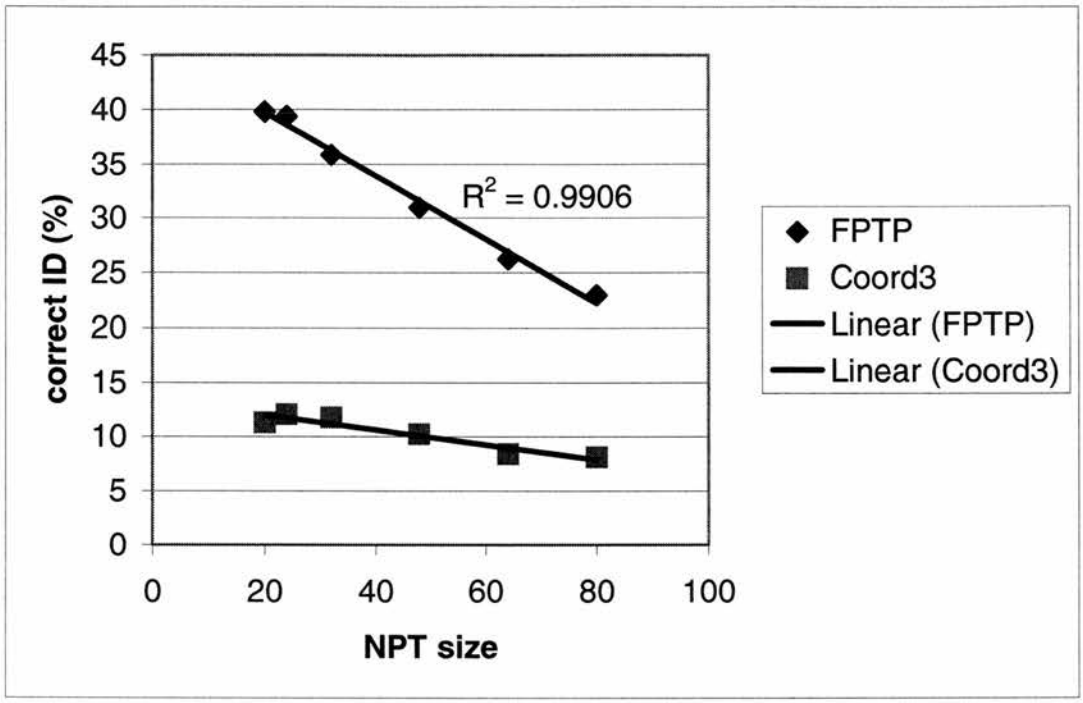


The combined accuracies for all raw image genera (labelled as 'ALL' in Table 4.13) decreased as NPT size increased (Fig. 4.48). The fit of the FPTP data to a linear trend line is almost perfect ($R^2 = 0.9906$), accuracy declining by around 20% as NPT size enlarges from 20 to 80. This produces a gradient of -0.3. The Coord3 data fit a linear trend line less well ($R^2 = 0.807$) and the mean identification accuracy only varied by 4% so the gradient is very shallow, just – 0.07.

So both the number of optimal genera and the mean accuracy values suggest a reduction in FPTP accuracy and no major effect on Coord3 accuracy as NPT size increases. As the smallest NPT sizes are

giving the highest identification accuracies it makes sense to focus on the smaller NPT sizes (20, 24 and 32) for the second set of analyses.

Fig. 4.48 – Mean identification accuracy for raw pollen images processed with different NPT sizes.



When the more processed image sets are considered over this smaller NPT range they differed in their most accurately identified NPT size but the differences in mean accuracy are consistently small (less than 6%) (Table 4.14).

Table 4.14 – Identification accuracy percentages FPTP with change in NPT size for four different image processing regimes. The responses to NPT size are small and vary between processing regimes, the highest accuracy in each case is highlighted.

	Method		Raw	Cleaned, LF	Cleaned, DF	Cleaned, LF, sized
	T size		8 to 67	10	10	10
	Pool size		33	15	15	47
NPT	FPTP (%)	20	39.83	48.67	75.33	59.79
		24	39.40	50.00	74.00	65.11
		32	35.87	49.33	75.33	65.32
	Coord3 (%)	20	11.35	26.00	52.87	23.62
		24	12.10	25.33	55.33	27.23
		32	11.78	25.33	52.00	28.09

4.8.4 Discussion

In the first set of analyses (with raw pollen) pollen genera differed markedly in their sensitivity to NPT size. *Anthericum* grains have so few surface features (Fig. 4.49) that the degree of sub-sampling is largely irrelevant. They were slightly (5%) better identified at the smallest NPT sizes, but when this best performance is only 5% accuracy there is little scope for accuracy reduction at larger NPT sizes.

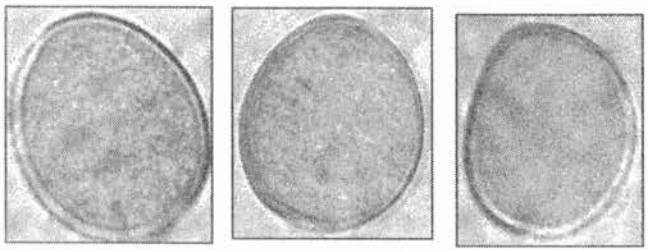


Fig. 4.49 – Three grains of *Anthericum*, the genus that was mis-identified irrespective of NPT size.

Becium is harder to explain. This genus achieved accuracy in the range 26-30% irrespective of NPT size. It has long surface furrows and what appears to be huge intraspecific variation due to dye absorption and orientation (Fig. 4.50).

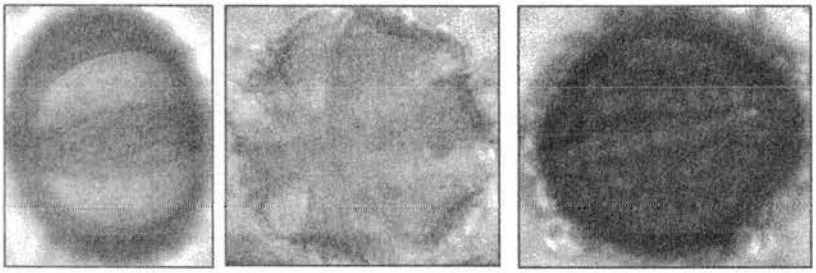


Fig. 4.50 – Three grains of *Becium*, a genus that had 26-30% FPTP accuracy irrespective of NPT size.

This genus does tend to have large patches of similar colour when imaged, so the degree of sub-sampling within such large colour blocks would have little influence on accuracy.

Justicia was the genus with the greatest NPT influence. It was identified 72% more accurately with NPT20 than with NPT80. This makes intuitive sense as *Justicia* has a distinctive elliptical shape with a double band of darker tissue around the edge (presumably the exine) and the central region of consistent colour (pale pink for *Justicia diclipteroides* and deeper pink for *Justicia* sp.1 (Fig. 4.51). The more the images of *Justicia* are sub-sampled the more similar they will become so the greater the likelihood of a correct FPTP identification.

Fig. 4.51 – *Justica diclipteroides* (left) and *Justicia* sp.1 (middle and right).

As different genera achieve highest accuracy from different NPT sizes it may be wise to consider several NPT sizes in an analysis. These outputs could then either be used as a voting ensemble or the greatest certainty identification could be taken from whichever NPT size produces it.

When all genera of raw images are considered the general trend was for smaller NPT sizes to give greatest accuracy for the most genera and the highest mean accuracy values. The decrease in the number of optimal genera from NPT20 – NPT32 was large (the steep part of the concave curve in Fig. 4.47) but there was only a small decrease in mean FPTP accuracy (linear). This suggests that the difference in accuracy between the optimal NPT size and neighbouring sizes was only small. The Coord3 results were less sensitive to NPT size as accuracy was low irrespective.

It may be that the optimal NPT size was not tested (e.g. 22) but as the differences in mean accuracy in the NPT range 20-32 were consistently small (less than 6%) so the merits of fine-tuning NPT may not be worth the effort involved (Table 4.14).

4.9 Summary and the way forward

4.9.1 Summary

Pollen cleaning by acetolysis (section 4.3.3), dark-field microscopy (section 4.4.3) and maintaining relative size (section 4.6.3) all increased the identification accuracy of pollen using DAISY so an image processing regime that combines these three is likely to give the greatest identification accuracy. Unfortunately, the importance of relative size was not appreciated when a visit was made to the NHM to image the DF pollens so smaller pollens were imaged with more magnification (camera zoom) than larger pollens. When the importance of relative size later became apparent these images were not suitable for analysis but there was insufficient time to return to London to re-image the DF pollen with standardised magnification (i.e. camera zoom). For this reason, this combination of factors has yet to be tested.

The optimal NPT size of this regime is not known but is likely to be 32, as this gave the highest FPTP accuracy for the dark-field image set and the true size image set (section 4.8.3). Better still, a voting ensemble or screening for high certainty in a wider range of NPT sizes would allow individual taxa to exploit their optimal NPT size. A training set of 20 - 50 pollen is a good compromise between accuracy

and time constraints, especially as dark-field pollens benefited less from increase in TS size than their light-field counterparts (section 4.5.3).

Pool size had a very large impact on pollen identification accuracy, much more than Gauld *et al.* (2000) observed with wasp wing venation. This is a major limitation, as it is likely that many different genera-identifiers (or species-identifiers) will be needed to identify pollen loads. The FPTP accuracy reached an asymptote at above 50% accuracy in the LF, true size trial (section 4.7.3) so it is likely to asymptote at higher accuracy than this in a DF true size set. This accuracy is likely to be further improved by alterations to Floret (the DAISY classification component, see section 2.3), such as hierarchical identification and manifold reduction. While accuracy in the range of 50 - 60% is not enough to identify single grains with high certainty it would usually allow a user to identify the more important pollens in a pollen load.

This accuracy is well behind that of the pollen-centred systems of Trelour *et al.* (2004) and Li *et al.* (2004) which achieved 95% and 100% accuracy. These systems have concentrated on pollen surface structure and textural detail (rather than the characteristics of the equatorial plane). This approach has yet to be evaluated for the DAISY system and should perhaps be the next to be investigated.

The identification of single grains with light microscopy is the most versatile way to identify pollen. However, two other methods can be suggested that may substantially improve the accuracy of pollen identification, the partition function approach and laser-scanning confocal microscopy.

4.9.2 Partition functions

When a pollen grain is identified by DAISY the list of closest neighbours can be obtained as output. When the nearest neighbour is of an incorrect taxa DAISY has produced a mis-identification for that grain. The partition function approach makes use of single grain mis-identifications. If many grains of that taxa are identified the set of mis-identifications can form a 'fingerprint' for that taxa, comprising the other taxa it closely resembles. The partition function approach compares the mis-identification fingerprint of a group of pollens that appear to be a single taxa with the fingerprints of sets of training pollens.

Provisional investigations by Dr Mark O'Neill suggest that the partition function approach could produce identification accuracy above 80% for pollen images like these (O'Neill, pers. comm.). Training could be done using the same sets of images as those used for FPTP / Coord identification. However, there are two weaknesses to this approach. First, the pollen to be identified must occur many

times in a load for a partition function to be obtained. Second, a mixed load must be separated into morphotaxa before these can be identified. This manual separation would be time consuming and would require palynological experience to do well. It would be very easy to mis-group similar looking taxa as a single morphotaxon or split very different orientation views of a single taxon. Thus, this is unlikely to be a useful approach for the non-specialist.

4.9.3 Laser-scanning confocal microscopy

The image seen through a standard light microscope includes the in-focus portion and the out-of-focus portions below and above the plane of focus. The blur caused by the out-of-focus planes is a natural consequence of the optics of the microscope. Confocal microscopy removes out-of-focus blur by passing the light through one or more small apertures, leaving only a thin, highly focused plane. The distance between the specimen and the microscope objective is then changed and new focal planes imaged in the same way. Once a series of planes has been imaged, individual slices can be examined separately or the whole specimen digitally reconstructed as a three-dimensional object. The removal of out of focus haze can increase image resolution allowing improved measurements and visualisation (VayTek, 2005).

Culverhouse and Williams (2003) report a preliminary trial, applying confocal microscopy to zooplankton, and seek partners to form a large project consortium, the first aim of which was to confocal image large numbers of zooplankton species (Figs. 4.52 and 4.53). Despite this proposal, the doubt was still raised that confocal scan rates and depth of field for full 3D large field applications may not yet be appropriate.

Fig. 4.52 – Confocal images showing polymorphism in the the dinoflagellate species, *Dinophysis caudate.* Culverhouse & Williams (2003).

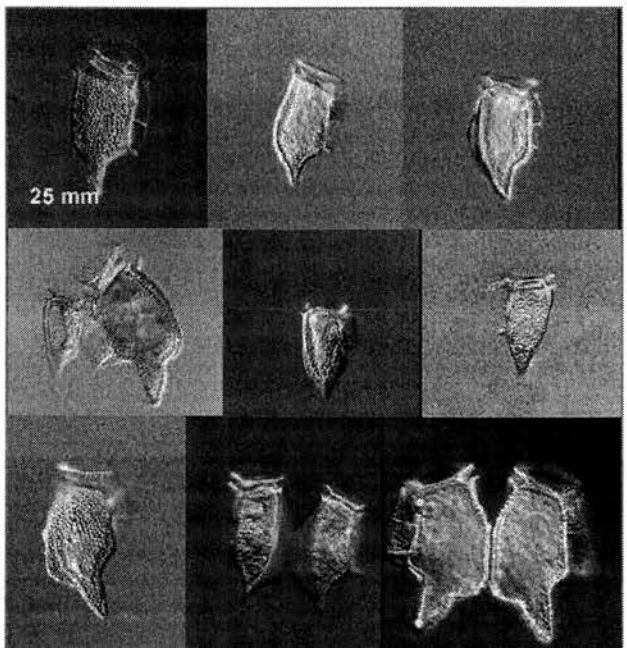
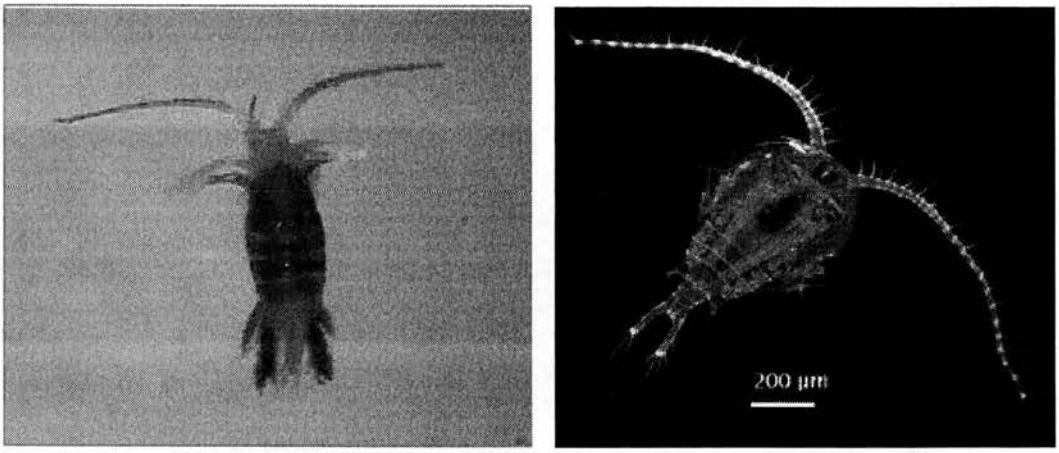


Fig. 4.53 – Copepod imaged with normal light microscopy (left) and confocal microscopy (right). Culverhouse & Williams (2003).



When automatic confocal microscopy was used to image forams for DAISY identification accuracy increased substantially. Arrangements are now being made for a proof-of-concept trial for DAISY, using confocal microscopy to identify larger pollens in the NHM collections (O'Neill, pers. comm.).

The main disadvantage of laser scanning confocal microscopy is that is accessible to very few researchers. Laser scanning confocal microscopes cost at least £60,000 (O'Neill, pers. comm.) so only large research facilities will be able to afford one. As long as a few central laboratories had the laser scanning confocal microscopes this could be a useful pollen identification application for DAISY. Confocal microscopes are likely to get cheaper in the next few years. However, currently they are unlikely to be accessible to ecologists working on pollination ecology.

Attempts have been made to achieve similar results to laser scanning confocal microscopy using a normal light microscope and computer software; this is known as the 'deconvolution approach'. VayTek (http://www.vaytek.com/FAQ.html) produces the MicroTome software. A normal digital image from a light microscope is inputted and image enhancement algorithms (such as nearest neighbour classification) 'deconvolve' the image, i.e. remove the blur from out-of-focus image planes. As no new hardware is necessary this is a much cheaper alternative. However, VayTek admits thick or semi-transparent non-living specimens that require powerful laser light to penetrate into the material will be best imaged by a laser scanning confocal microscope (VayTek, 2005). Some of the larger pollen grains, such as *Abutilon,* may fall into that category. If such an approach does increase image quality and hence the accuracy of identification then it may be worth trying to introduce deconvolution into DAISY.

Chapter 5 – Network analysis of pollen loads

5.1 Introduction

5.1.1 Pollen load analysis

The healthy functioning of ecosystems is essential to humankind (e.g. Costanza *et al.,* 1997) and ecosystem functioning is positively related to biodiversity (e.g. Loreau *et al.,* 2001). Flowering plants and pollinating insects are very diverse, representing approximately one-third of all described species (e.g. Kearns *et al.,* 1998) and it is estimated that more than 90% of plants are pollinated by animals. There is evidence that pollinator loss can lead to extinction of plant species (e.g. Bond, 1995), and the need to conserve pollination interactions has been recognised by the International Pollinator Initiative (São Paulo Declaration on Pollinators, 1999). Pollination systems are under increasing threat from anthropogenic sources: habitat fragmentation, changes in land use, use of pesticides and herbicides, and invasions of non-native plants and animals (Hughes *et al.,* 1997; Sala *et al.,* 2000). A "pollination crisis" is suggested by declines of honeybees (e.g. Watanabe, 1994) and native bees (Allen-Wardell *et al.,* 1998; Kevan, 1999) and the extent of this threat has been hotly debated in recent publications (especially Ghazoul (2005) and Steffan-Dewenter *et al.* (2005)). If we are to manage pollination in natural and agricultural ecosystems we must first understand the basic aspects of plant-pollinator interactions (Kearns *et al.,* 1998; Potts *et al.,* 2003). Pollen load analysis is an important approach to such work as, unlike visitation observation, which relates to a single plant, the quantities of pollens carried may inform the researcher of the flower visits of an entire foraging trip, or even previous trips if "contaminant" pollens escape grooming.

Pollen load analysis is a well-established approach in pollination ecology (e.g. Villaneuva-G, 2002; Bastos *et al.,* 2004; Price *et al.,* 2004). Pollen can be sampled from whole-body washes, from scopa loads (Almeida-Muradian *et al.,* 2005), from faecal pellets (Eltz *et al.,* 2001) or in the case of social bees from pollen traps at hive entrances (Goodwin & Perry, 1992; Langenberger & Davis, 2002). Goodwin & Perry (1992) found that the amount of staminate pollen in the corbiculae (leg sacs) of *Apis mellifera* was closely related to the proportion of staminate pollen on their body, indicating that corbiculae pellets can be used as a reasonable measure of visitation patterns. Many plants tend to use multiple pollinators and *vice versa* (e.g. Waser *et al.,* 1996) and the many plant-pollinator connections involved in pollination make it a complex system to analyse (Green *et al.,* 2005). Recent pollination studies (e.g. Memmott *et al.,* 2004; Devoto *et al.,* 2005) have suggested that network analysis may be a useful approach, as it allows manipulation and analysis of large amounts of data. Many insect-plant interactions can be considered at once so this is a good way to investigate potential secondary

extinctions, i.e. extinction of species A caused by extinction of species B, on which it depends for an ecosystem function (food source or pollination).

5.1.2 Network analysis

A network is a set of items, known as 'nodes', with connections between them, termed 'edges' (Fig. 5.1) (Newman, 2003).

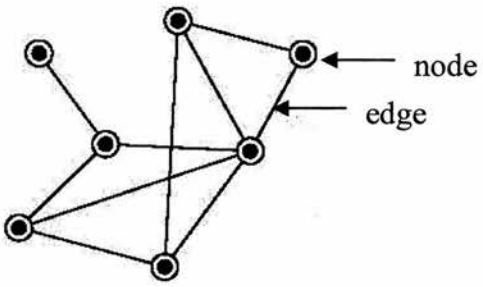


Fig. 5.1 – A small example network with seven nodes and ten edges. Adapted from Newman, 2003.

Networks have been used in the social sciences since the 1930s to address issues such as centrality (which individuals are best connected or have the most influence) and connectivity (whether or how individuals are connected to one another) (Newman, 2003). Watts (2004) discussed recent uses of networks in sociology. Sociological networks are often small enough to interpret visually.

In recent years the availability of computers and communication networks has allowed us to gather and analyse data on a much larger scale. It is not uncommon to see networks with millions or even billions of nodes (Newman, 2003). In its most useful form, a network approach focuses on components of network structure that cannot be recreated from observing individual nodes alone (Proulx *et al.,* 2005). Information networks generally need to be very large, the two best-studied information networks being paper citations and website hyperlinks. A network representing the citations between academic papers (e.g. White *et al.,* 2003) is acyclic as papers can only cite other papers that have already been written. A network representing websites that hyperlink each other (e.g. Albert *et al.,* 1999) is cyclic, there is no natural ordering of sites and nothing to prevent closed loops. Technological networks are similarly large. These are networks designed to distribute a commodity or resource, for example, electrical power grids (Watts and Strogatz, 1998) and airline routes (Amaral *et al.* 2000). However, the best-studied technological network is the network of physical connections between computers that makes up the Internet.

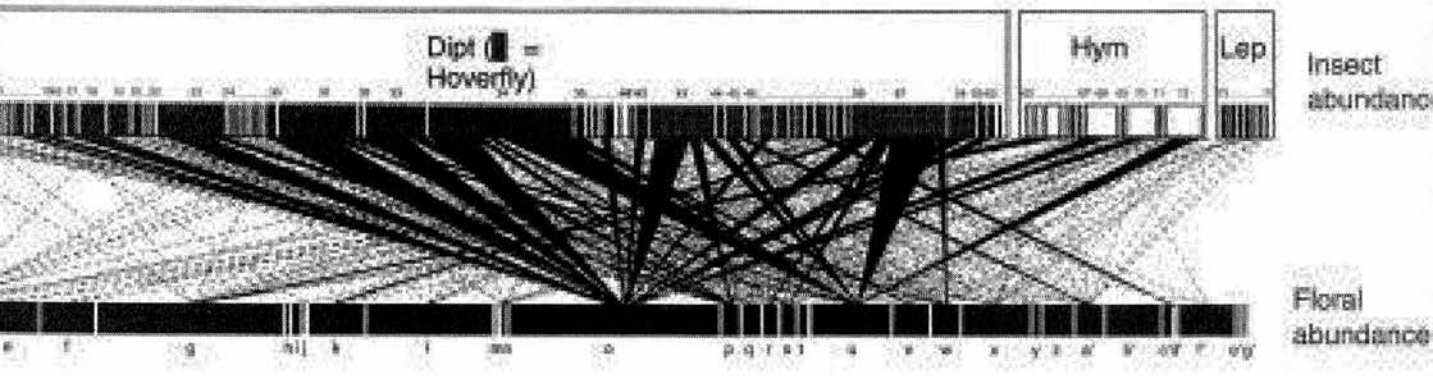
It is impossible to represent very large networks visually; it is more meaningful to assess them statistically, e.g. path length and hub size (how many connections are made with a single node). This change in scale also brings with it a change in analytical approach. Rather than asking which node is most critical to the integrity of the network, it makes more sense to ask what percentage of nodes must

be removed to significantly affect network integrity (Newman, 2003). This approach was taken in a pollination study by Memmott *et al.* (2004) and will be used here so that comparisons can be drawn with Memmott *et al.* (2004).

A number of authors have advocated a food web approach to pollination biology. Jordano (1987) calculated the connectivity of 36 pollinator communities to study the different modes of mutualism. Waser *et al.* (1996) reviewed the levels of generalisation in plant-pollinator interactions, concluding that generalisation was the rule rather than the exception, so that even a small number of species would provide a web of interactions. Kearns *et al.* (1998) pointed out that these richly connected webs make the task of conserving pollination systems more "subtle and complex". They raised the concept of resilience in plant-pollinator interaction webs and considered this approach a priority for future work.

Memmott (1999) illustrated how contemporary methods of web construction and analysis could be applied to plant-pollinator communities. Her networks take the form of connections between two linear planes, one for plants and the other for pollinators (Fig. 5.2). While this is a useful way to reflect diversity and abundance as well as connections it is poorly suited to large networks of hundreds of species.

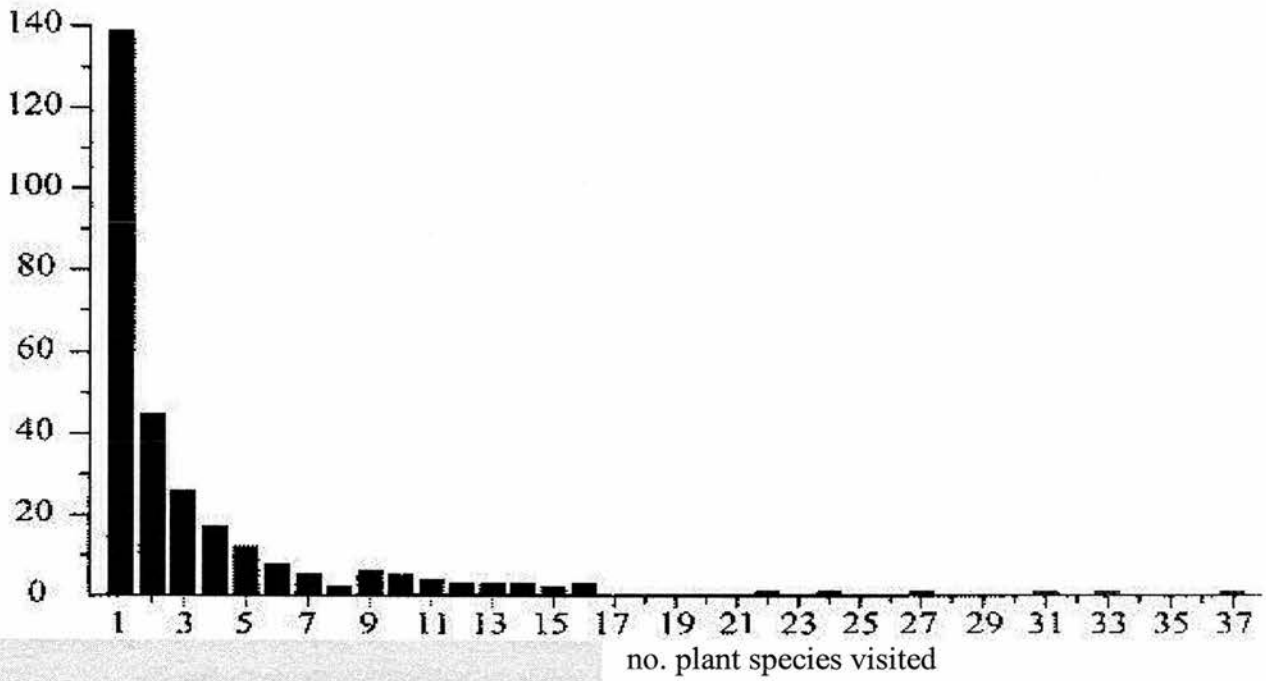
Fig. 5.2 – The style of pollination web used by Memmott (1999), here showing insect visitation. Each species is represented by a rectangle. The lower line represents flower abundance and the upper line insect abundance. The widths of the rectangles and the sizes of interactions between them are proportional to their abundance at the field site. Interactions shown by a dotted line were observed less than 10 times.



Memmott *et al.* (2004) investigated the tolerance of pollination networks to species extinction using data from the large flower visitation surveys of Clements & Long (1923) (918 interactions) and Robertson (1929) (15 265 interactions). The distributions of plant species visited per pollinator species were characterised by long-tailed distributions, many insects visiting only one plant, most visiting fewer than ten and few visiting many plants (Fig. 5.3). Memmott *et al.* (2004) simulated the removal of

different percentages of the pollinator community and recorded the consequent loss of the plants that depend on these pollinators for reproduction (Fig. 5.4). If pollinators were removed at random the decline in plants visited began slowly but accelerated steadily, with the bulk of plant extinctions occurring only after 70-80% of pollinator species had been removed. If the least-linked pollinators were removed first there was minimal decline until more than 90% of the pollinators had been removed, at which point came catastrophic decline. The greatest decline in plant species diversity occurred when the most-linked pollinator species (bumble-bees and some solitary bees) were removed first, this produced a decline that was described as 'essentially linear' (Fig. 5.4). This general tolerance to extinction contrasted with catastrophic declines reported from standard food webs and suggested that there was substantial redundancy in pollinators per plant (Memmott *et al.*, 2004).

Fig. 5.3 – The distribution of plant species visited per pollinator species (Clements & Long, 1923) fitted a strongly concave distribution. Memmott *et al.* (2004).



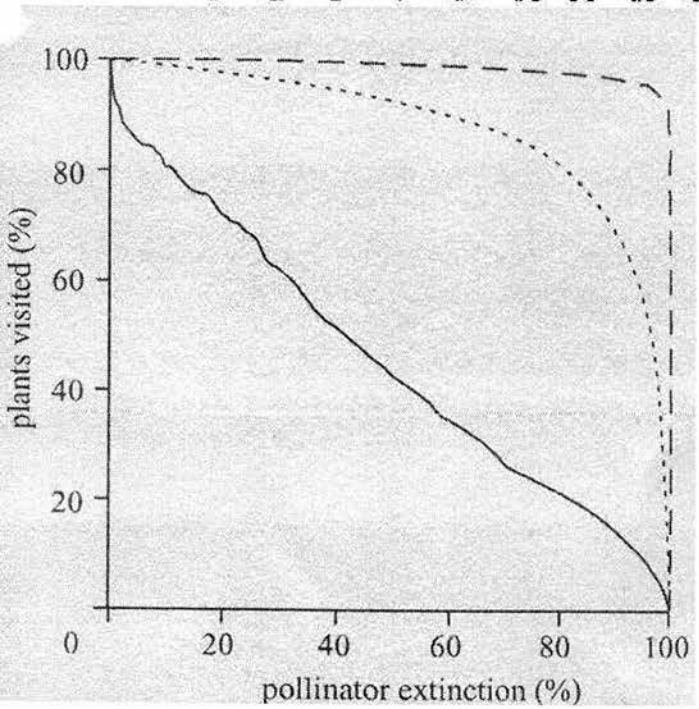


Fig. 5.4 – Extinction pattern for the pollination network of Robertson (1929) in Memmott *et al.* (2004). The solid line showed the deletion impact of the most to least linked pollinators, the dashed line showed least to most and the dotted line is random extinctions.

Devoto *et al.* (2005) produced eight plant-pollinator networks along a steep rainfall gradient in Patagonia. Comparing these networks they concluded that flies dominated the wetter end of the gradient while at the drier end bees played the greater role. Comparison of networks is an interesting idea that will be developed further in this research.

Rather than arranging insects in one plane and plants in another (as in Memmott, 1999), the approach used here arranges plant and insect nodes loosely in space so that highly-connected nodes cluster in the centre and poorly-connected nodes are on the periphery. Connections can be assessed visually and connectivity statistics are generated. This approach is akin to information and technological networks. It would easily scale to very large data sets whilst still allowing taxa to be analysed individually. The analysis has been carried out in collaboration with Dr Mark O'Neill of the University of Newcastle. This type of network has not been applied to ecological data before.

Three sets of networks have been analysed:

1) A large network, including all insect-plant connections, even those representing a single pollen grain on a single insect. Quantification involves the 'hub size' (number of connections) of individual genera and the percentage damage to the network when individual nodes are removed ('damage').

2) A set of networks of those insects that carry substantial amounts of pollen (bees and hoverflies). The amount of a particular pollen that must be carried before we consider an insect to be important for pollination is a subjective figure. By assessing networks with different thresholds it may be possible to study this in a more quantitative way. These networks have load proportion thresholds, set at 10%, 25%, 50% and 75% of total load. Thus, in the 75% network, a connection will only be included if the insect genus carries that pollen genus as 75%+ of at least one load. Only nodes that link to *Acacia* are included in the networks. Removal of individual nodes assesses importance of certain pollinators at different levels of foraging fidelity.

3) A set of networks involving bees and hoverflies carrying any amount of pollen. The effect of different extinction combinations on overall network integrity was assessed by systematic removal of nodes. A similar approach was taken to that of Memmott *et al.* (2004) but importance was assessed by connective damage rather than hub size and a genetic algorithm ('Smasher') generated the most and least damaging combinations. Three scenarios were investigated.

In the first, the combinations of insect genera least important for network integrity are removed, beginning with a single genus and building up 12 genera (out of 13). This simulates a probable extinction sequence, because specialist pollinators, which also tend to be the rarest species (Vázquez & Aizen, 2003), appear at greatest risk of real-world extinction (Rathcke & Jules, 1993). In the second, the groups of genera to be removed are selected at random. This is the null model (Memmott *et al.,* 2004).

In the final scenario, the combinations of insects most important to network integrity are removed. This explores the 'attack tolerance' of networks to the loss of highly connected nodes (Dunne *et al.,* 2002; Memmott *et al.,* 2004). This is a 'worst case scenario' but with simultaneous declines in highly-connected pollinators, such as bumblebees (e.g. Williams, 1982) and honeybees (e.g. Watanabe, 1994) it is feasible.

5.2 Methods

5.2.1 Catching and pollen load sampling

The insect genera involved (numbers in brackets relate to the number of loads) were:

BEES

Superfamily Apoidea

Family Apidae
Amegilla (12)
Apis (107)
Macrogalea (17)
Plebeina (10)
Tetralonia (3)

Family Halictidae
Lipotriches (11)
Patellapis (7)
Pseudapis (14)

Family Megachilidae
Heriades (12)
Megachile (55)

WASPS

Superfamily Vespoidea

Family Eumenidae
Delta (12)

Superfamily Sphecoidea

Family Sphecidae
Bembix (6)
Philanthus (6)
Sphex (6)

FLIES

Family Bombyliidae
Bombylius (17)

Family Calliphoridae
Chrysomya (4)
Hemipyrellia (10)
Rhynia (9)
Rhyncomya (53)
Stegosoma (3)

Family Syrphidae
Ceriana (6)
Eristalinus (18)
Phytomia (18)

Family Tachinidae
Dejeania (16)

Insects were hand-netted at *Acacia* inflorescences and surrounding herbs. Loads were taken either from the leg corbicula of live social bees (*Apis* and *Plebeina*) anaesthetised with CO_2, or from the scopa and bodies of pinned insects. Visible pollen loads were gently dislodged from specimens using a mounted

needle and the pollen adhered onto fuchsin gel. The small amounts of body pollen on wasps, flies and bees without scopa were collected directly onto the fuchsin gel by dabbing the specimen with the gel. The gel was melted and a cover slip placed on top (as in section 4.2). Slides were clearly labelled with insect type, the plant on which the specimen was collected and the date of collection.

5.2.2 Pollen Identification

The initial aim was to use DAISY for pollen identification but this was not viable for two reasons. First, there was no opportunity to clean the pollen loads by acetolysis so the unclean pollen approach of section 4.2 would have been the only option. Second, DAISY identification is still not accurate enough to give pollen identifications with a high level of certainty.

Instead, each load slide was examined under an optical light microscope and identified by comparison with the reference images of 4.2 (unclean) and 4.3 (clean) and the diameter measurements in Fig. 4.36 of section 4.6.1. The correct genus was determined using a hierarchical approach:

1. Several grains of a morphotype were imaged with a x40 lens and minimum camera zoom, then grain diameters measured, averaged and compared with the pollen diameter graph (Fig. 4.36 in section 4.6.1). Genera in that size range were retained for further consideration.
2. Grain shape was compared with that of the remaining genera (using a photo guide) and the closest genus deemed the provisional match.
3. The provisional match was confirmed by comparison with the full image set for that genus.

Stages 1 and 2 each reduced the pool of possible genera and stage 3 confirmed a likely identification. If no provisional match was found, or it was refuted at stage 3, then the process was modified, first matching shape to the full set of genera then considering if the possible shape matches were of a plausible size. In a few cases morphotypes did not match any of the training genera; they were considered 'alien' and excluded from analyses.

5.2.3 Estimation of contribution to the load

Once the pollen morphotypes in a bee or hoverfly load had been identified to genus their relative contributions to the pollen load were estimated as percentages. Those genera contributing at least 10%, 25%, 50% and 75% were considered increasingly important to the carrier insect and interactions over these thresholds were listed for inclusion in threshold networks. The bombyliid, calliphorid and tachinid flies and the wasps only carried very small pollen loads. Therefore, they only appear in the

largest network. The smaller networks (where 10%+, 25%+, 50%+ or 75%+ of pollen must come from a certain plant) only include bees and the hoverflies *Eristalinus* and *Phytomia.*

5.2.4 Network analyses

Six lists of insect-plant combinations were produced (in the format of Table 5.1). These lists were sent to Dr Mark O'Neill (MAO) to be the inputs for six networks. Each insect or plant genus became a node and each interaction became an edge in the network.

Insect	Plant
Apis	*Acacia*
Apis	*Barleria*
Apis	*Echiochilon*
Delta	*Kleinia*
Xylocopa	*Acacia*
Xylocopa	*Commelina*

Table 5.1 – The insect-plant combinations were listed in two columns for entry into the network software.

The six interaction lists sent to MAO as network inputs were:

1. *bfw1* Bee, fly and wasp loads, connections made from any amount of pollen.
2. *bs1* Bee and hoverfly loads, connections made from any amount of pollen.
3. *bs10* Bee and hoverfly loads, 10%+ of at least one load.
4. *bs25* Bee and hoverfly loads, 25%+ of at least one load.
5. *bs50* Bee and hoverfly loads, 50%+ of at least one load.
6. *bs75* Bee and hoverfly loads, 75%+ of at least one load.

MAO was requested to process these lists to produce three sets of outputs: visualised networks, node by-node statistical tables and combination removal tables. I then analysed these outputs to produce the results of section 5.3 (the collaboration sequence is summarised in table 5.2).

The visualised networks were two-dimensional representations of pollination interactions in the form of jpeg files (e.g. Fig. 5.5). These networks had no labels to indicate which node was for which insect / plant. In the smallest networks the labels could be determined by comparison with the interaction lists so they were added manually (to bs50 and bs75). The bee and hoverfly threshold networks (2 – 6) were progressively smaller with each increase in threshold. It is easier to interpret the structure of such related networks when comparable structural regions are on the same side of each network. This was achieved by reflecting networks about the y-axis.

The node-by-node tables assessed the connectivity of each node and the network damage caused by single node removals. Each plant or insect genus was listed with four statistical measures; two of these,

'hub size' (number of connections to that hub) and 'damage' (% of the network that would collapse if that node were removed, i.e. the ripple effects of an extinction), were chosen as the basis for discussion. The insects and plants were organised into separate tables and these tables ordered by declining hub size or damage. These data were tabulated for discussion and the hub size data of the largest network (bfw1) plotted for comparison with Fig. 2a of Memmott *et al.* (2004) (included as Fig. 5.3).

The largest network of bees and hoverflies (bs1) was selected for the removal of combinations of nodes. MAO removed increasing proportions of the pollinator genera without replacement, such that they included 1 – 12 of the 13 insect genera. The node combinations were selected using three different algorithms: random, least damaging combination and most damaging combination, making use of Smasher, a commercially available genetic algorithm. This differs from the method of Memmott *et al.* (2004) where nodes were removed at random, least connected (i.e. smallest hub size) first or largest hub size first, without consideration of combined effects. Percentage insect removal was then plotted against % network damage for comparison with Fig. 1b of Memmott *et al.* (2004) (included as Fig. 5.4).

Table 5.2 – The collaborative sequence between the researcher (ATW) and Dr Mark O'Neill (MAO). ATW provided six lists of interactions, the input for MAO's network algorithms, the outputs of which were analysed by ATW.

Time →

ATW	MAO	ATW
six interaction lists	visualised networks	* manually added data labels to bs50 & bs75 * reflected bs50 so that similarity to bs75 clearer
	node-by-node statistical tables	* plants nodes and insect nodes separated * selected hub size and connective damage as metrics * ordered according to hub size and damage * tabulated for discussion (tables 5.3 and 5.4) * graphed for comparison with Memmott *et al.* (2004)
	combination removals - random - least damaging - most damaging	* graphed for comparison with Memmott et al. (2004)

5.3 Results

5.3.1 All common *Acacia* visiting insects

The network including all common *Acacia*-visiting insects, irrespective of the amount of pollen they carried, is shown as Fig. 5.5. Nine of the insects and eight of the pollens had over 10 connections so it is unsurprising that the network is visually very complex. The generalist pollinators that carried pollen from many different plant genera are located centrally. The genera that carried a small range of plant pollens are on the periphery; it is important to realise that these may be specialist insects, but are more likely to be generalists that visit flowers only rarely for an occasional input of nectar

Rather more informative than the visualisations are the statistical outputs. These outputs take the form of data tables in which each node (plant or insect genus) has its connectivity quantified. The two most useful output metrics were 'hub size' and 'damage'. Hub size is the number of connections a single node makes, i.e. how many insects visit that plant or how many plants that insect visits. This is the metric used in Memmott *et al.* (2004). It is useful because it directly relates to pollination behaviour. Damage is the percentage of network links that would be lost if that node were removed, i.e. it quantifies the ripple effects of an extinction. Hub size and damage often differ in the genera they determine as important, so they are useful in combination.

Relatively few loads from relatively few places and times have been analysed, so any interpretation of network outputs can only be preliminary suggestions for further investigation. When all common visitors are included in the analysis (network bfw1 and Table 5.3) the insect genera that carried the greatest numbers of most pollen genera were *Megachile* (23), *Apis* (20), *Amegilla* (17), *Macrogalea* (15) and *Lipotriches* (15) (Table 5.3). These bee genera, representing all three bee families included in the analysis, have been observed to frequently collect pollen from flowers. They are medium to large bees with scopa in which to carry pollen. All but *Lipotriches* are long-tongued, enabling them to feed from flowers with longer corollas.

The insect genera that carried the fewest pollens were *Tetralonia* (3), *Ceriana* (3), *Hemipyrellia* (3), *Dejeania* (3) and *Sphex* (2). *Tetralonia* is an apid bee with scopa and *Ceriana* is a hoverfly (the hoverflies are generally considered to be important flower visitors (e.g. Sutherland *et al.,* 1999)); both of these genera were under sampled (only three and six loads analysed respectively) and this under sampling may have been a reason for such poor connectivity. *Hemipyrellia* (Calliphoridae) and *Dejeania* (Tachinidae) are generalist flies, *Dejeania* has long spikes but neither genus has many hairs

on which to transport pollen. *Sphex* is a sphecid wasp, a predatory insect that is thought to visit a single flower occasionally to drink nectar for quick-release energy.

The removal of no single insect genus would cause substantial damage to the network. *Apis* is the most important, with 7 % damage (Table 5.3). *Megachile* had more network connections than *Apis* but its removal only caused 3 % damage, grouping it with all the other genera (2 – 3 % damage). This is probably because *Apis* was the only insect to carry pollen of four plant genera (*Anthericum, Crotolaria, Kleinia* and *Sida*), while *Megachile* was the unique carrier of only *Sphaeranthus.* Thus, if *Apis* were to be removed (as suggested by the *Apis mellifera* decline of Watanabe, 1994) the network suggests that four plants may be left without pollination, while a similar removal of *Megachile* would only leave one genus unpollinated and this would contribute to make the damage of *Apis* removal greater.

The distribution of pollen genera per insect genera, comparable to Fig. 2a in Memmott *et al.* (2004) (Fig. 5.3), is not strongly concave (Fig. 5.6). No insects carried just a single pollen genus and the maximal number of insect genera for any one pollen number is only four. Three-quarters of insects carry fewer than twelve pollen genera but the fine-scale pattern appears quite random.

Fig. 5.5 – Visualised pollen load network for all common *Acacia* visiting insects, irrespective of the amount of pollen they carried. This includes the wasps and the bombyliid, calliphorid and tachinid flies, as well as the bees and syrphids. Green on black is the only format currently available.

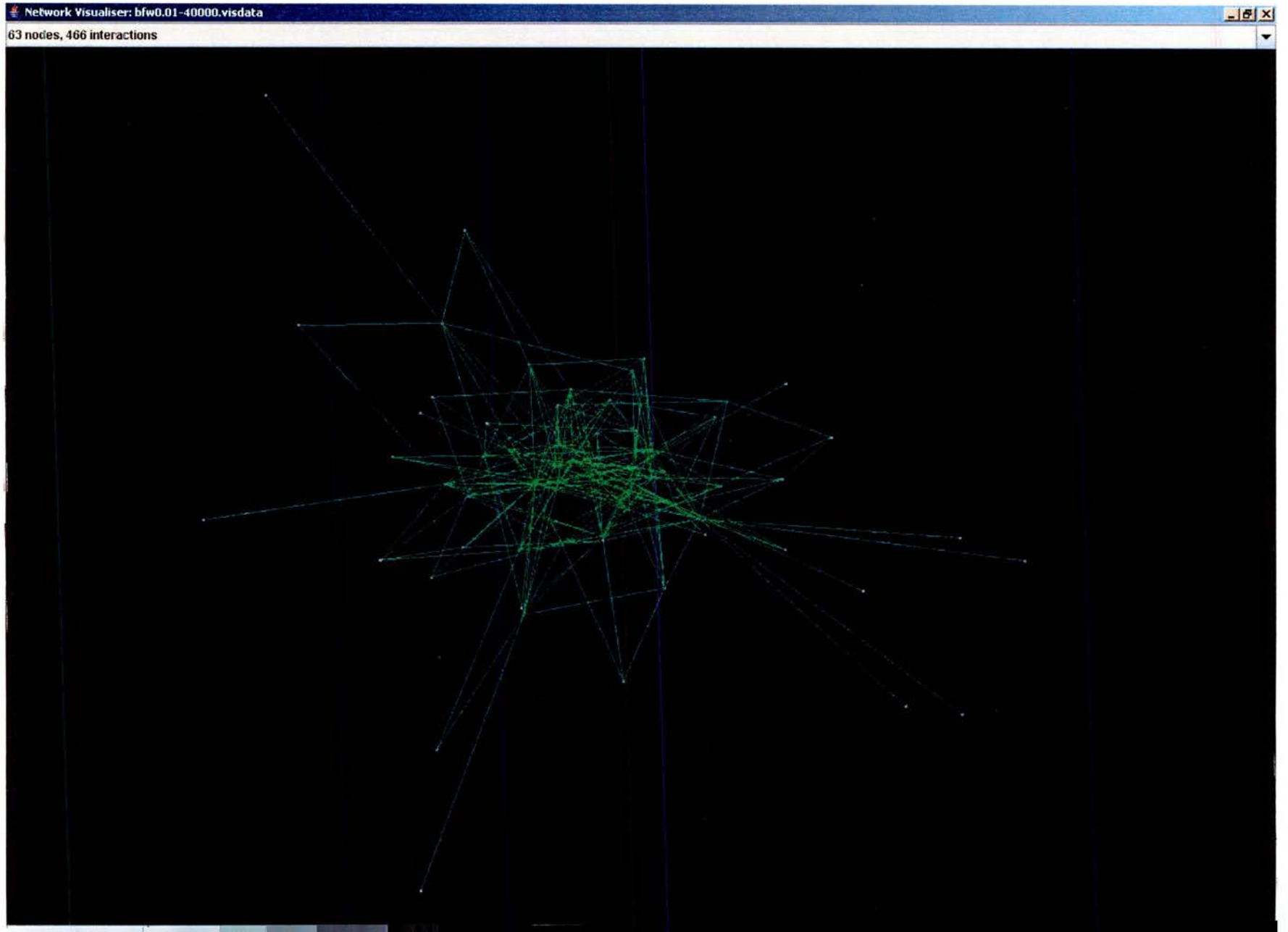
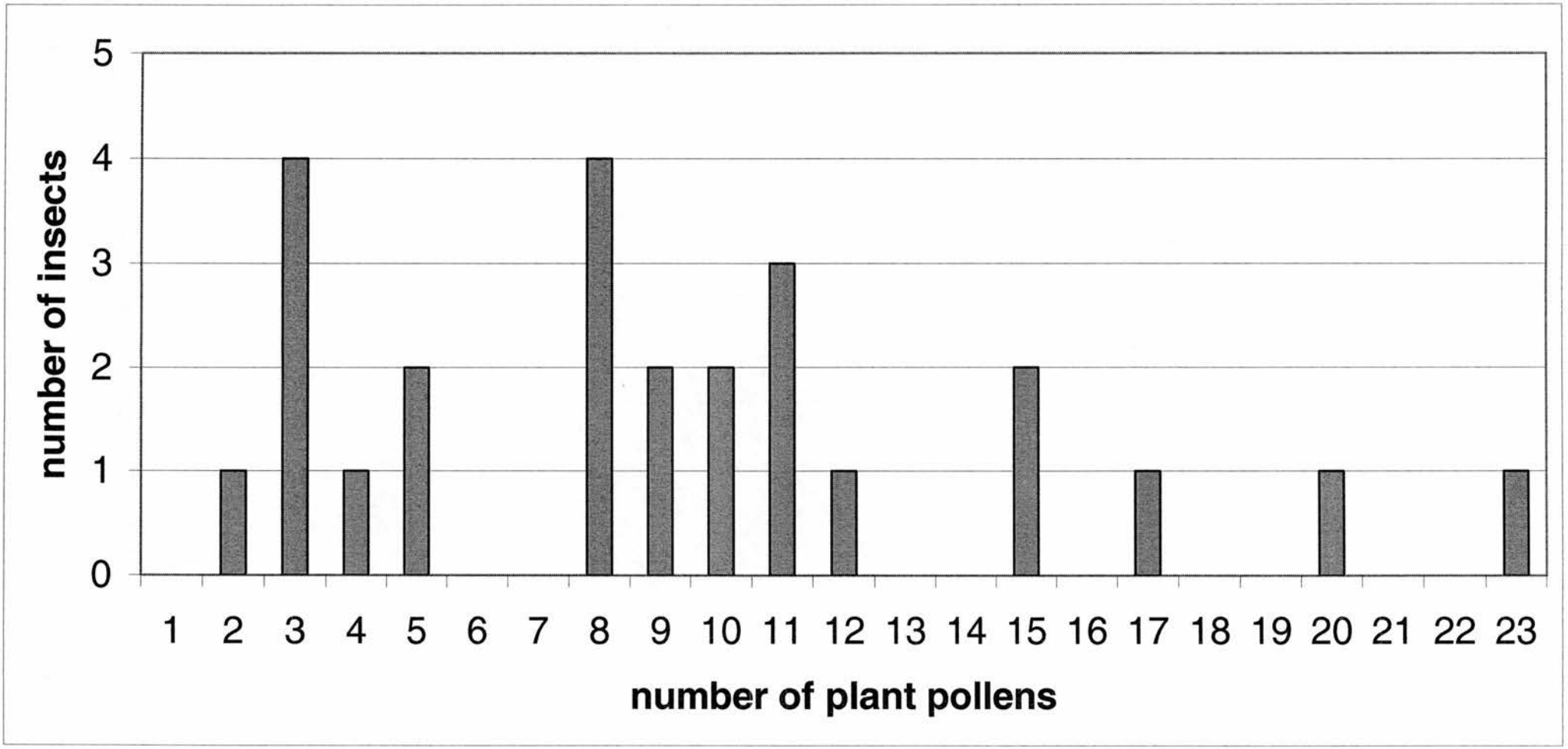


Table 5.3 – Hub size and connective damage (%) for all the common *Acacia* visiting insects. They have been ordered according to decreasing hub size. The colours in the damage column highlight the different damage values.

Insect	Hubsize	Damage (%)
Megachile	23	3
Apis	20	7
Amegilla	17	2
Macrogalea	15	3
Lipotriches	15	2
Heriades	12	2
Rhyncomya	11	3
Bombylius	11	2
Phytomia	11	2
Delta	10	3
Eristalinus	10	2
Xylocopa	9	3
Patellapis	9	2
Rhinia	8	3
Chrysomya	8	2
Bembix	8	2
Pseudapis	8	2
Philanthus	5	2
Plebeina	5	2
Stegosoma	4	2
Tetralonia	3	2
Hemipyrellia	3	2
Ceriana	3	2
Dejeania	3	2
Sphex	2	2

Fig. 5.6 - The distribution of pollen genera per insect genera, comparable to Fig. 2a in Memmott *et al.* (2004) (Fig. 5.3), is not strongly concave. No insects carry just a single pollen genus but ¾ carry less than 12.



The pollens carried by the most insect genera were *Acacia* (23), *Leucas* (18), *Ocimum* (17) and *Helichrysum* (14) (Table 5.4) in the families Fabaceae, Lamiaceae and Asteraceae (Fig. 5.7). Most of the catching was done around flowering acacias so it is unsurprising that *Acacia* is the most carried genus. These plants were observed to flower extensively for long time periods. They have cream / white flowers (or yellow in *Acacia nilotica* and *Acacia seyal*). *Acacia* and *Helichrysum* have small actinomorphic flowers, presented in compound inflorescences, which could be probed by short-tongued insects. *Leucas* and *Ocimum,* however, are less morphologically "generalist", with zygomorphic flowers that would only provide resources to moderate-long tongued insects. *Leucas* (24 – 28 µm) and *Helichrysum* (15 – 29 µm) have very small pollen and the pollen of *Acacia* (38 – 57 µm) and *Ocimum* (59 – 70 µm) is moderate in diameter (Fig. 5.7). A pollen size limitation to insect transport may be present, as the largest of the pollens (*Abutilon*) was only ever carried by largest of the insects (*Xylocopa*).

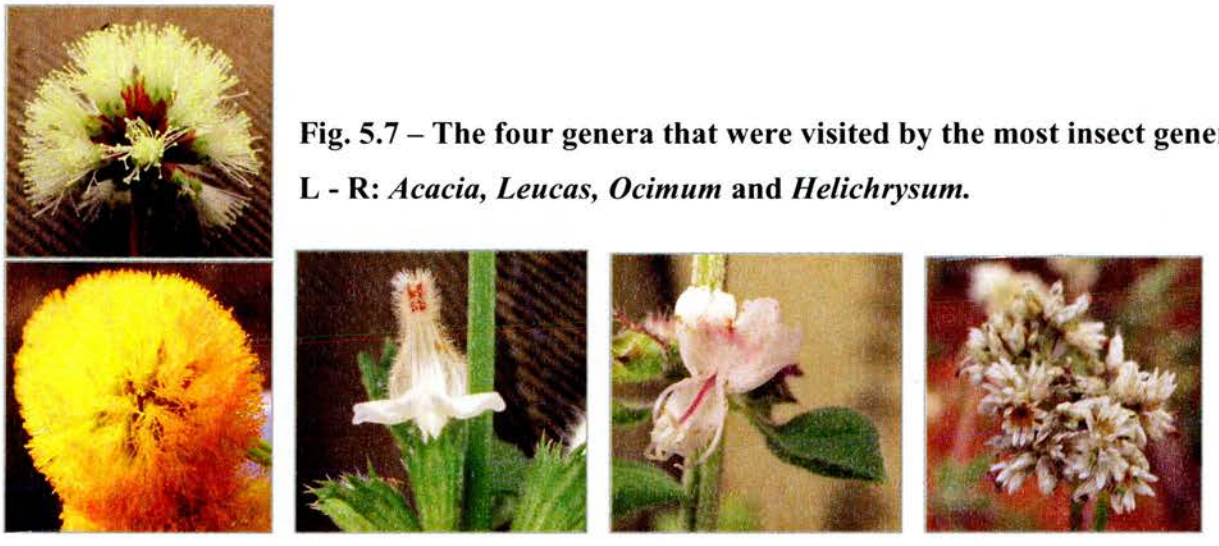


Fig. 5.7 – The four genera that were visited by the most insect genera. L - R: *Acacia, Leucas, Ocimum* and *Helichrysum.*

Ten pollen genera were carried by only a single insect genus; they were carried by *Apis* (*Anthericum, Crotolaria, Kleinia, Sida*), *Delta* (*Tribulis*), *Megachile* (*Sphaeranthus*), *Macrogalea* (*Commicarpus*), *Rhinia* (*Becium*), *Rhyncomya* (*Cynodon*) and *Xylocopa* (*Abutilon*) (Table 5.4). As all of these insects forage on at least eight plant genera they seem unlikely to go extinct so these specialist plants are unlikely to be at great risk of secondary extinction. The pollens all caused just 2% damage if singly removed from the network. No insect carried just a single pollen genus and those carrying two or three pollen genera tended to forage on well-connected plants, e.g. *Sphex* carried *Acacia* (23 pollens) and *Heliotropium* (8). The blowflies and wasps are rarely given serious consideration as pollinators, as they carry little pollen on their bodies, but these pollen load data suggest that *Becium, Cynodon* and *Tribulis* would be lost from the network without *Rhinia, Rhyncomya* and *Delta* (although *Apis* was also observed to collect pollen from *Cynodon*).

The distribution of insect genera per pollen genus is weakly concave (Fig. 5.8), and this is better seen when means of ranges are shown in Fig. 5.9. It is closer to the strongly concave distribution of Memmott *et al.* (2004) than was the case for the distribution of plant genera per insect genera (Fig, 5.6).

Plant	Hub size	Damage (%)
Acacia	23	2
Leucas	18	2
Ocimum	17	2
Helichrysum	14	2
Carissa	13	2
Gynandropsis	12	2
Commelina	12	2
Gutenbergia	12	2
Justicia	10	2
Ipomoea	9	2
Plectranthus	9	2
Heliotropium	8	2
Melhania	7	2
Grewia	7	2
Pentanisia	6	2
Aloe	5	2
Hibiscus	5	2
Gloriosa	5	2
Kalanchoe	5	2
Lippia	4	2
Achyranthes	3	2
Chascanum	3	2
Athroisma	3	2
Indigofera	3	2
Composite	3	2
Barleria	3	2
Erogrostis	2	2
Pavonia	2	2
Abutilon	1	2
Anthericum	1	2
Becium	1	2
Commicarpus	1	2
Crotolaria	1	2
Cynodon	1	2
Kleinia	1	2
Sida	1	2
Sphaeranthus	1	2
Tribulis	1	2

Table 5.4 – Hub size and connective damage (%) for pollens carried by all the common *Acacia* visiting insects.

Fig. 5.8 - The distribution of insect genera per pollen genera is weakly concave.

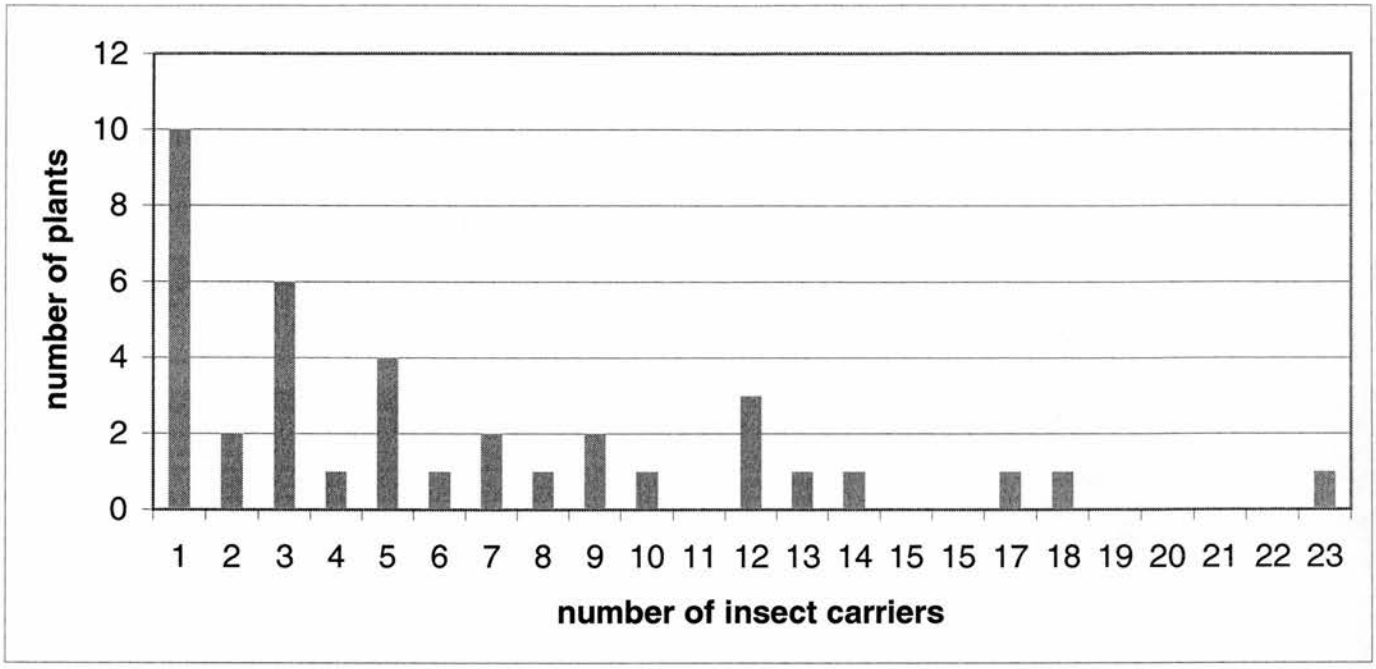
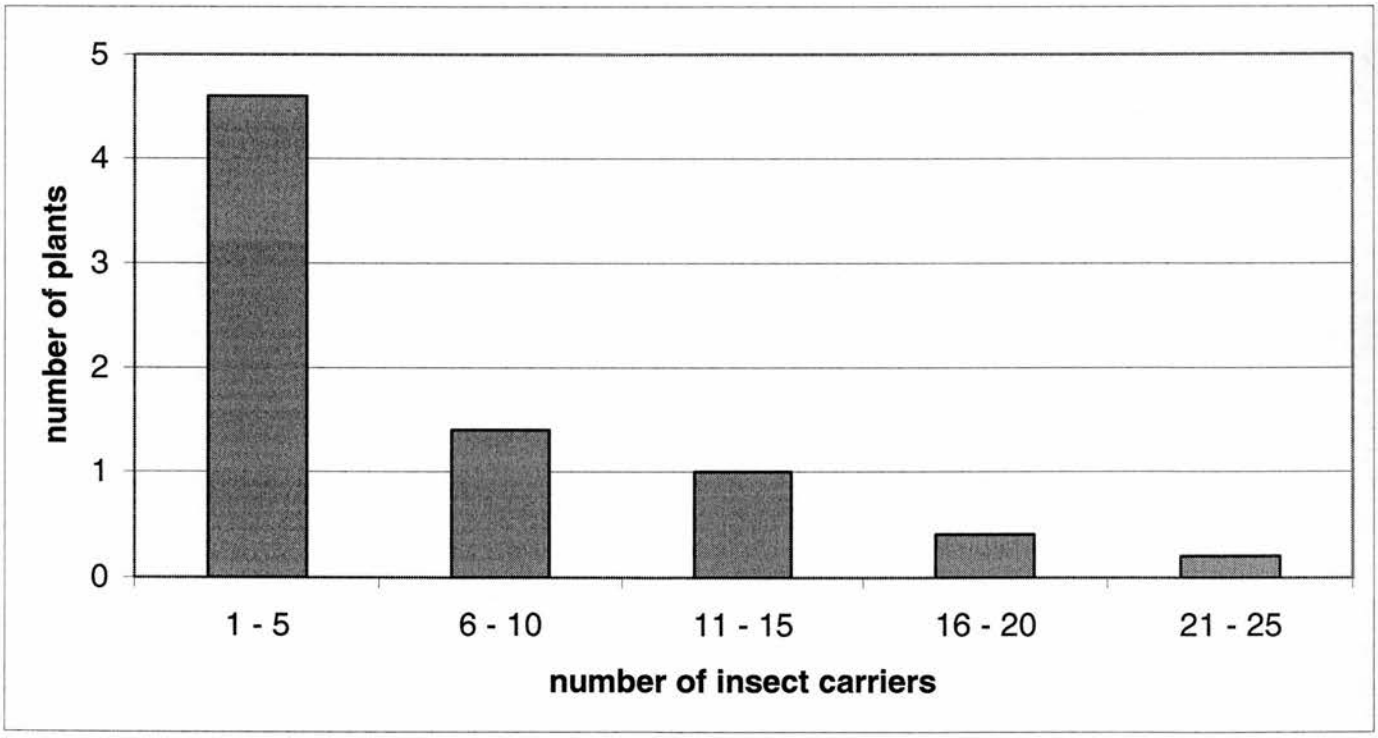


Fig. 5.9 – The concave distribution of insect genera per pollen genera can be seen more clearly when mean values are calculated for 1-5, 6-10, 11-15, 16-20 and 21-25.



5.3.2 Bees and hoverflies with fidelity inclusion thresholds

The networks of bees and hoverflies carrying pollens over 10%, 25%, 50% and 75% of the load are less complex at higher thresholds because fewer interactions are included (Fig 5.10a-d). These networks are nested (see Bascompte *et al.* (2003) for discussion of nested plant-animal networks), with the same core genera present in all four and less dominant genera dropping away each time the threshold is raised. The visualisation for over 50% has been mirrored about the y-axis so that the similarity of the cluster on the right to the 75%+ cluster is easier to see; this cluster is centred on *Acacia.*

The basic visualisations are difficult to analyse by eye, as the nodes are too densely clustered to be legibly labelled with generic names. It should soon be possible to add labels automatically (where they would be useful), once the software is developed further (O'Neill, pers. comm.). Labels have been manually added to the smallest two networks, those representing pollens comprising at least 50% and 75% of total load, in Fig. 5.11. In both networks there is a group of insects that forage primarily on *Acacia* (at least some of the time). This is inevitable, as insects were mainly caught around flowering *Acacia.* Only the 50%+ network also includes foraging at *Leucas, Ocimum* and *Helichrysum.* While some of the loads did have these pollen genera at 75%+ there was no direct connection (*Phytomia* in the 50%+ network) from them to *Acacia* so they have been omitted from the 75%+ network.

Fig. 5.10a – Visualised pollen load networks of bees and syrphids in which connections represent 10%+ of total load.

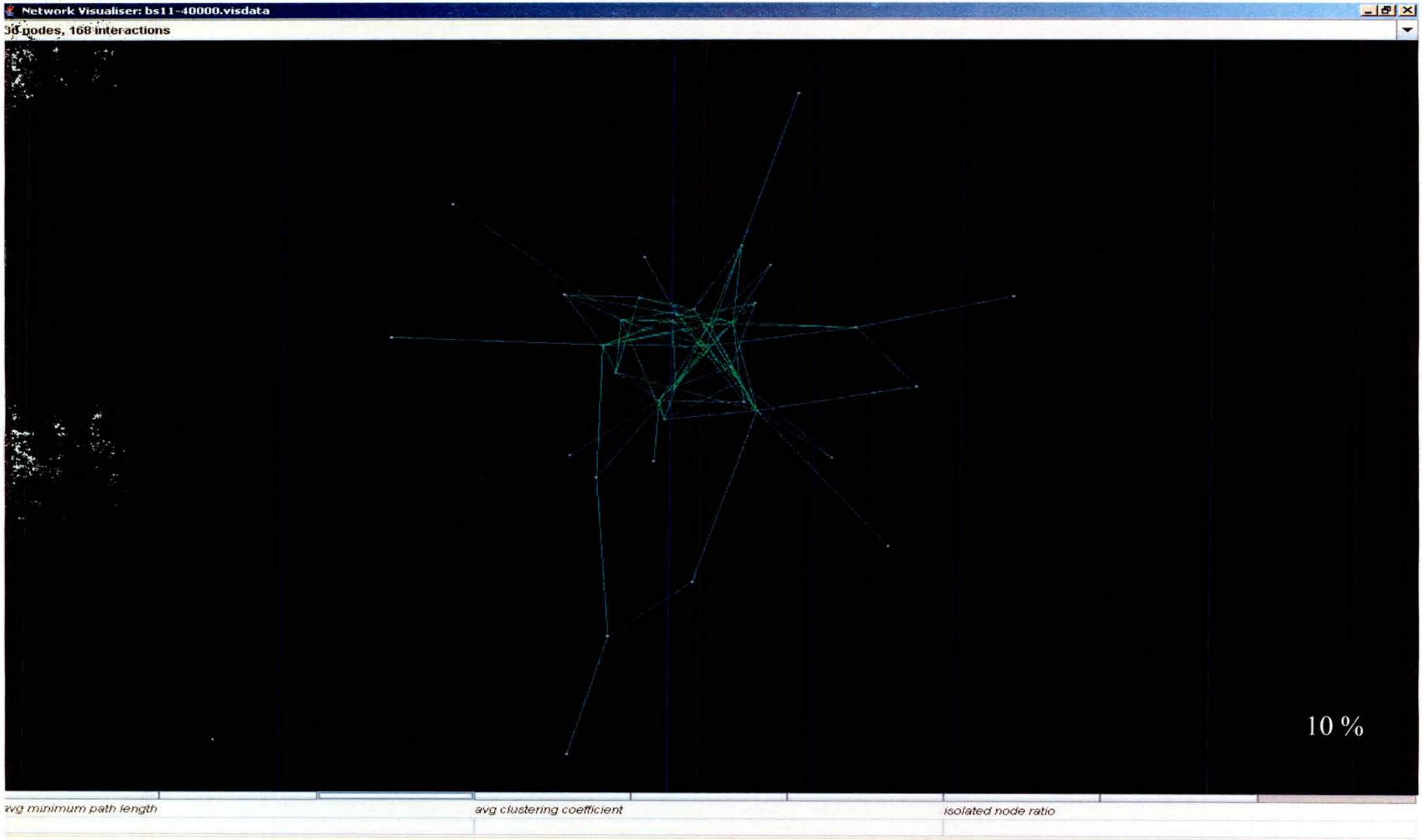


Fig. 5.10b – Visualised pollen load networks of bees and syrphids in which connections represent 25%+ of total load.

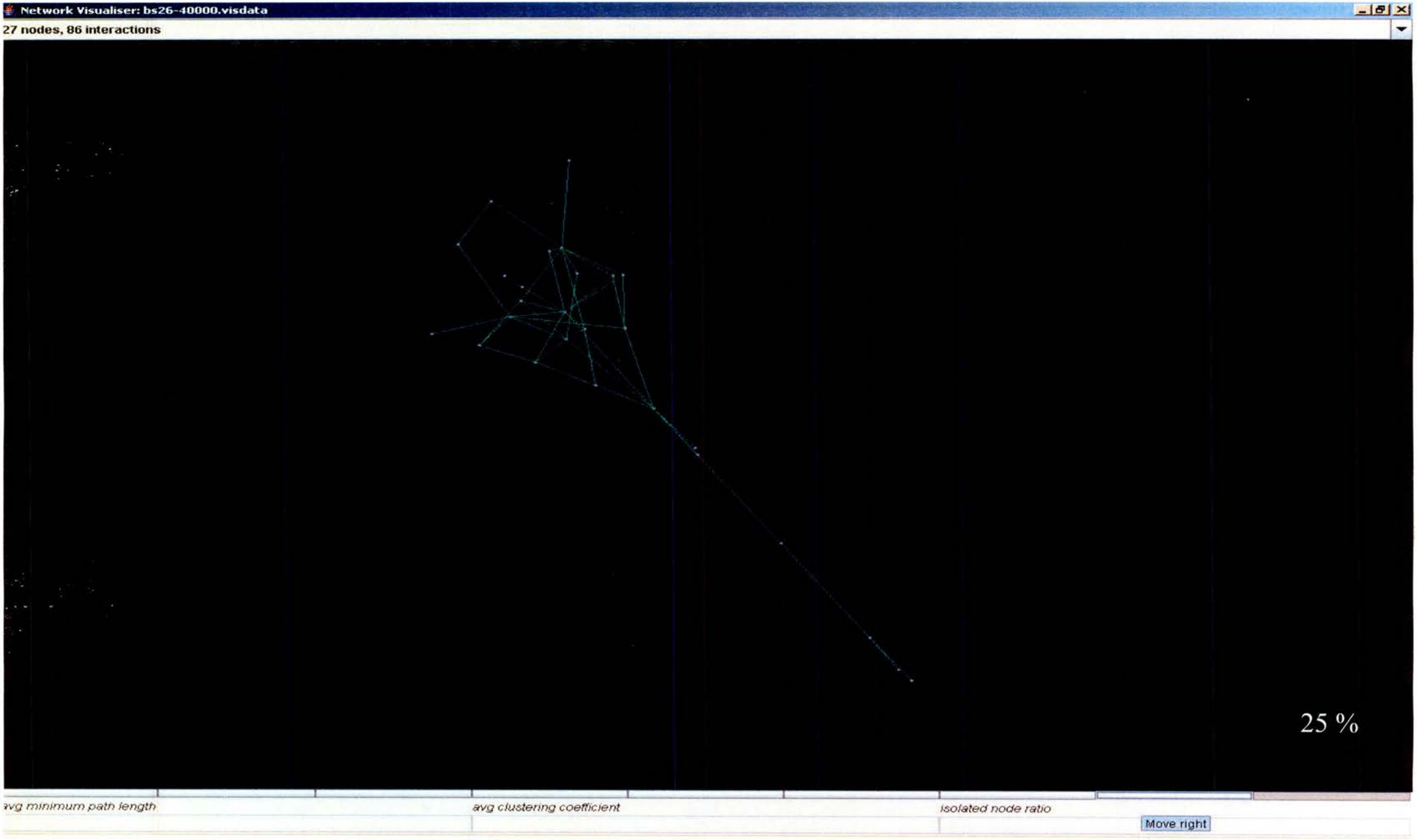


Fig. 5.10c – Visualised pollen load networks of bees and syrphids in which connections represent 50%+ of total load.

Fig. 5.10d – Visualised pollen load networks of bees and syrphids in which connections represent 75%+ of total load.

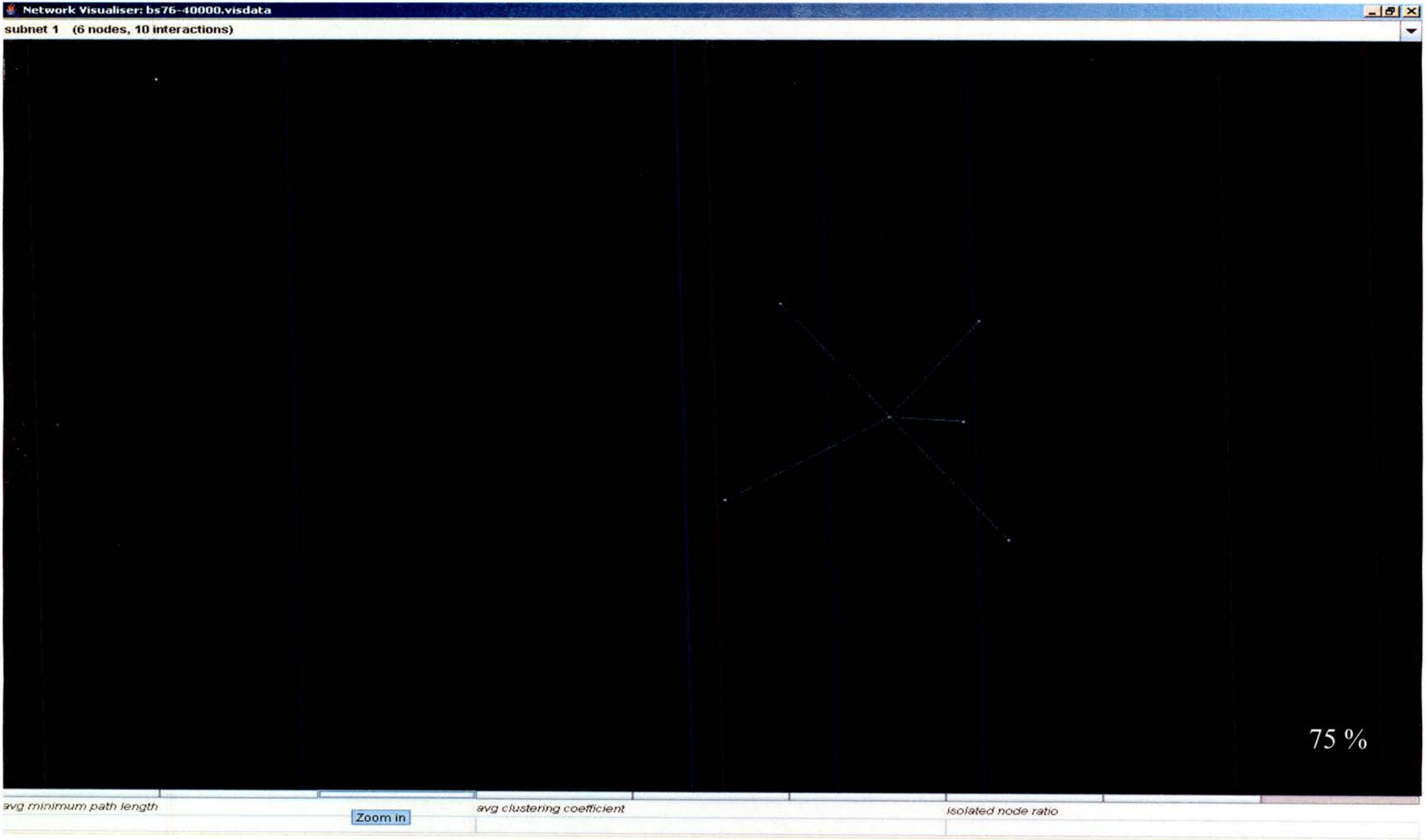
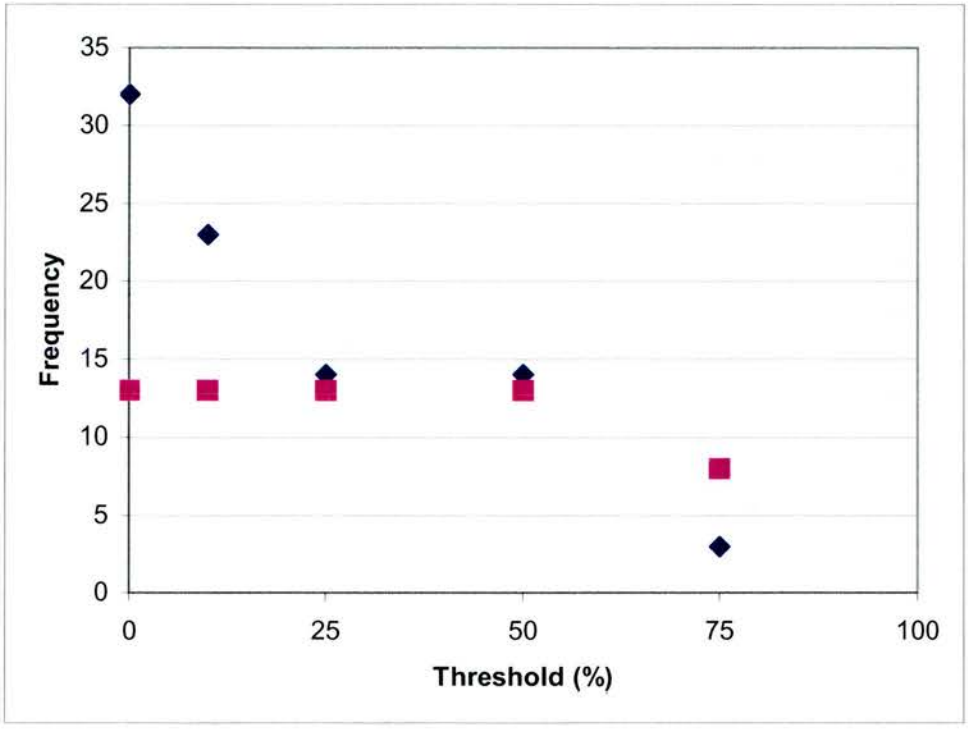


Fig. 5.11 – Pollen networks for pollens dominating bee and syrphid pollen loads with labels manually added: 50%+ (top) and 75%+ (bottom). Yellow signifies a plant and white an insect.

When no load threshold is set, 13 bee and hoverfly genera collect 32 pollen genera (Fig. 5.12). As the threshold is increased to 25% more than half the plants are no longer represented. The number of plant

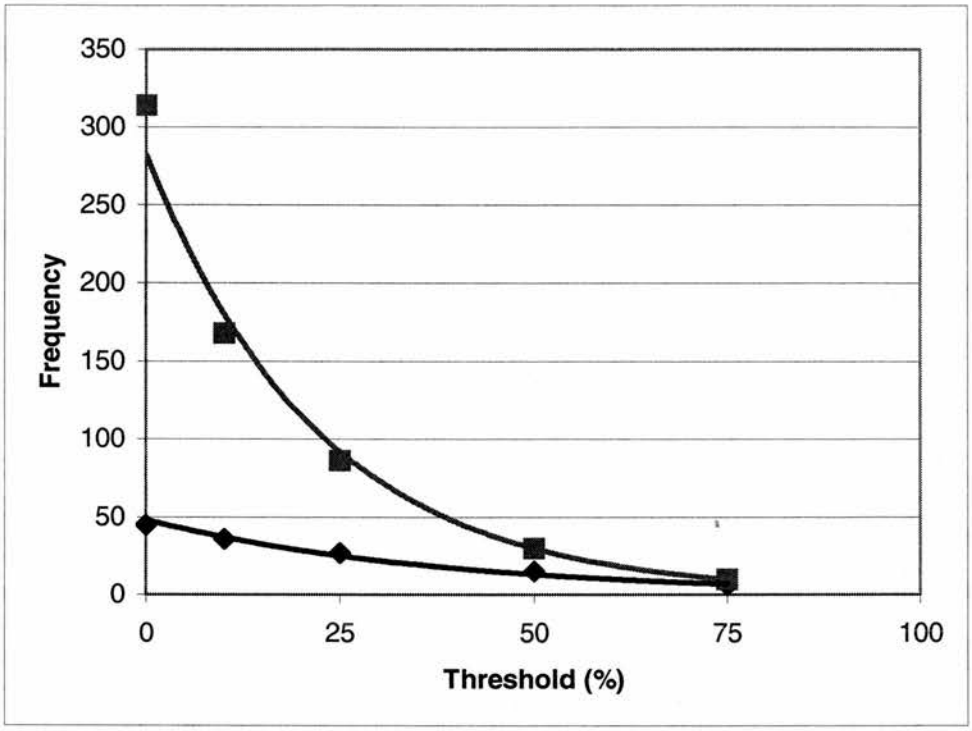
pollens falls linearly but the insect number remains unchanged, as all 13 insect genera carry at least one pollen genus at 25%+. The same plants and insects also satisfy the 50% threshold (i.e. the dominant pollen in the load). As the threshold is raised from 50% to 75% five further insect genera are excluded leaving a final eight. Nevertheless, there are twice as many insects involved as pollens. Only three pollen genera achieve 75%+ load domination. When 75%+ domination is enforced, only *Acacia* (carried by *Eristalinus, Phytomia, Macrogalea, Plebeina* and *Pseudapis*), *Ocimum* (carried by *Lipotriches* and *Patellapis*) and *Leucas* (carried by *Lipotriches* and *Xylocopa*) remain.

Fig. 5.12 – As the load composition for inclusion increases the number of plants and insects included decreases. Plants are shown as blue diamonds and insects are shown as pink squares.



Another way to analyse the importance of load threshold is to count the network nodes and connections that are linked to *Acacia*. As the threshold is increased from 0% to 75% both the number of *Acacia*-linked nodes and the number of *Acacia*-linked connections declines exponentially ($R^2 > 0.98$) (Fig. 5.13). In practical terms this means that increasing the composition threshold from 0% to 10% decreased the complexity of the network much more than increasing the threshold from 50% to 75% (even though the threshold change is much smaller). Many plants form part of the mixed loads of generalists but very few are sufficiently attractive and common to dominate loads.

Fig. 5.13 – As the load composition for inclusion increases the number of *Acacia*-linked nodes (diamonds) declines exponentially ($R^2 = 0.9825$), with just 13% of nodes remaining by the 75% threshold. The number of *Acacia*-linked interactions (squares) also declines exponentially ($R^2 = 0.9972$). This decline is more dramatic and only 3% of interactions remain by the 75% threshold.



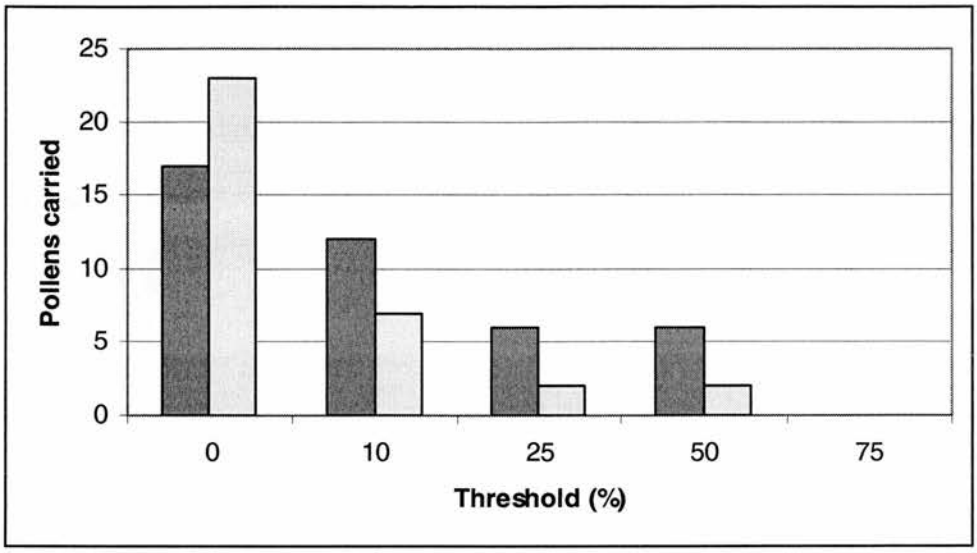
When hub size is assessed with load threshold for each insect genus, different insect genera appear to be most important at different thresholds. In Table 5.5, the most important genus at each threshold is highlighted in yellow and other important genera in green. When a single grain of pollen is enough to produce a connection the most connected insect genus is *Megachile* (23), followed by *Apis* (20); but once a pollen has to comprise at least 10% of a load these genera lose their importance, being replaced by *Amegilla* (12), *Patellapis* (9) and *Lipotriches* (9). *Patellapis* (7) and *Amegilla* (6) continue to be the most connected genera at the 25%+ and 50%+ thresholds. Only *Lipotriches* (2) carries more than one pollen genus at 75%+ of a load.

Table 5.5 – The number of plant pollens carried by bee and hoverfly genera at different minimum percentages a load. Most important genus pale, other important genera darker.

Insect	Hub size				
	0%+	10%+	25%+	50%+	75%+
Amegilla	17	12	6	6	
Apis	20	5	2	2	
Eristalinus	10	5	3	3	1
Heriades	12	8	4	4	
Lipotriches	15	9	4	4	2
Macrogalea	15	8	2	2	1
Megachile	23	7	2	2	
Patellapis	9	9	7	7	1
Phytomia	11	5	3	3	1
Plebeina	5	2	1	1	1
Pseudapis	8	5	4	4	1
Tetralonia	3	3	3	3	
Xylocopa	9	6	2	2	1

The number of pollens carried by each insect declines as the threshold increases (Table 5.5) but the gradient of this decline varies between insect genera. This could be a response to different foraging strategies. The genera that visit many different plants (often only collecting a small part of their load from that plant genus) tend to lose more than 50% hub size between 0 and 10% thresholds (*Megachile* and *Apis*). Those that are more focused ('flower constant') in their foraging, collecting fewer pollens in small amounts, tend to lose less than 50% hub size (*Amegilla, Heriades, Lipotriches, Patellapis, Pseudapis, Tetralonia* and *Xylocopa*). These possible strategies are illustrated by *Megachile* and *Amegilla* in Fig, 5.14. Seventy percent of the pollen genera that *Megachile* collects make up less than 10% of a load. As the threshold increases further the decline in pollen number is gentler. In *Amegilla*, only 29% of pollens make up less than 10% of a load so 71% of pollen genera are collected with greater fidelity.

Fig. 5.14 – *Megachile* (unshaded) and *Amegilla* (shaded) may be examples of different pollen collection strategies. In *Megachile* the initial decline with threshold is steep as most pollen genera comprise less than 10% of a load. In *Amegilla* the initial decline is more gradual as most pollens comprise more than 10% of a load.

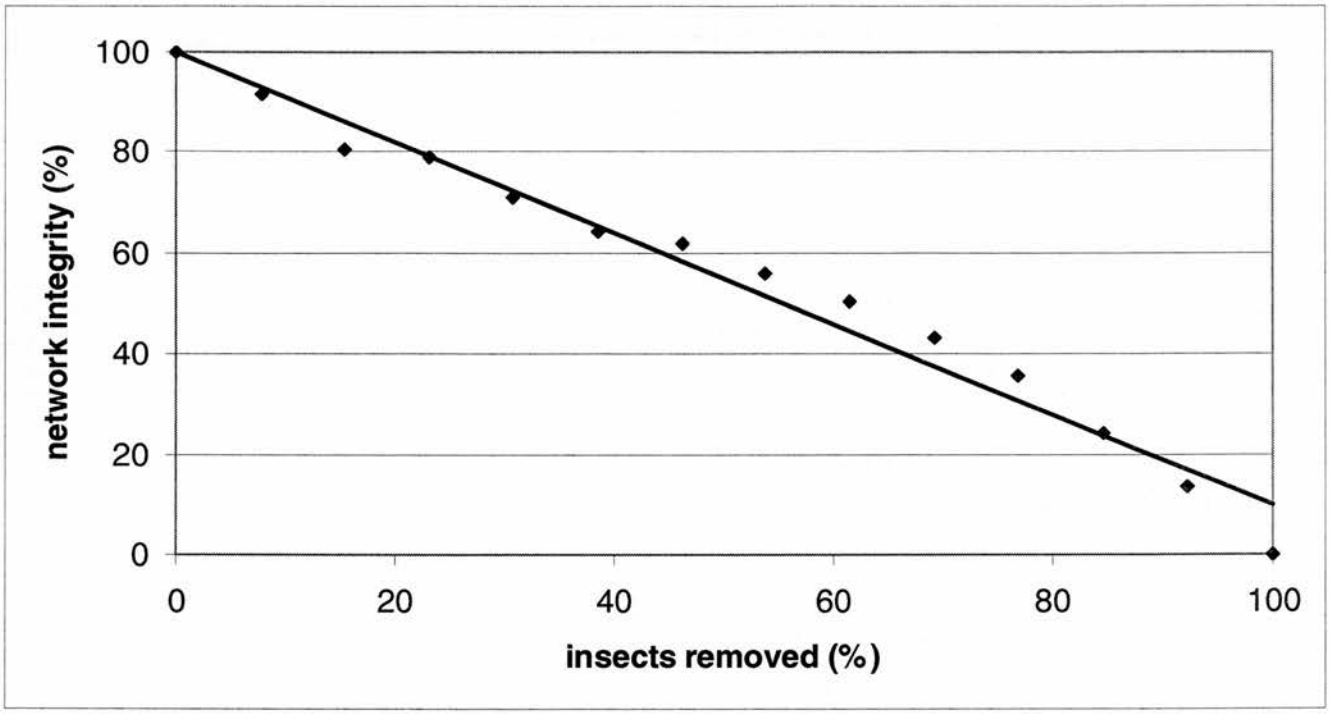


5.3.3 Removal of pollinator combinations

If nodes are not discriminated into plants and insects and the most vital node combinations are removed, then these node combinations are composed entirely of insects. This is unsurprising as the insects *Apis, Megachile, Macrogalea, Xylocopa* and *Phytomia* were the only nodes to cause more than 2% damage when removed individually, and once these insects had been removed then the other insects would become increasingly important. *Acacia* pollen was present in the loads of 23 insect genera but all of these also carried other pollens.

When node combinations for removal are selected at random the rate of network damage could be described as linear ($R^2 = 0.9762$) with a gradient close to –1 (Fig. 5.15). As insects for removal are selected at random and means calculated, the highly damaging removals could be cancelled out by the removals that cause little damage so all insect removals could be considered to cause moderate additional damage. This is different to the gently convex curve reported by Memmott *et al.* (2004). However, this trend also suggests a shallow sigmoidal curve, where the first few and last few insect species to be removed have more impact on network integrity than the others (producing a steeper curve).

Fig. 5.15 – The reduction in network integrity when random combinations of insects are removed could be described as linear ($R^2 = 0.9762$) with gradient close to –1, or shallowly sigmoidal.



The small combinations of insects that were most damaging to remove were:

Number	Genera	Damage (%)
1	*Apis*	17
2	*Apis* and *Lipotriches*	32
3	*Apis, Lipotriches* and *Megachile*	38
4	*Apis, Lipotriches, Megachile* and *Macrogalea*	46

These are short and long-tongued bees from all of the three bee families included. They are all genera that carry many different pollen genera (Table 5.6).

Table 5.6 – The total number of connections, number of unique collections and number of connections with prior removals for the four most damaging insect genera.

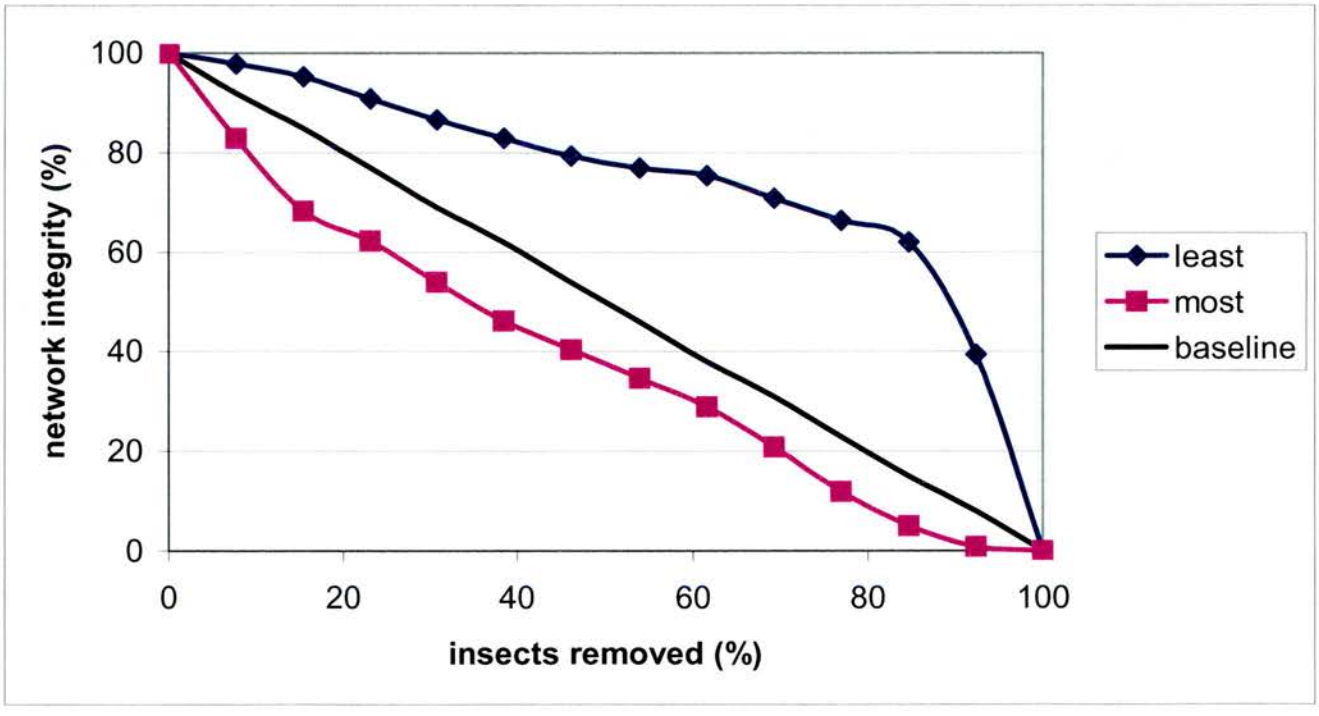
Insect	Hub size	Unique	Unique with prior removals
Apis	20	4	4
Lipotriches	15	0	0
Megachile	23	2	2
Macrogalea	15	1	2

Apis, Megachile and *Macrogalea* make intuitive sense, as they are the unique carriers of pollen genera (Table 5.6). Removing *Macrogalea* alone would cause one pollen genera to not be carried

(*Commicarpus*). However, if *Apis* and *Macrogalea* have already been removed it will also be the only carrier of *Barleria.* The selection of *Lipotriches* is harder to understand as it is the unique carrier of no pollen genera.

When the number of insects included in the most damaging combination increases, network integrity decreases (shown in pink on Fig 5.16). This decrease has a good linear fit ($R^2 = 0.9748$) with gradient close to –1. However, when compared to a linear baseline it appears that the first two removals cause more damage than the others (with steeper gradient of integrity loss) and the final two cause less damage than the others (with more shallow gradient of integrity loss). This subtle departure from an inversely proportional relationship may be because the insects making up the smallest removal groups will be the most damaging ones, causing a large initial impact. By the time the final insects are being added to the largest removal groups the only insects remaining are the least damaging genera that make little difference to this remnant network.

Fig. 5.16 – Reduction in network integrity when combinations of insects are removed.



The small combinations of insects that were least damaging to remove were:

Number	Genera	Damage (%)
1	*Eristalinus*	2
2	*Eristalinus* and *Tetralonia*	5
3	*Eristalinus, Tetralonia* and *Amegilla*	9
4	*Eristalinus, Tetralonia, Amegilla* and *Pseudapis*	13

Eristalinus is a hoverfly. *Tetralonia* and *Amegilla* are large, long-tongued apid bees and *Pseudapis* is a short-tongued halictid bee.

For the least damaging removal combinations two forces act in concert, to create a convex curve (shown in blue on Fig. 5.16). The first genera to be removed are the specialists or infrequent flower visitors (whose plants are also visited by generalists), thus the initial decline is relatively gentle. The loss of the first half of the insect genera causes only a 25% reduction in network integrity. As the final two insect genera are added into removal combinations a critical point is crossed and a sudden shift occurs, similar to the final catastrophic decline described in Memmott *et al.* (2004). These extreme generalists are both the most damaging in their own right and the final remaining insects, so their removal causes network integrity to dramatically reduce.

5.4 Discussion

This discussion is arranged in three sections. The first covers the many limitations of these data. The second and third bear those limitations in mind to discuss the implications of single removals and combination removals.

5.4.1 Data limitations

Some of the basic caveats discussed in Memmott *et al.* (2004) are also appropriate to this study.

A) Simulated removal of insects assumes that all pollen carriers are effective pollinators, so that a plant must lose all its visitors before its reproduction fails. This assumption was not supported by the data, as many insect genera (wasps and most of the flies) only ever carried very small amounts of pollen and would not have been able to support the pollination requirements of plant communities without the bees and hoverflies. Violation of this caveat means that the importance of some insects may have been overestimated while that of others was underestimated. However, this assumption was necessary because there is a lack of literature on the relative effectiveness of flower visitors (Memmott *et al.,* 2004).

B) It was assumed that all plants require pollination to reproduce but some plants can propagate asexually, through cloning or self-pollination. This assumption was considered reasonable because, in the long-term, sexual reproduction is important to avoid extinction (Holsinger, 2000).

C) It was assumed that reproduction is either a total success or a total failure; this is a massive simplification as a plant may reproduce poorly (with few individuals producing offspring) but still continue at lower density than before rather than going extinct.

D) Pollinators remaining after an insect extinction were assumed not to expand their floral diets, 'rescuing' plants that would otherwise have gone unpollinated (Kondoh, 2003). New links could potentially arise during extinctions, if pollinators are released from competitive exclusion (Memmott *et al.*, 2004).

Memmott *et al.* (2004) considered that despite the above assumptions the network approach provides a useful approximation of actual pollination systems. This study also has many limitations that did not apply to Memmott *et al.* (2004), and which provide one explanation for differences in results from Memmott *et al.* (2004).

a) Only 308 pollen loads were analysed so many network links will have gone undetected. Pollen load sampling needs to be much more extensive before any interpretations can be considered seriously.

b) Those loads that have been analysed in this study were not evenly spread between insect genera. Therefore, genera that were sampled more than most, such as *Apis* (107 loads), *Megachile* (55) and *Xylocopa* (24), would be expected to carry a great diversity of pollen genera, and those poorly sampled, such as *Tetralonia* (3) and *Patellapis* (7), to carry few. Even sampling is difficult to achieve, as the more specialist genera tend to be the rarer species (Vázquez & Aizen, 2003), and therefore to provide fewer pollen loads for analysis. Nevertheless, some data suggest that high load number may not be essential for a broad range of pollens to be encountered, e.g. *Patellapis* had only seven loads analysed yet still had more pollen genera than any other insect genus above the 25% and 50% load thresholds.

c) When insects deposit pollen on flowers some of this pollen may be from the incorrect species. If another insect then visits the flower it could pick up this contaminant pollen, so its load may contain pollen from a plant genus it has never visited. The largest network analysed (Fig. 5.5) has no threshold for pollen inclusion. A single grain was enough to form a connection so some of the connections may have been due to contaminant pollen. This problem can be largely overcome be setting a threshold at a percentage of total load. When a 10% threshold was added to the interactions of the 13 bee and hoverfly species in Table 5.5, the number of interactions halved. This was probably too high a threshold as some little visited plants may have been

excluded. 1 – 5% would be more appropriate. Once the very weak connections (possible contaminants) have been removed more insect genera will carry just a single pollen genus.

d) The pollen loads for this study were collected during only part of a single year, the period following the heavy rains (rains mainly come in April-May and November). Therefore, they provide only part of the annual set of pollination interactions, e.g. during the wet season the same insects may forage from a very different assortment of plants. The large visitation datasets of Clements & Long (1923) and Robertson (1929) analysed by Memmott (2004) were collected over 11 and 22 years respectively, providing many more interactions and a much fuller picture.

e) The large networks of Memmott *et al.* (2004) were resolved to species. This was straightforward as they were based on visitation observations rather than on pollen load analysis, and identifying adult plants to species-level is generally easier than pollen grains. In the present study it was decided to identify pollens and insects no more precisely than to genus level, as species-level identifications could not be guaranteed accurate. However, this approach results in a great oversimplification. Species are true biological units but genera are artificial human constructs. Species within a genus can differ substantially in their size, shape and interactions. This was illustrated by *Hibiscus cicatricosa* and 'the little *Hibiscus*' in Fig. 4.21, Section 4.3.4. It is here that DAISY identification would be especially valuable, if DAISY could be trained using herbaria pollen specimens, identified to species by experts, and developed to provide sufficient accuracy. I expect that had these data been resolved to species level there would have been many more plant and insect nodes and interactions, which may have provided a more accurate reflection of power law fit.

5.4.2 Single removals

One of the most distinctive features of this present study is the use of the node importance measure 'damage'. Damage is the percentage of network integrity that would be lost if a certain node were removed. This is more sophisticated than connectivity (or hub size), the measure used by Memmott *et al.* (2004) and takes fuller advantage of the ability to summarise many connections that makes network analysis useful. Hub size is the number of connections to other hubs but damage is also influenced by the hub size of those connecting hubs, i.e. the ripple effects of an extinction. Hub size and damage frequently conflicted over the most important genera for single removals. The insect genus with the greatest hub size in Table 5.3 is *Megachile* (23). Although *Apis* (20) has fewer network connections it caused more than twice the damage of *Megachile* because it was the unique carrier of four pollen genera, whereas *Megachile* was the unique carrier of only one. Removing individual nodes generally produced very little connective damage (2 - 3%). The most damaging node to remove was the

honeybee, *Apis*, which caused 7% damage (Table 5.3). This suggests that no one genus is essential to network integrity. This will be discussed more fully later, in reference to combination extinctions.

The threshold for inclusion in a pollen load network is very important as different thresholds tell very different stories. If no threshold is set then the network includes many species that may have negligible contribution to pollination and it appears at though the network is very tolerant to extinctions of genera. At this level *Apis* has the largest connective damage as it collects pollen from a wide range of flowers (20) and visits more unusual flowers than *Megachile.* The most important insect genera in combination are long-tongued megachilids and apid bees: *Megachile, Apis, Macrogalea* and *Xylocopa.* However, if a pollen type must dominate at least one load to be included in the network then the halictids, *Lipotriches* and *Patellapis,* become the most important pollen carriers and removing these two genera can substantially damage network integrity. It is important that such distinct foraging behaviours (also illustrated by Fig 5.14) are recognised in future pollination studies. In this study I examined intraload composition threshold, e.g. does a pollen make up 25%+ of at least one load? It would also be interesting to set thresholds on an interload basis, e.g. is a pollen carried by that insect in at least five loads (ideally from different times of day and parts of the season)? This would approach fidelity from a different angle and might produce interesting insights.

The blowflies (Calliphoridae) are rarely given serious consideration as pollinators, as they carry little pollen on their bodies. However, these pollen load data suggest that *Becium* and *Cynodon* would be lost from the largest network (Fig. 5.5) without *Rhinia* and *Rhyncomya. Rhyncomya*, a calliphorid, was observed to be a very common visitor to *Acacia* and was the third most sampled genus for pollen loads, with 53 loads analysed. While each insect may carry little pollen the huge numbers of individuals may compensate for this. The Calliphoridae and the Muscidae are closely related families, both in the Calypterae. The Muscidae have been described as important pollinators in arctic and sub arctic ecosystems where bees and butterflies are largely absent (Pont, 1993). They are generalist flower visitors, feeding on nectar and, in some species, on pollen gathered using the fore tarsi (Pont, 1993). These are only preliminary data, but if these insects do pollinate plants that bees rarely visit perhaps they should be given greater consideration in future pollination studies.

A network is connected according to the power law when most nodes in the network interact with just one other node, but a few interact with tens or hundreds of nodes (Proulx *et al.,* 2005), so a graph of number of nodes against number of interactions is strongly concave (e.g. Fig. 5.3 from Memmott *et al.,* 2004). Such networks are known as 'small-world' networks, as any two hubs can be connected through just a small number of their neighbours. The visitation observation data of Memmott *et al.* (2004) suggested that pollination interactions fit the power law and may be considered small world networks,

which is reflected in the strongly concave curve of Fig. 5.3, taken from Memmott *et al.* (2004). While the data analysed here include substantial differences in hub size (i.e. number of pollens carried) between different insect genera the interactions seem too generalist to fit the power law. No insect genus carried just one pollen genus and the insects-interactions graph appears more random than concave (Fig. 5.6). This may be because this study used pollen load analysis not visitation observations (thereby representing brief and infrequent visits more fully) or because it was carried out in Kenya, an ecosystem where drought and overgrazing may have strongly selected for generalists. However, it is more likely that the poor power function fit was a result of the small, incomplete data set. Jordano *et al.* (2003) found that small plant-pollinator networks sometimes deviated from the power law with fewer super-generalists than predicted. The networks analysed here are much smaller (and less complete) than those in Memmott *et al.* (2004) so Jordano's 'truncated' power law may be a better description of these *Acacia* pollination community data. If these networks were enlarged with additional pollen loads they may come to be scale-free. This idea is supported by Figs. 5.8 and 5.9 where a concave curve is suggested through the chaos and ten pollens were only carried by a single insect genus. The insects and pollens are currently discussed as genera, not as species, which makes interactions less precise so they appear more generalist. If the pollen loads of insect species were considered separately it may be found that most insect species do carry a single plant species. Likewise, if a 1-5% of total load threshold was set for an interaction to be included this may remove many weak connections so more insects carry just a single pollen.

5.4.3 Combined removals

The insects that were selected by the 'Smasher' genetic algorithm to be part of the smallest most-damaging or least-damaging groupings were not of a single family, neither did they share guild characteristics such as body size and tongue length. This suggests that family and guild may be inappropriate groupings when considering pollen transport, thus the value of automated identification in pollination ecology (to genus or species level) may be even greater than previously thought.

Removal of pollinator combinations produced an inversely proportional relationship for random removals (Fig, 5.15), a slightly concave curve for most-damaging removals and a strongly convex curve (with network collapse only when the last few insects are removed) for least-damaging removals (Fig. 5.16). Possible reasons for these trends were considered in section 5.3.3. They are similar in basic architecture to the networks for visitation observations in the Rocky Mountains of Colorado (Clements & Long, 1923) and the prairie-forest of Illinois (Robertson, 1929) by Memmott *et al.* (2004). It is possible that the differences in architecture are due to the under sampling of these pollen load data, as discussed in 5.4.1.

If the differences in architecture are due to the weak data set then visitation observations and pollen load analyses seem to provide comparable reflections of pollination activity, and similar pollination communities in very different geographical regions seem to be similarly tolerant to insect extinction. It would be interesting to collaborate with pollination ecologists working in different ecosystems (e.g. Mediterranean, tundra and rainforest) to see if the basic architecture holds true.

Memmott *et al.* (2004) prioritised insects for combination removal by hub size (the number of plants they visited), while this work based removal order on the more sophisticated measure, damage. The two methods produced similar results. This suggests that hub size alone (without the ripple effects included in damage) is effective in describing damage caused by combined insect extinctions.

The loss of a small number of highly connected nodes in networks like the Internet may cause network breakdown (Solé & Montoya, 2001; Newman, 2003). Alternatively, Memmott *et al.* (2004) suggested that the loss of network integrity in pollination networks is essentially linear. The analyses of combined removals in these data support Memmott *et al.* (2004). The most serious decline in network integrity occurred when the most damaging combinations of genera were removed first. This decline is only slightly concave rather than catastrophic (Fig. 5.16). This tolerance to extinction can be explained by two interacting factors, redundancy and nestedness (Memmott *et al.*, 2004). As most plants and insects in the system studied here are generalists there is functional redundancy. A single plant may have its pollen carried by many insect genera and a single insect may gather pollen from many plants, thus the extinction of a single node is unlikely to lead to other extinctions. This tolerance to extinction may support Ghazoul's school of thought, that there is no serious pollinator crisis (Ghazoul, 2005), but the many limitations of these data should not be forgotten.

Bascompte *et al.* (2003) describe nestedness when the connections of the specialists are nested within those of the generalists so the community is organised cohesively around a central core of interactions. Specialised interactions tend to be a subset of generalist interactions (Memmott *et al.*, 2004). Bascompte *et al.* (2003) found that 52 mutualistic plant-animal networks were highly nested and suggested that the larger the network the more nested it would be. Put in terms of ecosystem functioning many of the specialists can be lost without significant damage to ecosystem function, as their roles are also covered by generalists. This is important, as specialists are likely to be the first to go extinct (Rathcke & Jules, 1993). In the data analysed here, the insects that carried few pollen genera tended not to be specialists but generalists that contacted pollen only infrequently (e.g. wasps that were mainly predatory but visited an occasional flower to get an energy boost from nectar) or had morphology poorly suited to carrying pollen (e.g. blowflies). Nevertheless, the same pattern is appropriate, as the interactions of these poorer quality pollen carriers were a subset of generalist

interactions. Nestedness confers three important properties, seen in the present data. First, pollens that are carried by few insects are carried by insects that carry many pollens. All ten of the pollen genera that were only carried by a single insect genus were carried by generalist insects: *Megachile* (carried 23 pollens), *Apis* (20), *Macrogalea* (15), *Rhyncomya* (11), *Delta* (10), *Xylocopa* (9) and *Rhinia* (8) (Table 5.3). Second, insects that carry few pollens tend to forage on pollens that are carried by many other insects. This trend was also noted, e.g. *Sphex* (a predatory wasp) carried just two pollen genera but these were both pollens that were widely carried, *Acacia* interacted with 23 insects and *Heliotropium* with eight. The third property is the absence of compartments of disconnected interactions, typified by pollination syndromes such as bee-flowers and fly-flowers (Dicks *et al.,* 2002). As it has already been said that nodes with few connections always connected to those with many connections this property has already been satisfied.

5.4 Summary and the way forward

Network analysis is a useful tool for pollination ecology, as it allows the combined effects of pollen transport interactions to be considered. 'Damage' is an informative network metric, more powerful than hub size as it summarises both the direct and the ripple effects of node removals.

The network data were too incomplete for node removals to be analysed with confidence. It is important to analyse many more loads, sampling more balanced numbers from each insect, at many times and on many dates. Both insects and pollens should be resolved to species level and a load threshold of 1% should be in place to prevent contaminant pollens from producing false connections. This fuller investigation will resolve whether the rather random distribution seen in Fig. 5.6 would approximate a concave curve or retain distinctive differences from the Memmott *et al.* (2004) data due to the methodology (pollen loads not visitation analyses) or ecosystem.

Pollen load analysis is a successful basis to study combined removals from pollination networks, providing similar results to much larger datasets of flower visitation observations. Combined removals suggest that the *Acacia* pollination community at Mpala Research Centre is tolerant to extinctions. Fifty percent of insect genera could be removed at random with only a 25% reduction in network integrity. When the least damaging groups were selected the initial loss to network integrity was small, 85% of insect genera could be removed with only a 40% loss in integrity. However, the loss of the final two genera then caused catastrophic network breakdown. Even when the most damaging combinations were removed the network damage only a shallowly concave curve. This tolerance to insect removals can be explained by the generality and nested nature of the network.

Chapter 6 – Conclusions

This research has produced three original studies applying new software to the study of pollination ecology. First, two alternative methods for insect wing identification by DAISY have been investigated. They were not only better suited to the practicalities of fieldwork, but also almost as accurate as the established method. Second, it was hoped that DAISY would also be an effective tool for pollen identification, a totally new challenge for this software. Unfortunately, these data suggested that DAISY was not the best software for pollen identification. Third, small-scale network analyses investigated insect removals from pollen load data for the first time. These analyses suggested that the insect communities involved in acacia pollination are stable and highly tolerant to insect removal.

DAISY wing identification gave highly accurate results overall, with many taxa identified with 100% accuracy. The standard method, S (in which wings were removed from specimens and wing cells highlighted polygonally) was 93% accurate when the closest match was taken to be the identification (FPTP) and 79% accurate when the Coord3 metric, which demanded high certainty (as the closest three matches had to 'agree' an a classification), was used.

The attached wing method (A) provided identification accuracy that was not significantly lower (Wilcoxon's signed ranks test, $P > 0.05$) than S. This held true for family, genus or species level identification and even when the Coord3 metric demanded high certainty. This negligable accuracy difference when wings remained attached is encouraging for the use of live and museum specimens.

The box-crop method (B) also showed great potential. This approach is much quicker than S as the user need only mark two landmark points on each image to highlight the region of interest (ROI), rather than the 20+ points necessary to draw a polygonal ROI. For FPTP the accuracy of B was not significantly lower (Wilcoxon's signed ranks test, $P > 0.05$) than that of S, whether identification was at family, genus or species level. However, when the Coord3 metric demanded high certainty, B did perform with significantly lower accuracy (Wilcoxon's signed ranks test, $P < 0.05$) (at family, genus and species level). This reduction in accuracy was only relatively small (8% at most in the mean values). Different users of the system must make an informed decision about the certainty they require and the time they are willing to invest to achieve that.

Changing the size of the Normalised Polar Thumbnails had a dramatic impact on the identification accuracy of some species but only a small impact on mean accuracy. Overall a 24 pixel thumbnail was identified more accurately than the 32 pixel thumbnail previously recommended. Generally the addition of extra specimens to training sets produced a curve of diminishing returns. This supported the findings of previous DAISY studies (Weeks *et al.*, 1997; Weeks *et al.*, 1999b; Gauld *et al.*, 2000; Pajak, 2001;

Watson *et al.*, 2004). The asymptote was reached with just six or seven training specimens. This is encouraging data for wing identification applications of DAISY as large training sets are very time consuming to produce and require specimens that may be unavailable to a field ecologist.

Pollen proved to be less suited to DAISY classification than insect wings. Cleaning pollen by acetolysis, imaging pollen with dark-field microscopy and maintaining the relative sizes of different pollens all improved the accuracy of identification but not enough to give satisfactorily accurate results. The highest accuracy obtained was around 73%. However, this was for a set of just 27 pollen genera. When larger numbers of pollens were considered (80+ genera) the accuracy fell to 50-60%. Pollen is small, three-dimensional and often translucent. However, the greatest challenge for DAISY seemed to be the lack of fixed points for orientation, i.e. a spherical pollen grain has no intuitive top, bottom, front or back. DAISY identifies normalised polar thumbnails on a pixel-by-pixel basis. Without fixed points for orientation even a single grain would provide very different pixel parameters at different orientations. DAISY pollen accuracy is well behind that of the pollen-centred systems of Trelour *et al.* (2004) and Li *et al.* (2004) which achieved 95% and 100% accuracy so pollen is probably best left to these systems and DAISY developed for the objects it suits best.

Network analysis allows many pollen transport interactions to be considered in combination so it is a useful tool in pollination ecology. A few large networks have been produced in the past (e.g. Memmott *et al.*, 2004) but these were based on observed insect visits to flowers. The current networks were instead based on analysis of pollen loads. While this is still insufficient to be direct evidence of pollination, it reflects an entire foraging trip more accurately than visitation observations. Unfortunately, these network data were too incomplete for node removals to be analysed with confidence. It is important to discriminate insects and pollens to species level, analyse many more loads and sample evenly from each insect taxa, at many times of day and on many dates. A load threshold of 1%+ of total load would prevent contaminant pollens from producing false connections. This fuller investigation would resolve whether the rather random distribution of pollen genera carried by insect genera found here (Fig. 5.6) would approximate a concave curve or show distinctive differences from the Memmott *et al.* (2004) trend.

Combined removals suggest that the *Acacia* pollination community at Mpala Research Centre is tolerant to extinctions. Fifty percent of insect genera could be removed at random with only a 25% reduction in network integrity. When the least damaging groups were selected the initial loss to network integrity was small, 85% of insect genera could be removed with only a 40% loss in integrity. However, the loss of the final two genera then caused catastrophic network breakdown. Even when the most damaging combinations were removed the network damage produced only a shallow concave curve. This tolerance to insect removals can be explained by the generality and nested nature of the network and suggests that the extinction of a small subset of pollen carrying insects in the Kenyan

savannah environment is unlikely to lead to a serious pollination crisis. However, these networks do not discriminate flower visitors from pollinators so in-depth fieldwork as well as network modelling is required to give decisive data on this issue.

References

Amaral, L. A. N., Scala, A., Barthélémy, M. and Stanley, H. E. 2000. Classes of small- world networks. *Proceedings of the National Academy of Sciences of the United States of America,* **97**, 11149-11152.

Albert, R., Jeong, H. and Barabási, A. –L. 1999. Diameter of the world-wide web. *Nature,* **401**, 130-131.

Albert, R., Jeong, H. and Barabási, A. –L. 2000. Error and attack tolerance of complex networks. *Nature,* **406**, 378-382.

Allen-Wardell, G., Bernhardt, P, Bitner, R., Burquez, A., Buchmann, S., Cane, J., Cox, P. A., Dalton, V., Feinsinger, P., Ingram, M., Inouye, D., Jones, C. E., Kennedy, K., Kevan, P., Koopowitz, H., Medellin, R., Medellin-Morales, S., Nabhan, G. P., Pavlik, B., Tepedino, V., Torchio, P and Walker, S. 1998. The potential consequences of pollinator declines on the conservation of biodiversity and stability of crop yields. *Conservation Biology*, **12**, 8-17.

Almeida-Muradian, L. B., Pamplona, L. C., Coimbra, S. and Barth, O. M. 2005. Chemical composition and botanical evaluation of dried bee pollen pellets. *Journal of Food Composition and Analysis,* **18** (1), 105-111.

Arbuckle, T., Schroder, S., Steinhage, V. and Wittman, D. 2001. Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In: *15th International Symposium on Informatics for Environmental Protection*, Zurich. pp. 425- 430.

Arce, L. R. and Banks, H. 2001. A preliminary survey of pollen and other morphological characters in neotropical *Acacia* subgenus *Aculeiferum* (Leguminosae: Mimosoideae). *Botanical Journal of the Linnean Society*, **135,** 263-270.

Armstrong, J. A. 1979. Biotic mechanisms in the Australian flora – a review. *New Zealand Journal of Botany,* **17**, 467-508.

Baldock, K. C. R., Bulleid, R. J., Cant, E. T. & Nelson, I. L. 2004. University of Bristol Pollination Ecology Expedition 2001 Report. University of Bristol Expedition Report Series #6.

Barber, D. 2004. Learning from data: nearest neighbour classification. Online lecture notes of Amos Storkey, University of Edinburgh. www.anc.ed.ac.uk/~amos/lfd/lectures/lfd_2004_nearest_neighbour.pdf (last accessed 21/09/05).

Barrett, R. D. H. and Hebert, P. D. N. 2005. Identifying spiders through DNA barcodes. *Canadian Journal of Zoology – Revue Canadienne de Zoologie,* **83** (3), 481-491.

Bascompte, J., Jordano, P., Melian, C. J. and Olesen, J. M. 2003. The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America,* **100** (16), 9383-9387.

Bastos, D. H. M., Barth, O. M., Rocha, C. I., Cunha, I. B. D., Carvalho, P. D., Torres, E. A. S. and Michelan, M. 2004. Fatty acid composition and palynological analysis of bee (Apis) pollen loads in the states of Sao Paulo and Minas Gerais, Brazil. *Journal of Apicultural Research,* **43** (2), 35-39.

Beale, R. and Jackson, T. 1990. *Neural computing: an introduction.* Bristol: Institute of Physics Publishing.

Becker, S. 1991. Unsupervised learning procedures for neural networks. *International Journal of Neural Systems*, **2**, 17-33.

Bernhardt, P. 1986. Bee-pollination in *Hibbertia fasciculate* (Dilleniaceae). *Pollination, Systematics and Evolution,* **152**, 231-241.

Bernhardt, P. 1989. Floral ecology of Australian *Acacia*. In: *Advances in Legume Biology* (Ed. by J.L.Zarucci, C. H. S. A.), pp. 263-281: Missouri Botanical Garden.

Blundell, M. 1992. *The Collins Photo Guide to Flowering Plants of East Africa.* Collins, London.

Boddy, L. and Morris, C. W. 1993. Analysis of flow cytometry data – a neural network approach. *Binary*, **5**, 17-22.

Boddy, L. and Morris, C.W. 1997. Artificial neural networks for identification. First Bionet International Working Group on Automated Taxonomy. University of Wales, Cardiff. pp. 29-37.

Boddy, L.; Morris, C.W.; Wilkins, M.F.; Tarran, G.A. and Burkill, P.H. 1994. Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry* **15**, 283-93.

Boddy, L., Wilkins, M. K. and Morris, C. W. 2001. Pattern recognition in flow cytometry. *Cytometry* **44** (3), 195-209.

Bond, W. J. 1995. Assessing the risk of plant extinction due to pollinator and disperser failure. In *Extinction Rates* (Ed. by Lawton, J. H. and May, R. M.), pp. 131-146. Oxford University Press, Oxford.

Bonton, P., Boucher, A., Thonnat, M., Tomczak, R., Hidalgo, P. J., Belmonte, J. and Galan, C. 2001. Colour image in 2D and 3D microscopy for the automation of pollen rate measurement. *Image Anal Stereol,* **20**, 527-532.

Brauchli, K., Christen, H., Haroske, G., Meyer, W., Kunze, K. D. and Oberholzer, M. 2002. Telemicroscopy by the Internet revisited. *Journal of Pathology,* **196** (2), 238-243.

Chapman, J. W., Reynolds, D. R. and Smith, A. D. 2003. Vertical-looking radar: A new tool for monitoring high-altitude insect migration. *Bioscience,* **53**, 503-511.

Chesmore, E. D. 1997. Methodologies for automating the identification of species. In: *First Bionet International Working Group on Automated Taxonomy*, pp. 3-12. Cardiff University.

Chesmore, E. D. 1999. Technology transfer: applications of electronic technology in ecology and entomology for species identification. *Natural History Research*, **5,** 111-126.

Chesmore, E. D. and Ohya, E. 2004. Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. *Bulletin of Entomological Research,* **94** (4), 319-330.

Clements, R. E. & Long, F. L. 1923. *Experimental pollination: an outline of the ecology of flowers and insects.* Carnegie Institute of Washington, Washington DC.

Coe, M. and Beentje, H. 1991. *A Field Guide to the Acacias of Kenya*. Oxford: Oxford University Press.

Connie, T., Jin, A. T. B., Ong, M. G. K. and Ling, D. N. C. 2005. An automated palmprint recognition system. *Image and vision computing,* **23** (5), 501-515.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R. V., Paruelo, J., Raskin, R. G., Sutton, P and vandenBelt, M. 1997. The value of the world's ecosystem services and natural capital. *Nature,* **387**, 253-260.

Culverhouse, P. F., Simpson, R. G., Ellis, R., Lindley, J. A., Williams, R., Parisini, T., Reguera, B., Bravo, I., Zoppoli, R., Earnshaw, G., McCall, H. and Smith, G. 1996. Automated classification of field-collected dinoflagellates by artificial neural network. *Marine Ecology Progress Series*, **139,** 281-287.

Culverhouse, P. F. and Williams, R. 2003. Remote and local automatic analysis of zooplankton species. North Pacific Marine Science Organisation (PICES) conference poster. http://www.cis.plym.ac.uk/cis/publications/PICES-poster-2003.pdf (last accessed 21/09/05).

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V. and Gonzalez-Gil, S. 2003. Expert and machine discrimination of marine flora: a comparison of recognition accuracy of field-collected phytoplankton. *International Conference on Visual Information Engineering*, 177-183.

Culverhouse, P. F. 2005. Natural object recognition – machines versus humans. Symposium proceedings: *Algorithmic approaches to the identification problem in systematics.* 19 August 2005, The Natural History Museum, London.

Crowson, R. A. 1981. *The Biology of the Coleoptera.* Academic Press, London.

Dafni, A. 1992. *Pollination Ecology: a practical approach*. Oxford: Oxford University Press.

Daly, H. V. and Balling, S. S. 1978. Identification of Africanized honey bees in the Western Hemisphere by discriminant analysis. *Journal of the Kansas Entomological Society*, **51**, 857-869.

Daly, H. V., Hoelmer, K., Norman, P. and Allen, T. 1982. Computer-assisted measurement and identification of honey bees (Hymenoptera: Apidae). *Annals of the Entomological Society of America*, **75**, 591-594.

Davies, J. 1999. Beetles (Coleoptera) of Mkomazi. In: *Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna* (Ed. by Coe, M. J., McWilliam, N. C., Stone, G. N. and Packer, M. J.), pp. 249-268. London: The Royal Geographical Society (with The Institute of British Geographers).

Dicks, L. V., Corbet, S. A. and Pywell, R. F. 2002. Compartmentalization in plant-insect flower-visitor webs. *Journal of Animal Ecology*, **71**, 32-43.

De Sa-Otero, M. P., Gonzalez, A. P., Rodriguez-Damian, M. and Cernadas, E. 2004. Computer-aided identification of allergenic species of Urticaceae pollen. *GRANA,* **43** (4), 224-230.

Degrandi-Hoffman, G., Thorpe, R., Loper, G. and Eisikowitch, D. 1992. Identification and distribution of cross-pollinating honey-bees on almonds. *Journal of Applied Ecology*, **29,** 238-246.

Determann, H. and Lepusch, F. (1979). *The Microscope and Its Application.* Ernst Leitz Wetzlar GmbH. [Instruction manual accompanying the Leitz Dialux 20 compound microscope].

Do, M. T., Harp, J. M. and Norris, K. C. 1999. A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research,* **89** (3), 217-224.

Domingo, M. and Uriz, M. J. 1998. Design and development of SPONGIA, an expert system for sponges identification. *Scientia Marina,* **62** (1-2), 45-57.

Domingo, M., Martin-Baranera, M., Sanz, F., Sierra, C. and Uriz, M. J. 1999. Validating SPONGIA, an expert system for sponge identification. *Expert systems with applications,* **16** (4), 379-384.

Dunne, J. A., Williams, R. J. and Martinez, N. D. 2002. Network structure and biodiversity loss in food webs: Robustness increases with connectance. *Ecology Letters,* **5**, 558-567.

Dupraw, E. J. 1965. The recognition and handling of honey-bee specimens in non-Linnean taxonomy. *Journal of Apicultural Research*, **4**, 71-84.

Dytham, C. 1999. *Choosing and Using Statistics: A biologists guide.* Blackwell Science Ltd., Oxford.

Eardley, C. 2002. Afrotropical bees now: what next? In: *Pollinating Bees – TheConservation Link Between Agriculture and Nature* (Ed. by Fonseca, P. K. a. V. L. I.), pp. 97-104. Brasilia: Ministry of Environment / Brasilia.

Edwards, M. and Morse, D. R. 1995. The potential for computer-aided identification in biodiversity research. *Trends in Ecology and Evolution*, **10,** 153-158.

Ekenel, H. K. and Sankur, L. 2005. Multiresolution face recognition. *Image and vision computing,* **23** (5), 469-477.

Eltz, T., Bruhl, C. A., Van de Kaars, S. and Linsenmair, K. E. 2001. Assessing stingless bee pollen diet by analysis of garbage pellets: a new method. *Apidologie*, **32,** 341-353.

Evans, G. 1975. *The life of beetles*. 232 pp. London: George Allen and Unwin Ltd.

Faegri, K. 1992. *Textbook of pollen analysis*. John Wiley and Sons, Ltd.

Faegri, K. and van der Pijl, L. 1966. *The Principles of Pollination Ecology.* Pergamon, Oxford.

Flenley, D. R. 1968. The problem of Pollen Recognition. In: *Problems in Picture Interpretation* (Ed. by Clowes, M. B. and Penny, J. P.), pp.141-145. CSIRO, Canberra.

Flenley, D. R. 2003. Some prospects for lake sediment analysis in the 21st century. *Quaternary International* **105**, 77-80.

Forsythe, T. G. 1987. *Common Ground Beetles*. Richmond: The Richmond Publishing Co. Ltd.

France, I., Duller, A. W. G., Duller, G. A. T. and Lamb, H. F. 2000. A new approach to automated pollen analysis. *Quaternary Science review*, **19,** 537-546.

Galbraith, J. M. and Bryant, R. B. 1998. A functional analysis of soil taxonomy in relation to expert system techniques. *Soil Science*, **163,** 739-747.

Galbraith, J. M., Bryant, R. B. and Ahrens, R. J. 1998. An expert system for soil taxonomy. *Soil Science*, **163,** 748-758.

Gaston, K. J. and May, R. M. 1992. Taxonomy of taxonomists. *Nature*, **356,** 281-282.

Gaston, K. J. and O'Neill, M. A. 2004. Automated species identification - why not? *Philosophical Transactions of the Royal Society of London* B, **359** (1444), 655-667.

Gauld, I. D. 1991. The Ichneumonidae of Costa Rica, 1. *Memoirs of the American Entomological Institute*, **47**, 1-589.

Gauld, I. D., O'Neill, M. A. and Gaston, K. J. 2000. Driving Miss Daisy: the performance of an automated insect identification system. In *Hymenoptera: Evolution, Biodiversity and Biological Control* (Ed. by Austin, A. D. and Dowton, M.), pp. 303-312. CSIRO, Canberra.

Ghazoul, J. 2005. Buzziness as usual? Questioning the global pollination crisis. *Trends in Ecology and Evolution,* **20**, 367-373.

Gilbert, F. S. 1993. *Hoverflies*. Richmond: The Richmond Publishing Co. Ltd.

Godfray, H. C. J. 2002. Challenges for taxonomy. *Nature*, **417,** 17-19.

Goodwin, R. M. and Perry, J. H. 1992. Use of pollen traps to investigate the foraging behaviour of Honey-bee colonies in kiwifruit orchards. *New Zealand Journal of Crop and Horticultural Sciences*, **20** (1), 23-26.

Green, J. L., Hastings, A., Arzberger, P., Ayala, F. J., Cottingham, K. L., Cuddington, K., Davis, F., Dunne, J. A., Fortin, M. J., Gerber, L. and Neubert, M. 2005. Complexity in ecology and conservation: Mathematical, statistical, and computational challenges. *Bioscience,* **55** (6), 501-510.

Guinet, P. 1981. Mimosoidaea: the characters of the pollen grains. In: *Advances in Legume Systematics* (Ed. by Raven, R. M. P. a. P. H.), pp. 835-857. London: Royal Botanical Gardens Kew.

Guinet, P. 1986. Geographical patterns of the main pollen characters in the genus *Acacia* (Leguminosae), with particular reference to subgenus Phyllodineae. In: *Pollen and Spores: Form and Function* (Ed. by Blackmore, S. and Ferguson, I. K.), pp. 97-311. London: Academic Press.

Gullan, P. J. and Cranston, P. S. 1994. *The Insects: an Outline of Entomology.* Chapman & Hall, London.

Gunter, S. and Bunke, H. 2005. Off-line cursive handwriting recognition using multiple classifier systems - on the influence of vocabulary, ensemble, and training set size. *Optics and lasers in engineering,* **43** (3-5), 437-454.

Hajdaoud, A., Schroder, S. and Steinhage, V. 2005. ABIS: Automatic identification of bees. : Uni Bonn. http://www.informatik.uni-bonn.de/projects/ABIS/ (last accessed 21/09/05).

Hawkeswood, T. H. 1983. Observations on *Pyrgoides dryops* (Blackburn) Coleoptera: Chrysomelidae, a pollen-feeding beetle on *Acacia leiocalyx* (Domin.) Pedley, at Brisbane, south-east Queensland. *Victoria Naturalist,* **100**, 156-158.

Hawkeswood, T. H. 1985. The role of butterflies as pollinators of *Acacia bidwilii* (Mimosaceae) at Townsville, Northern Queensland. *Australian Journal of Botany,* **33**, 167-173.

Haykin, S. 1999. *Neural Networks: a comprehensive foundation.* Prentice-Hall International, Inc., New Jersey.

Hebert, P. D. N., Cywinska, A., Ball, S. L. and de Waard, J. R. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London* B, **270**, 313-321.

Heseltine, T., Pears, N. and Austin, J. 2002. Evaluation of image pre-processing techniques for eigenface based face recognition. *Proceedings of the Second International Conference on Image and Graphics, SPIE,* **4875**, 677-685.

Higgins, L. and Hargreaves, B. 1983. *The Butterflies of Britain and Europe*. London: Collins.

Hodges, D. 1984. *Pollen loads of the Honeybee: A guide to their identification by colour and form.* International Bee Research Association, Cardiff.

Holsinger, K. E. 2000. Reproductive systems and evolution in vascular plants. *Proceedings of the National Academy of Science of the United States of America,* **97**, 7037-7042.

Houle, D., Mezey, J., Galpern, P. and Carter, A. 2003. Automated measurement of drosophila wings. *BMC Evolutionary Biology,* **3**, art no. 25.

Hubbard, B. B. 1998. *The World According to Wavelets*. Second edition. A. K. Peters, Ltd., Natick, Massachusetts.

Huele, R. and Haes, H. U. 1998. Identification of individual sperm whales by wavelet transform of the trailing edge of the flukes. *Marine Mammal Science*, **14,** 143-145.

Hughes, J. B., Daily, G. C. and Ehrlich, P. R. 1997. Population diversity: its extent and extinction. *Science*, **278**, 689-692.

Janzen, D. H. 1974. Swollen-thorn *Acacia* species of Central America. *Smithsonian Contributions to Botany* **13**.

Jordano, P. 1987. Patterns of mutualistic interactions in pollination and seed dispersal-connectance, dependance, asymmetries and coevolution. *American Naturalist,* **129**, 657-677.

Jordano, P., Bascompte, J. and Olesen, J. M. 2003. Invariant properties in coevolutionary networks of plant-animal interactions. *Ecology Letters,* **6**, 69-81.

Kapp, R. O. 1969. *How to know pollen and spores*. Wm C Brown, Iowa.

Kearns, C. A., Inouye, D. W. and Waser, N. M. 1998. Endangered mutualisms: the conservation of plant-pollinator interactions. *Annual Review of Ecology and Systematics*, **29,** 83-112.

Kenrick, J. 2003. Review of pollen-pistil interactions and their relevance to the reproductive biology of *Acacia*. *Australian Systematic Botany*, **16,** 119-130.

Kenrick, J. and Knox, R. B. 1989. Pollen-pistil interactions in Leguminosae (Mimosoideae). In: *Advances in Legume Biology* (Ed. by Stirton, C. H. and Zarucchi, J. L.), pp. 127-156. St Louis, Missouri: Missouri Botanical Garden.

Kesseler, R and Harley, M. 2004. Pollen: The Hidden Sexuality of Flowers. Andreas Papadakis Ltd.

Kevan, P. G. 1999. Pollinators as bioindicators of the state of the environment: species, activity and diversity. *Agriculture, Ecosystems and Environment,* **74**, 373-393.

Kim, H. T., Ikeda, Y. and Choi, H. L. 2005. The identification of Japanese black cattle by their faces. *Asian-Australasian journal of animal sciences,* **18** (6), 868-877.

Kitching, I. J., Forey, P. L., Humphries, C. J. and Williams, D. 1998. *Cladistics: The theory and practice of parsimony analysis*. Second edition. Systematics Association Publication 11. Oxford University Press, Oxford.

Knapp, S., Bateman, R. M., Chalmers, N. R., Humphries, C. J., Rainbow, P. S., Smith, A.B., Taylor, P. D., Vane-Wright, R. I. and Wilkinson, M. 2002. Taxonomy needs evolution, not revolution. *Nature*, **419,** 559.

Knox, R. B. 1979. *Pollen and allergy.* Studies in Biology No. 107. Edward Arnold Ltd., London.

Knox, R. B., Bernhardt, P., Marginson, R., Beresford, G., Baker, I. and Baker, H. G. 1985. Extra-floral nectarines as adaptations for bird pollination in *Acacia terminalis. American Journal of Botany*, **72**, 1185-1196.

Koeniger, G., Koeniger, N., Mardan, M., Otis, G. and Wongsiri, S. 1991. Comparative anatomy of male genital organs in the genus *Apis*. *Apidologie*, **22,** 539-552.

Kohonen, T. 1989. *Self-Organization and Associative Memory.* 3rd edition. Springer-Verlag, NewYork.

Kohonen, T. 2001. *Self-Organizing Maps*. Springer, Berlin.

Kondoh, M. 2003. Foraging adaptation and the relationship between food-web complexity and stability. *Science*, **299**, 1388-1391.

Langenberger, M. W. and Davis, A. R. 2002. Honey bee pollen foraging in relation to flowering phenology of biennial caraway (*Carum carvi* L.). *Canadian Journal of Plant Science,* **82** (1), 203-215.

Lacave, C. and Diez, F. J. 2004. A review of explanation methods for heuristic expert systems. *Knowledge Engineering Review,* **19** (2), 133-146.

Lacey, A. J. 1999. Basic Optical Microscopy. In: *Light Microscopy in Biology* (Ed. by Lacey, A. J.). The Practical Approach Series. Oxford University Press, Oxford.

Lang, R. I. W. 2000. A Future for Dynamic Neural Networks. First year report, University of Reading.

Lang, R.I.W. 2001. Initial Study into the Plastic Self Organising Map. Second year report, University of Reading.

Lang, R.I.W. and Warwick, K. 2002. The Plastic Self Organising Map. *International Joint Conference on Neural Networks*, **1**, 727-732.

Lang, R.I.W. 2005. Plastic Self-Organising Maps. Symposium proceedings: *Algorithmic approaches to the identification problem in systematics.* 19 August 2005, The Natural History Museum, London.

Langford, M., Taylor, G. E. and Flenley, J. R. 1990. Computerised identification of pollen grains by texture analysis. *Review of Palaeobotany and Palynology*, **64,** 197-203.

LeFeuvre, P., Rose, G. A., Gosine, R., Hale, R., Pearson, W. and Khan, R. 2000. Acoustic species identification in the Northwest Atlantic using digital image processing. *Fisheries research,* **47** (2-3), 137-147.

Li, P., Treloar, W.J., Flenley, J.R. and Empson, L. 2004. Towards automation of palynology 2: the use of texture measures and neural network identification of optical images of pollen grains. *Journal of Quaternary Science,* **19** (8), 755-762.

Liao, S-H. 2005. Expert system methodologies and applications – a decade review from 1995 to 2004. *Expert Systems with Applications.* **28** (1), 93-103.

Lindo, A. C., Rodriguez, P. G., Avila, M. M., Antequera, T. amd Palacios, R. 2004. Computer vision algorithms versus traditional methods in food technology: The desired correlation. *Progress in Pattern Recognition, Image Analysis and Applications Lecture Notes in Computer Science,* **3287**, 59-66.

Lipscomb, D., Platnick, N. and Wheeler, Q. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends in Ecology and Evolution*, **18,** 65-66.

Longino, J. T. 1994. How to measure arthropod diversity in a tropical rainforest. *Biology International,* **28**, 3-13.

Longino, J. T., Colwell, R. K. 1997. Biodiversity assessment using structured inventory: Capturing the ant fauna of a tropical rainforest. *Ecological Applications,* **7**, 1263-1277.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., Hooper, D. U., Huston, M. A., Raffaelli, D., Schmid, B., Tilman, D. and Wardle, D. A. 2001. Ecology - Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science,* **294** (5543), 804-808.

Lucas, S. M. 1997. Face recognition with the continuous n-tuple classifier. *British Machine Vision Conference Proceedings,* **1**, 222-231.

Lynuxworks. 2005. POSIX. http://www.lynuxworks.com/products/posix/posix.php3 (last accessed 21/09/05).

MacLeod, N., O'Neill, M. A. and Walsh, S. A. In prep. Use of Unsupervised Neural Nets to address the taxonomic impediment in systematics.

Mallet, J. and Willmott, K. 2003. Taxonomy: renaissance or Tower of Babel? *Trends in Ecology and Evolution*, **18,** 57-59.

Maslin, B. R. 2001a. Introduction to *Acacia.* In *Flora of Australia,* vol. 11A, part 1 (Ed. by Orchards, A. E. and Wilson, A. J. G.), pp. 3-13. CSIRO Publishing, Melbourne, Australia.

Maslin, B. R. 2001b. Wattle Acacias of Australia. Perth: Australian Biological Resources Study, Canberra and Department for Conservation and Land Management, Perth.

Maslin, B. R. 2002. The role and relevance of taxonomy in the conservation and utilisation of Australian *Acacias*. *Conservation Science of Western Australia*, **4,** 1-9.

May, R. 2004. Tomorrow's taxonomy: collecting new species in the filed will remain the rate-limiting step. *Philosophical Transactions of the Royal Society of London* B, **359** (1444), 733-734.

McGavin, G.C. 1992. *The pocket guide to insects of the Northern Hemisphere*. Parkgate Books, London.

Memmott, J. 1999. The structure of a plant-pollinator food web. *Ecology Letters*, **2**, 276-280.

Memmott, J., Waser, N. M. and Price, M. V. 2004. Tolerance of pollination networks to species extinction. *Proceedings of the Royal Society of London Series B,* **271** (1557), 2605-2611.

Menon, A. 2004. Inequality's arrow: The role of greed and order in genetic algorithms. *Genetic and evolutionary computation, Proceedings Lecture Notes in Computer Science,* **3102**, 1352-1364.

Michener, C. D. 2000. *The Bees of the World.* Baltimore: The John Hopkins University Press, Baltimore.

Minelli, A. 2003. The status of taxonomic literature. *Trends in Ecology and Evolution*, **18,** 75-76.

Mirkin, G. R. and Bagdasaryan, L. L. 1972. The feasibility of identifying paleontological objects with the aid of optical analysing systems. *Paleontological journal* **6**, 103-108.

Montague, G. and Morris, J. 1994. Neural-network contributions in biotechnology. *Trends in Biotechnology*, **12**, 312-324.

Moore, A. and Miller, R. H. 2002. Automated identification of optically sensed aphid (Homoptera: Aphidae) wingbeat waveforms. *Annals of the Entomological Society of America,* **95**, 1-8.

Nabhan, G. P. and Buchmann, S. L. 1997. Services provided by pollinators. In *Nature's services: societal dependence on natural ecosystems.* (Ed. by Daly, G. C.), pp. 133-150. Island Press, Washington DC.

Ness, E. 2005. SPIDA-web: Artificial neural networks fill in for taxonomists. *Conservation in Practice,* **6**, 1.

New, T. R. 1984. *A Biology of Acacias*. Oxford University Press, in association with Latrobe University Press, Melbourne, Australia.

Newman, D. R. 1998. *Neural networks.* Online lecture notes, Queen University, Belfast. http://www.qub.ac.uk/mgt/intsys/nn.html (last accessed 21/09/05).

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM [Society for Industrial and Applied Mathematics] Review,* **45** (2), 167-256.

Nottingham University. 2005. Histogram equalisation. Computer science assignment. http://www.cs.nott.ac.uk/~tpp/G5AIVI/histogram.pdf (last accessed 21/09/05).

Obrist, M. K., Boesch, R., Fluckiger, P. F. 2004. Variability in echolocation call design of 26 Swiss bat species: consequences, limits and options for automated field identification with a synergetic pattern recognition approach. *Mammalia* **64** (4), 307-322.

Oliver, I., Pik, A., Britton, D., Dangerfield, M., Colwell, R. K. and Beattie, A. J. 2000. Virtual Biodiversity Assessment Systems. *Bioscience,* **50** (5), 441.

Oliver, I. and Beattie, A. J. 1993. A possible method for the rapid assessment of biodiversity. *Conservation Biology,* **7**, 562-568.

O'Neill, M. A., Burns, G. A. P. C. and Hilgetag, C. C. 2002. The PUPS-MOSIX environment: a homeostatic environment for neuro- and bio-informatic applications. In *Neuroscience databases: a practical guide* (Ed. by R. Kötter), pp. 187-202. Kluwer Academic Publishers, Boston.

O'Neill, M.A., Gauld, I.D., Gaston, K.J. and Weeks, P.J.D. 1997. DAISY: an automated invertebrate identification system using holistic vision techniques. First Bionet International Working Group on Automated Taxonomy. University of Wales, Cardiff. pp. 13-22.

O'Neill, M. A., Watson, A. T. and Kitching, I. J. 2005. DAISY: a vision based system for automated insect identification and biodiversity screening. Symposium proceedings: *Algorithmic approaches to the identification problem in systematics.* 19 August 2005, The Natural History Museum, London.

O'Neill, M. A. 2005. Automated insect identification project. http://chasseur.usc.edu/pups/projects/daisy.html (last accessed 21/09/05).

Osborne, J. L., Clark, S. J., Morris, R. J., Williams, I. H., Riley, R. J., Smith, A. D., Reynolds, D. R. and Edwards, A. S. 1999. A landscape-scale study of bumble bee foraging range and constancy, using harmonic radar. *Journal of Applied Ecology,* **36** (4), 519-533.

Overney, G. T. 2004. *Why I like darkfield illumination.* www.microscopy-uk.org.uk/mag/artmar04/godarkfield.html (last accessed 21/09/05).

Oxford University. 2005. *Handwritten Digit Recognition: Nearest Neighbour Classifier.* Online lecture notes. http://www.robots.ox.ac.uk/~dclaus/digits/neighbour.htm (last accessed 21/09/05).

Pajak, M. 2001. Biological Sciences Masters project, University of Oxford. Unpublished.

Patterson, D. J. 2003. Progressing towards a biological names register. *Nature*, **661,** 661.

Pedersen, K. S. and Lee, A. B. 2002. Toward a full probability model of edges in natural images. In *Proceedings of the 7th European Conference on Computer Vision (ECCV),* Copenhagen, May 2002. http://www.itu.dk/~kimstp/papers/jetstat.pdf (last accessed 21/09/05).

Penev, P. S. and Atick, J. J. 1996. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, **7**, 477-500.

Pennington, H. 1999. *GTK+/Gnome application development.* New Riders Publishing, Indianapolis.

Pentland, A., Moghaddam, B. and Starner, T. 1994. View-based and modular eigenspaces for face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington,* 84-91.

Pont, A. C. 1993. Observations on anthophilous Muscidae and other Diptera (Insecta) in Abisko National Park, Sweden. *Journal of Natural History*, **27**, 631-643.

Portable Application Standards Committee. 2005. The Portable Application Standards Committee of the Institute of Electrical and Electronics Engineers homepage. http://www.pasc.org (last accessed 21/09/05).

Potts, S. G., Vulliamy, B., Dafni, A., Ne'eman, G. and Willmer, P. 2003. Linking bees and flowers: how do floral communities structure pollinator communities? *Ecology*, **84,** 2628-2642.

Price, L. D., Chukwuma, F. and Adamczyk, J. J. 2004. Honey bee (Hymenoptera: Apidae) pollen load rate based on pollen grain size. *Journal of Entomological Science,* **39** (4), 677-678.

Proulx, S. R., Promislow, D. E. L. and Phillips, P. C. 2005. Network thinking in ecology and evolution. *Trends in Ecology and Evolution*, **20** (6), 345-353.

Prys-Jones, O. E. and Corbet, S. A. 1991. *Bumblebees*. Slough: Richmond Publishing Co. Ltd.

Ramsdale, C. D. and Ramsdale, P. D. C. 1998. Mosquito genera of the world. CABIKEY: Computer aided biological key. *Medical and Veterinary Entomology,* **12** (4), 438, 439.

Rasnitsyn, A. P., Basibuyuk, H. H. and Quicke, D. L. J. 2004. A basal chalcidoid (Insecta: Hymenoptera) from the earliest cretaceous or latest Jurassic of Mongolia. *Insect Systematics and Evolution,* **35** (2), 123-135.

Rathcke, B. J. and Jules, E. S. 1993. Habitat fragmentation and plant-pollinator interactions. *Current Science,* **65**, 273-277.

Raven, P. H. and Wilson, E. O. 1992. A fifty-year plan for biodiversity surveys. *Science*, **258,** 1099-1100.

Reynolds, A. P., Dicks, J. L., Roberts, I. N., Wesselink, J. J., de la Iglesia, B., Robert, V., Boekhout, T. and Rayward-Smith, V. J. 2003. Algorithms for identification key generation and optimization with application to yeast identification. *Applications of Evolutionary Computing Lecture Notes in Computer Science,* **2611**, 107-118.

Rinderer, T. E., Buco, S. M., Rubink, W. L., Daly, H. V., Stelzer, J. A., Riggio, R. M. and Baptista, F. C. 1993. Morphometric identification of Africanized and European honey bees using large reference populations. *Apidologie*, **24**, 569-585.

Roadknight, C. 1997. Transparent neural network data modelling. PhD thesis, Nottingham Trent University. Unpublished.

Robertson, C. 1929. *Flowers and insects: lists of visitors to four hundred and fifty-three flowers.* C. Robertson, Carlinville, IL.

Rodman, J. E. and Cody, J. H. 2003. The taxonomic impediment overcome: NSF's Partnerships for Enhancing Expertise in Taxonomy (PEET) as a model. *Systematic Biology,* **52** (3), 428-435.

Ronneberger, O., Heimann, U., Schulz, U., Dietze, V., Burkhardt, H. and Gehrig, R. 2000. Automated pollen recognition using gray scale invariants on 3D volume image data. In: *Second European Symposium on Aerobiology* . Vienna, Austria.

Roth, V., Steinhage, V., Schroder, S., Cremers, A. B. and Wittman, D. 1999. Pattern recognition combining de-noising and linear discriminant analysis within a real world application. *Computer Analysis of Images and Patterns. Lecture Notes in Computer Science,* **1689**, 251-258.

Ruppert, E. E. and Barnes, R. D. 1994. *Invertebrate Zoology*. Sixth Edition. Saunders College Publishing, Florida.

Russ, J. C. 1995. *The Image Processing Handbook*. Boca Raton, CRC Press.

Russell, K. N., Do, M. T. and Platnick, N. I. 2005. Introducing SPIDA-web: An automated identification system for biological species. Symposium proceedings: *Algorithmic approaches to the identification problem in systematics.* 19 August 2005, The Natural History Museum, London.

Russell-Smith, T., Stone, G. N. and van Noort, S. 1999. Invertebrate diversity in Mkomazi. In: *Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna* (Ed. by Coe, M. J., McWilliam, N. C., Stone, G. N. and Packer, M. J.), pp. 171-196. London: Royal Geographical Society (with The Institute of British Geographers).

Sala, O. E., Chapin, F. S., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L. F., Jackson, R. B., Kinzig, A., Leemans, R., Lodge, D. M., Mooney, H. A., Oesterheld, M., Poff, N. L., Sykes, M. T., Walker, B. H., Walker, M. and Wall, D. H. 2000. Global biodiversity scenarios for the year 2100. *Science*, **287**, 1770-1774.

São Paulo Declaration on Pollinators. 1999. *Report on the recommendations of the workshop on the conservation and sustainable use of pollinators in agriculture with emphasis on bees.* Brazilian Ministry of the Environment, Brasilia, Brazil.

Sawyer, R. 1981. *Pollen identification for beekeepers*. University College Cardiff Press, Cardiff.

Saxena, R., Zachariah, S. G. and Sanders, J. E. 2002. Processing computer tomography bone data for prosthetic finite element modelling: A technical note. *Journal of Rehabilitation Research and Development,* **39** (5), 609-613.

Schmull, M., Heinrichs, J., Baier, R., Ullrich, D., Wagenitz, G., Groth, H., Hourtocolon, S. and Gradstein, S. R. 2005. The type database at Gottingen (GOET) – a virtual herbarium online. *Taxon*, **54** (1), 251-254.

Scholtz, C. H. and Holm, E. 1985. *Insects of Southern Africa*. Butterworths, Durban.

Schulmeister, S. 2003. Review of morphological evidence on the phylogeny of basal Hymenoptera (Insecta), with a discussion of the ordering of characters. *Biological Journal of the Linnean Society,* **79** (2), 209-243.

Seberg, O., Humphries, C. J., Knapp, S., Stevenson, D. W., Peterson, G., Scharff, N. and Anderson, N. M. 2003. Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends in Ecology and Evolution*, **18**, 63-65.

Sharkey, M. J. and Roy, A. 2002. Phylogeny of the Hymenoptera: a reanalysis of the Ronquist *et al.* 1999 reanalysis, emphasising wing venation and apocritan relationships. *Zoologica Scripta*, **31** (1), 57-66.

Shelly, T. E. and Villalobos, E. 2000. Buzzing bees (Hymenoptera: Apidae, Halictidae) on *Solanum* (Solanaceae): Floral choice and handling time track pollen availability. *Florida Entomologist,* **83** (2), 180-187.

Shepherd, G. M. and Koch, C. 1990. Introduction to synaptic circuits. In: *The Synaptic Organisation of the Brain* (Ed. by Shepherd, G. M). pp. 3-31. Oxford University Press, New York.

Sherry, S. P. 1971. *The Black Wattle*. University of Natal Press, Pietermaritzburg, South Africa.

Smart, I. J. and Knox, R. B. 1979. Aerobiology of grass pollen in the city atmosphere of Melbourne; quantitative analysis of seasonal diurnal changes. *Australian Journal of Botany,* **27**, 317-331.

Smith, K.G.V. and Vockeroth, J.R. 1980. Family Syrphidae. In *Catalogue of the Diptera of the Afrotropical Region.* Ch. 38, pp.488-510. British Museum (Natural History), London.

Solé, R. V. and Montoya. J. M. 2001. Complexity and fragility in ecological networks. *Proceedings of the Royal Society of London B,* **268**, 2039-2045.

Steffan-Dewenter, I., Potts, S. G. and Packer, L. 2005. Pollinator diversity and crop pollination services are at risk. *Trends in Ecology and Evolution,* **20** (12), 651-652.

Sterrenburg, F. A. S. 1997. Unpublished, part of discussion on University of Indiana website. www.indiana.edu/~diatom/pleuro.dis (last accessed 21/09/05).

Stevenson, R. D., Haber, W. A. and Morris, R. A. 2003. Electronic field guides and user communities in the eco-informatics revolution. *Conservation Ecology,* **7** (1), art. no. 3.

Stillman, E. C. and Flenley, J. R. 1996. The needs and prospects for automation in palynology. *Quaternary Science Reviews*, **15,** 1-5.

Stone, G. N., Raine, N. E., Prescott, M. and Willmer, P. G. 2003. Pollination ecology of acacias (Fabaceae, Mimosoideae). *Australian Systematic Botany*, **16,** 103-118.

Stone, G. N., Willmer, P., Rowe, J. A., Nyundo, B. and Abdallah, R. 1999. The pollination ecology of Mkomazi *Acacia* species. In: *Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna* (Ed. by Coe, M. J., McWilliam, N. C., Stone, G. N. and Packer, M. J.), pp. 337-357 London: The Royal Geographical Society (with The Institute of British Geographers).

Sutherland, J. P., Sullivan, M. S. and Poppy, G. M. 1999. The influence of floral character on the foraging behaviour of the hoverfly, *Erisyrphus balteatus. Entomologia Experimentalis et Applicata,* **93**, 157-164.

Tardival, G. M. and Morse, D. R. 1997. The role of the user in computer-based species identification. In: *Information Technology, Plant Pathology and Biodiversity* (Ed. by Scott, P., Bridge, P., Jeffries, P. and Morse, D. R.). Wallingford: CAB International.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. and Vogler, A. P. 2002. DNA points the way ahead in taxonomy. *Nature*, **418,** 478.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. and Vogler, A. P. 2003. A plea for DNA taxonomy. *Trends in Ecology and Evolution,* **18**, 70-74.

Terry, A. M. R. and McGregor, P. K. 2002. Census and monitoring based on individually identifiable vocalizations: the role of neural networks. *Aminal Conservation* **5**, 103-111.

Thomas, B. A., Cleal, C. J. and Barthel, M. 2004. Palaeobotanical applications of incident-light darkfield microscopy. *Palaeontology,* **47** (6), 1641-1645.

Thompson, F. C. 1994. Bar codes for specimen data management. *Insect Collection News,* **9**, 2-4.

Tian, J., He, Y. L., Yang, X., Li, L. and Chen, X. J. 2004.Improving fingerprint recognition performance based on feature fusion and adaptive registration pattern. *Advances in biometric person authentication, proceedings lecture notes in computer science*, **3338**, 57-66.

Tofiliski, A. 2004. DrawWing, a program for numerical description of insect wings. *Journal of Insect Science,* **4**, 17.

Tong, C. S., Yuen, P. C. and Wong, Y. Y. 2002. Dividing snake algorithm for multiple object segmentation. *Optical Engineering,* **41** (12), 3177-3182.

Treloar, W.J. 1994. Automation of paleontology using image processing and pattern recognition techniques. Final report to Massey University, New Zealand.

Treloar, W.J., Taylor, G.E. and Flenley, J.R. 2004. Towards automation of palynology 1: analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning electron microscope images. *Journal of Quaternary Science* **19** (8), 745-754.

Turk, M. and Pentland, A.P. 1991. Eigenfaces for recognition. *Journal of Cognitive Neurosciences* **3**, 71-86.

Vanden Berghe, E. 2005. MASDEA – Marine species database for Eastern Africa. *Indian Journal of Marine Sciences,* **34** (1), 128-135.

Van Hout, R. and Katz, J. 2004. A method for measuring the density of irregularly shaped biological aerosols such as pollen. *Journal of Aerosol Science,* **35**, 1369-1384.

van Noort, S. and Stone, G. N. 1999. Butterflies (Lepidoptera: Hesperioidea and Papilionoidea) of Mkomazi. In: *Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna* (Ed. by Coe, M. J., McWilliam, N. C., Stone, G. N. and Packer, M. J.), pp. 281-298. London: The Royal Geographical Society (with The Institute of British Geographers).

Vázquez, D. P. and Aizen, M. A. 2003. Null model analyses of specialization in plant-pollinator interactions. *Ecology*, **84**, 2493-2501.

VeyTek. 2005. What is a confocal microscope? FAQ webpage. http://www.vaytek.com/FAQ.html (last accessed 21/09/05).

Villanueva-G, R. 2002. Polliniferous plants and foraging strategies of *Apis mellifera* (Hymenoptera: Apidae) in the Yucatán Peninsula, Mexico. *Revista de Biología Tropical,* **50** (3-4).

Walker, D. 1999. *Studying Pollen.* www.microscopy-uk.org.uk/mag/artjul99/pollen.html (last accessed 21/09/05)

Waser, N. M., Chittka, L., Price, M. V., Williams, N. M. and Ollerton, J. 1996. Generalization in pollination systems and why it matters. *Ecology,* **77**, 1043-1060.

Watanabe, M. E. 1994. Pollination worries as honey bees decline. *Science*, **265**, 1170.

Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of "small-world" networks. *Nature,* **393**, 440-442.

Watts, D. J. 2004. The "new" science of networks. *Annual Review of Sociology,* **30**, 243-270.

Weiss, D. G., Maile, W., Wick, R. A. and Steffen, W. 1999. Video Microscopy. In: *Light Microscopy in Biology* (Ed. by Lacey, A. J.). The Practical Approach Series. Oxford University Press, Oxford

Wang, H. B. and Culverhouse, P. F. 2004. The categorisation of similar non-rigid biological objects by clustering local appearance patches. *Lecture Notes in Computer Science*, **3177**, 65-70.

Watson, A. T., O'Neill, M. A. and Kitching, I. J. 2004. Automated identification of live moths (Macrolepidoptera) using Digital Automated Identification SYstem. *Systematics and Biodiversity*, **1** (3), 287-300.

Weber, R. W. 1998. Pollen identification. *Annals of Allergy, Asthma and Immunology,* **80**, 80141-80147.

Weeks, P. J. D., Gauld, I. D., Gaston, K. J. and O'Neill, M. A. 1997. Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research*, **87,** 203-211.

Weeks, P. J. D. and Gaston, K. J. 1997. Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. *Biodiversity and Conservation*, **6**, 263-274.

Weeks, P. J. D., O'Neill, M. A., Gaston, K. J. and Gauld, I. D. 1999a. Automating insect identification: exploring the limitations of a prototype system. *Journal of Applied Entomology*, **123,** 1-8.

Weeks, P. J. D., O'Neill, M. A., Gaston, K. J. and Gauld, I. D. 1999b. Species-identification of wasps using principle component assosiative memories. *Image and Vision Computing*, **17,** 861-866.

Wheeler, Q. D. 2003. Transforming taxonomy. *The Systematist* **22**, 3-5.

Wheeler, Q. D. 2005. Digital innovation and taxonomy's finest hour. Symposium proceedings: *Algorithmic approaches to the identification problem in systematics.* 19 August 2005, The Natural History Museum, London.

White, H. D., Wellman, B. and Nazer, N. 2003. *Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group.* University of Toronto, Canada.

White, J. 1999. *Pollen, its Collection and Preparation for the Microscope*. Northern Biological Supplies, Ipswich.

Williams, I. H. 2003. The Convention on Biological Diversity adopts the International Pollinator Initiative. *Bee World*, **84,** 27-31.

Williams, P. H. 1982. The distribution and decline of British bumble bees (*Bombus* Latr.). *Journal of Apicultural Research*, **21**, 236-245.

Willmer, P. 1993. *Bees, Ants and Wasps: a key to the genera of the British Aculeates*. Field Studies Council Publications, Shrewsbury.

Wilson, E. O. 2003. The encyclopedia of life. *Trends in Ecology and Evolution*, **18,** 77-80.

Witte, H. J. L. 1988. Preliminary research into possibilities of automated pollen counting. *Pollen et Spores* **30**, 111-124.

Wolf, G., Petersen, D., Dietel, M. and Petersen, I. 1998. Telemicroscopy via the internet. *Nature,* **391**, 613-614.

Wythoff, B., Xiao, H. K., Levine, S. P. and Tomellini, S. A. 1991. Computer-assisted infrared identification of vapour-phase mixture components. *Journal of chemical information and computer sciences*, **31** (3), 392-399.

Yoon, J. S., Park, J. C., Jang, S. W and Kim, G. Y. 2004. A shakeable snake for estimation of image contours. *Computational Science and its Applications – ICCSA 2004, pt 1, Lecture Notes in Computer Science,* **3043**, 9-16.

Yu, D. S., Kokko, E. G., Barron, J. R., Schaalje, G. B. and Gowen, B. E. 1992. Identification of Ichneumonid wasps using image analysis of wings. *Systematic Entomology*, **17**, 389-395.

Appendix 1 – Insect specimens donated to The National Museums of Kenya, University of St Andrews, Mpala Research Centre, Connal Eardley, the Natural History Museum, The National Museum of Scotland and the National Museums and Galleries of Wales.

FLIES

Order	Family	Subfamily	Genus	Species	Authority	Identified by	Total
Diptera	Bombyliidae	Bombyliineae	Bombylius	acrophylax	Greathead	D. Greathead	18
Diptera	Bombyliidae	Bombyliineae	Bombylella	ornata	(Wiedemann)	D. Greathead	1
Diptera	Bombyliidae	Bombyliineae	Bombylella	auricoma	(Bezzi)	D. Greathead	1
Diptera	Bombyliidae	Bombyliineae	Parisus (now called E	luteipennis	(Bezzi)	D. Greathead	1
Diptera	Bombyliidae	Bombyliineae	Systoechus	canicapillis	Bowden	D. Greathead	2
Diptera	Bombyliidae	Bombyliineae	Bombomyia	discoidea	(Fabricius)	D. Greathead	1
Diptera	Bombyliidae	Bombyliineae	Gonarthrus	sp.		D. Greathead	1
Diptera	Bombyliidae	Anthracinae	Exoprosopa	punctulata	Macquart	D. Greathead	4
Diptera	Bombyliidae	Anthracinae	Exoprosopa	sp. nov. nr. batracho	Bezzi	D. Greathead	1
Diptera	Bombyliidae	Anthracinae	Heteralonia	katonae	(Bezzi)	D. Greathead	2
Diptera	Bombyliidae	Anthracinae	Spogostylum	incisurale	(Macquart)	D. Greathead	2
Diptera	Bombyliidae	Anthracinae	Spogostylum	ventrale	Bezzi	D. Greathead	1
Diptera	Bombyliidae	Anthracinae	Anthrax	busonicus	Francois	D. Greathead	2
Diptera	Bombyliidae	Anthracinae	Exhyalanthrax	abruptus	(Loew)	D. Greathead	3
Diptera	Bombyliidae	Anthracinae	Exhyalanthrax	flammiger	(Walker)	D. Greathead	6
Diptera	Bombyliidae	Anthracinae	Exhyalanthrax	alliopterus	(Hesse)	D. Greathead	1
Diptera	Bombyliidae	Anthracinae	Thyridanthrax	elegans subperspicill	Bezzi	D. Greathead	1
Diptera	Bombyliidae	Anthracinae	Thyridanthrax	perspicillaris	Loew	D. Greathead	2
Diptera	Bombyliidae	Anthracinae	Villa	paniscoides	Bezzi	D. Greathead	1
Diptera	Bombyliidae	Anthracinae	Villa	sp.	Lioy	D. Greathead	2
Diptera	Bombyliidae	Anthracinae	Petrorossia	sp. fulvipes gp.		D. Greathead	2
Diptera	Bombyliidae	Phthiriinae	Phthiria	crocogramma	Hesse	D. Greathead	1
							56
Diptera	Calliphoridae	Calliphorinae	Chrysomya	regalis	Robineau-Desvoidy	J. Deeming	4
Diptera	Calliphoridae	Calliphorinae	Chrysomya	chloropyga	(Wiedemann)	J. Deeming	31
Diptera	Calliphoridae	Calliphorinae	Lucilia	cuprina	Wiedemann	J. Deeming	1
Diptera	Calliphoridae	Calliphorinae	Hemipyrellia	fernandica	(Macquart)	J. Deeming	10
Diptera	Calliphoridae	Calliphorinae	Phumosia	sp.	Robineau-Desvoidy	J. Deeming	1
Diptera	Calliphoridae	Rhiniinae	Stegosoma	vinculatum	Loew	J. Deeming	3
Diptera	Calliphoridae	Rhiniinae	Pararhyncomyia	cribiformis	Becker	J. Deeming	1
Diptera	Calliphoridae	Rhiniinae	Rhyncomya	trispina	Villeneuve	J. Deeming	1
Diptera	Calliphoridae	Rhiniinae	Rhyncomya	forcipata	Villeneuve	J. Deeming	26
Diptera	Calliphoridae	Rhiniinae	Rhyncomya	pruinosa	Villeneuve	A. Watson	33
Diptera	Calliphoridae	Rhiniinae	Rhyncomya	cassotis	(Walker)	J. Deeming	58
Diptera	Calliphoridae	Rhiniinae	Rhyncomya	soyauxi	Karsch	J. Deeming	3
Diptera	Calliphoridae	Rhiniinae	Rhinia	apicalis	(Wiedemann)	J. Deeming	6
Diptera	Calliphoridae	Rhiniinae	Rhinia	coxendix	(Villeneuve)	J. Deeming	1
Diptera	Calliphoridae	Rhiniinae	Rhinia	nigricornis	Macquart	J. Deeming	6
Diptera	Calliphoridae	Rhiniinae	Isomyia	tristis	(Bigot)	N. Wyatt	5
Diptera	Calliphoridae	Rhiniinae	new genus			J. Deeming	4
Diptera	Calliphoridae	Rhiniinae				A.Pont	6
Diptera	Calliphoridae	Calliphorinae	Bengalia	peuhi	Villeneuve	J. Deeming	2
Diptera	Calliphoridae	Calliphorinae	Bengalia	depressa	Walker	J. Deeming	1
Diptera	Calliphoridae	Calliphorinae	Hemigymnochaeta	sp.		J. Deeming	1
Diptera	Calliphoridae	Calliphorinae				A.Pont / J.Deeming	3
							203
Diptera	Muscidae	Coenosiinae	Coenosia	simulans	Paterson	A. Pont	7
Diptera	Muscidae	Muscinae	Musca	aethiops gabonensis	Stein	A. Pont	4
Diptera	Muscidae	Muscinae	Musca	biseta	Hough	A.Pont	1
Diptera	Muscidae	Muscinae	Musca	domestica calleva	Walker	A.Pont	1
Diptera	Muscidae	Muscinae	Musca	lusoria	Wiedemann	A. Pont	65
Diptera	Muscidae	Muscinae	Musca	sp. nr. lusoria		A. Pont	2
Diptera	Muscidae	Muscinae	Musca	nevilli	Kleynhans	A. Pont	9
Diptera	Muscidae	Muscinae	Musca	alpesa	Walker	A. Pont	3
Diptera	Muscidae	Muscinae	Musca	conducens	Walker	A. Pont	3
Diptera	Muscidae	Muscinae	Musca	tempestatum	Bezzi	A. Pont	1
Diptera	Muscidae	Muscinae	Musca	confiscata	Speiser	A. Pont	1
Diptera	Muscidae	Muscinae	Musca	lasiophthalna	Thompson	A. Pont	1
Diptera	Muscidae	Muscinae	Musca	xanthomelaena	Wiedemann	A.Pont	2
Diptera	Muscidae	Muscinae	Musca	sp. indet.	Linnaeus	A. Pont	2
Diptera	Muscidae	Muscinae	Neomyia	albigena	Stein	A.Pont	1
Diptera	Muscidae	Muscinae	Mitroplatia	smaragdina	Seguy	A. Pont	3
Diptera	Muscidae	Muscinae	Pyrellia	scintillans	Bigot	A. Pont	1
Diptera	Muscidae	Muscinae	Pyrellia	sp. nov.	Robineau-Desvoidy	A.Pont	14
Diptera	Muscidae	Phaoniinae	Helina	coniformis	Stein	A. Pont	3
Diptera	Muscidae	Phaoniinae	Atherigona	sp.	Rondani	A. Pont	1
Diptera	Muscidae	Limnophorinae	Lispe	leucospila	Wiedemann	A. Pont	3
Diptera	Muscidae	Stomoxyinae	Stomoxys	niger	Macquart	A. Pont	1

FLIES (continued)

Order	Family	Subfamily	Genus	Species	Authority	Identified by	Total
Diptera	Sarcophagidae	Miltogramminae	?Hoplacephala	spp.	Macquart	A. Watson	2
Diptera	Sarcophagidae	Miltogramminae	?Hilarella		Rondani	J. Deeming/A. Wats	6
Diptera	Sarcophagidae	Miltogramminae	?Senotainia		Zumpt	A. Watson	2
Diptera	Sarcophagidae	Miltogramminae				A.Pont	1
Diptera	Sarcophagidae	Sarcophaginae	Sarcophaga	sensu lato	Meigen	J. Deeming	1
							12
Diptera	Stratiomyidae					A.Pont	1
Diptera	Tachinidae	Dexiinae	Billaea	spp.	Robineau-Desvoidy	N. Wyatt	2
Diptera	Tachinidae	Tachininae	Peleteria	spp.	Robineau-Desvoidy	N. Wyatt	4
Diptera	Tachinidae	Tachininae	Dejeania	bombylans	Fabricius	A. Whittington	35
Diptera	Tachinidae	Tachininae	Linnaemya	neavei	Curran	N. Wyatt	1
Diptera	Tachinidae	Tachininae	Nemoraea	sp.	Robineau-Desvoidy	N. Wyatt	2
Diptera	Tachinidae	Goniinae	Blepharella	spp.	Macquart	N. Wyatt	1
Diptera	Tachinidae	Goniinae	Palexorista	spp.	Townsend	N. Wyatt	3
Diptera	Tachinidae	Goniinae	Pales	sp.	Robineau-Desvoidy	N. Wyatt	2
Diptera	Tachinidae	Goniinae	Exorista	spp.	Meigen	N. Wyatt	3
Diptera	Tachinidae	Goniinae	Kiniatiliops	sp.	Meshil	N. Wyatt	1
Diptera	Tachinidae	Goniinae	Pexopsis	sp.	Brauer & Bargenstamm	N. Wyatt	1
Diptera	Tachinidae	Goniinae	Carcelia	sp.	Robineau-Desvoidy	N. Wyatt	1
Diptera	Tachinidae	Goniinae	Aplomya	sp.	Robineau-Desvoidy	N. Wyatt	1
Diptera	Tachinidae	Goniinae				N. Wyatt	3
Diptera	Tachinidae					A.Pont	2
							62
Diptera	Conopidae		unknown			A. Watson	4
Diptera	Chloropidae	Oscinellinae	new genus		new genus	J. Deeming	2
Diptera	Syrphidae	Milesiinae	Eristalinus	?nr. taeniops		A. Watson	7
Diptera	Syrphidae	Milesiinae	Eristalinus	nr taeniops spp.	(Wiedemann)	A. Whittington	11
Diptera	Syrphidae	Milesiinae	Phytomia	natalensis	(Macquart)	A. Whittington	2
Diptera	Syrphidae	Milesiinae	Senaspis	haemorrhoa	(Gerstaecker)	A. Whittington	2
Diptera	Syrphidae	Milesiinae	Phytomia	incisa	(Wiedemann)	A. Whittington	31
Diptera	Syrphidae	Syrphinae	Ischiodon	aegyptius	(Wiedemann)	A. Whittington	3
Diptera	Syrphidae	Syrphinae	Allobaccha	sapphirina	(Wiedemann)	A. Whittington	1
Diptera	Syrphidae	Milesiinae	Ceriana	sp. only marginate known from Kenya		A. Whittington	6
Diptera	Syrphidae		Chrysotoxum	continuum		A. Whittington	4
Diptera	Syrphidae		Eupeodes	corollae	(Fabricius)	A. Whittington	3
							70
Diptera	Asilidae	Asilinae	Alcimus	spp.	Loew	A. Watson	2
Diptera	Asilidae	Asilinae	Hippomachus	sp.	Engel	J. Londt	1
Diptera	Asilidae	Asilinae	Neolophonotus	sp.	Engel	J. Londt	1
Diptera	Asilidae	Asilinae	Promachus	sp.	Loew	J. Londt	1
Diptera	Asilidae	Dasypogoninae	Scylaticus	sp.	Loew	J. Londt	1
Diptera	Asilidae	Laphriinae	Laxenecera	sp.	Macquart	J. Londt	1
Diptera	Asilidae	Leptogastrinae	Lasiocnemus	sp.	Loew	J. Londt	1
Diptera	Asilidae			3 spp.		A. Watson	4
							12
Diptera	Anthomyidae					A. Pont/unknown (N.	4
Diptera	Anthomyidae		Anthomyia			A.Pont	1
							5
Diptera	Theridae					unknown (N. Wyatt?	1
Diptera	Anthomyiidae/Theridae					unknown (N. Wyatt)	1
Diptera	Milichiidae	Milichinae	Milichia	sp.		J. Deeming	2

562

BEETLES

Order	Family	Subfamily	Genus	Species	Authority	Identified by	No.
Coleoptera	**Bruchidae**		Various			M. Brendell	1
Coleoptera	**Buprestidae**		Anthaxia (or sim.)	sp.		K. Baldock	2
Coleoptera	**Buprestidae**			sp.1		A. Watson	3
Coleoptera	**Buprestidae**			sp.2		A. Watson	1
Coleoptera	**Buprestidae**			sp.3		A. Watson	1
Coleoptera	**Buprestidae**			sp. 4		A. Watson	1
							8
Coleoptera	**Cerambycidae**	Cerambycinae	Helymaeus	insignis	Gerstaecker	A. Watson	3
Coleoptera	**Cerambycidae**	Cerambycinae	Closteromerus	claviger	Dalm.	A. Watson	10
Coleoptera	**Cerambycidae**	Cerambycinae	Promeces	suturalis	Harold	A. Watson	6
							19
Coleoptera	**Cleridae**					M. Brendell	1
Coleoptera	**Coccinellidae**		Cheilomenes	propinqua	Muls.	A. Watson	1
Coleoptera	**Coccinellidae**		Cheilomenes	aurora	Gerstaecker	A. Watson	1
							2
Coleoptera	**Chrysomelidae**	Galerucinae	Monolepta	ephippiata	Gerstaecker	A. Watson	2
Coleoptera	**Chrysomelidae**	Galerucinae	Monolepta	sp.1		A. Watson	1
Coleoptera	**Chrysomelidae**	Galerucinae	Monolepta	gossypii	Bryant	A. Watson	2
Coleoptera	**Chrysomelidae**	Galerucinae	Megalognatha	meruensis		S. Shute	34
Coleoptera	**Chrysomelidae**	Galerucinae	type 4			S. Shute (W&B)	2
Coleoptera	**Chrysomelidae**	Galerucinae	type 6			S. Shute (W&B)	2
Coleoptera	**Chrysomelidae**	Galerucinae	sim to type 9			A. Watson	1
Coleoptera	**Chrysomelidae**	Galerucinae	ATW sp.1	sp.1		A. Watson	1
Coleoptera	**Chrysomelidae**	Galerucinae	ATW sp.3	sp.3		A. Watson	1
Coleoptera	**Chrysomelidae**	Galerucinae	ATW sp. 4	sp. 4		A. Watson	1
Coleoptera	**Chrysomelidae**	Alticinae	type 1			S. Shute/K. Baldock	15
Coleoptera	**Chrysomelidae**	Clytrinae	?Gynandropthalma	possibly 2+ species (smaller/bigger)., 3 po		S. Shute/K. Baldock	14
Coleoptera	**Chrysomelidae**	Clytrinae		sp.1		S. Shute/A. Watson	1
Coleoptera	**Chrysomelidae**	Cryptocephalinae	sp. 4			S. Shute/W&B	3
Coleoptera	**Chrysomelidae**	Cryptocephalinae	sp. 5			A. Watson	1
Coleoptera	**Chrysomelidae**	Eumolpinae				A. Watson	1
							82
Coleoptera	**Curculionidae**		Nematocerus	sp. (F)	Reiche	R.T. Thompson	1
Coleoptera	**Curculionidae**		genus indet.			R.T. Thompson	1
Coleoptera	**Curculionidae**		Polyclaeis	maculatus	Boheman	R.T. Thompson	1
							3
Coleoptera	**Lagriidae**	Lagriinae	Lagria	villosa	(Fabricius)	A. Watson	15
							15
Coleoptera	**Lycidae**		Lycus	trabeatus	Guer	A. Watson	3
Coleoptera	**Lycidae**		Lycus	serenus	Kln	A. Watson	9
Coleoptera	**Lycidae**		Lycus	constrictus	Fahr	A. Watson	1
Coleoptera	**Lycidae**		Lycus	sp.1	(Fabricius)	A. Watson	17
							30
Coleoptera	**Meloidae**		Coryna	?apicornis	Guer.	W&B	51
Coleoptera	**Meloidae**		Coryna	?parenthesis	Gerstaecker	W&B	3
Coleoptera	**Meloidae**		Coryna	?chevrolati	Beauc.	W&B	11
Coleoptera	**Meloidae**		Coryna	?arussima	Gestro.	W&B	6
Coleoptera	**Meloidae**		Coryna	sp. 2		A. Watson	1
							72
Coleoptera	**Melyridae**			spp.		A. Watson	4
Coleoptera	**Nitidulidae**		unknown			M. Brendell	1
Coleoptera	**Prionoceridae**		Idgia	sp.		W&B	25
Coleoptera	**?Rhipiphoridae**			spp.		M. Brendell	3
Coleoptera	**Scarabaeidae**	Cetoniinae	Dichista	cincta	de Geer	W&B	14
Coleoptera	**Scarabaeidae**	Cetoniinae	Pachnoda	elegantissima	Gory & Percheron	W&B	16
Coleoptera	**Scarabaeidae**	Cetoniinae	Pachnoda	ephippiata	Gerstaecker	W&B	2

BEES

Order	Family	Subfamily	Genus	Subgenus	Species	Authority	Identified by	No.
Hymenoptera	Apidae	Apinae	Amegilla		rapida	(Smith)	C. Eardley	1
Hymenoptera	Apidae	Apinae	Amegilla		fallax	(Smith)	C. Eardley	31
Hymenoptera	Apidae	Apinae	Amegilla		fallax	(Smith)	A. Watson	4
Hymenoptera	Apidae	Apinae	Amegilla		calens	(Lepeletier)	C. Eardley	2
Hymenoptera	Apidae	Apinae	Amegilla		calens	(Lepeletier)	A. Watson	1
Hymenoptera	Apidae	Apinae	Amegilla		acraensis	(Fabricius)	C. Eardley	3
Hymenoptera	Apidae	Apinae	Amegilla		cymatilis	Eardley	C. Eardley	5
Hymenoptera	Apidae	Apinae	Amegilla		caelestina	(Cockerell)	C. Eardley	1
Hymenoptera	Apidae	Apinae	Apis		mellifera	Linnaeus	C. Eardley	19
Hymenoptera	Apidae	Apinae	Braunsapis		bouyssoui	(Vachal)	C. Eardley	3
Hymenoptera	Apidae	Apinae	Liotrigona		parvula	Darchen	C. Eardley	2
Hymenoptera	Apidae	Apinae	Pachymelus	Pachymelopsis	reichardti	Stadelmann	C. Eardley	5
Hymenoptera	Apidae	Apinae	Plebeina		hildebranti	(Friese)	C. Eardley	42
Hymenoptera	Apidae	Apinae	Tetralonia	(Eucara)	boharti	Eardley	C. Eardley	1
Hymenoptera	Apidae	Apinae	Tetralonia		nigropilosa	Friesa	C. Eardley	4
Hymenoptera	Apidae	Apinae	Tetraloniella		pulverosa	(Friese)	C. Eardley	3
Hymenoptera	Apidae	Apinae	Tetraloniella		sp. 2		C. Eardley	4
Hymenoptera	Apidae	Apinae	Thyreus		calceatus	(Vachal)	C. Eardley	3
Hymenoptera	Apidae	Apinae	Thyreus		brachyaspis	(Cockerell)	C. Eardley	1
Hymenoptera	Apidae	Apinae	Thyreus		axillaris	(Vachal)	C. Eardley	1
Hymenoptera	Apidae	Apinae	Thyreus		delumbatus	(Vachal)	C. Eardley	2
Hymenoptera	Apidae	Apinae	Thyreus		pretextus	(Vachal)	C. Eardley	1
Hymenoptera	Apidae	Apinae	Thyreus		sp.		A. Watson	1
Hymenoptera	Apidae	Xylocopinae	Ceratina	(Ctenoceratina)	moerenhouti	(Vachal)	C. Eardley	4
Hymenoptera	Apidae	Xylocopinae	Ceratina	(Pithitis)	sp.		C. Eardley	2
Hymenoptera	Apidae	Xylocopinae	Ceratina		sp.		A. Watson	14
Hymenoptera	Apidae	Xylocopinae	Macrogalea		candida	(Smith)	C. Eardley	62
Hymenoptera	Apidae	Xylocopinae	Xylocopa		somalica	Magretti	C. Eardley	35
Hymenoptera	Apidae	Xylocopinae	Xylocopa	Xylomelissa	hottentotta	Smith	C. Eardley	2
Hymenoptera	Apidae	Xylocopinae	Xylocopa		flavorufa	(GeGeer)	C. Eardley	7
Hymenoptera	Apidae	Xylocopinae	Xylocopa		calens	Lepeletier	C. Eardley	1
Hymenoptera	Apidae	Xylocopinae	Xylocopa		spp. (male)		C. Eardley	2
								269

Hymenoptera	Halictidae	Halictinae	Halictus	(Seladonia)	sp. 1		C. Eardley	1
Hymenoptera	Halictidae	Halictinae	Halictus	(Seladonia)	sp. 2		C. Eardley	6
Hymenoptera	Halictidae	Halictinae	Halictus	(Seladonia)	sp. 3		C. Eardley	1
Hymenoptera	Halictidae	Halictinae	Halictus	(Seladonia)	sp. 4		C. Eardley	2
Hymenoptera	Halictidae	Halictinae	Lasioglossum	(Dialictus)	sp. 1		C. Eardley	8
Hymenoptera	Halictidae	Halictinae	Lasioglossum	(Dialictus)	sp. 2		C. Eardley	2
Hymenoptera	Halictidae	Halictinae	Lasioglossum	(Dialictus)	sp. 3		C. Eardley	2
Hymenoptera	Halictidae	Halictinae	Lasioglossum	(Dialictus)	sp. 4		C. Eardley	1
Hymenoptera	Halictidae	Halictinae	Lasioglossum		sp. 5		C. Eardley	1
Hymenoptera	Halictidae	Halictinae	Lasioglossum		spp.		A. Watson	2
Hymenoptera	Halictidae	Halictinae	Patellapis	(Zonalictus)	sp. 1		C. Eardley	9
Hymenoptera	Halictidae	Halictinae	Patellapis	(Zonalictus)	sp. 2		A. Watson	7
Hymenoptera	Halictidae	Halictinae	Patellapis	(Zonalictus)	sp. 2		C. Eardley	1
Hymenoptera	Halictidae	Halictinae	Patellapis	(Chaetalictus)	sp.		C. Eardley	2
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 1		C. Eardley	1
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 2		C. Eardley	2
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 9		C. Eardley	1
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 10		C. Eardley	1
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 11		C. Eardley	1
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 12		C. Eardley	1
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(Lipotriches)	sp. 13		C. Eardley	4

BEES (continued)

Order	Family	Subfamily	Genus	Subgenus	Species	Authority	Identified by	No.
Hymenoptera	Halictidae	Nomiinae	Lipotriches	(?Trinomia)	sp. 1		C. Eardley	2
Hymenoptera	Halictidae	Nomiinae	Lipotriches		spp.		C. Eardley / AW	11
Hymenoptera	Halictidae	Nomiinae	Nomia	(Acunomia)	sp. 1		C. Eardley	1
Hymenoptera	Halictidae	Nomiinae	Nomia		sp.		A. Watson	1
Hymenoptera	Halictidae	Nomiinae	Pseudapis	(Pseudapis)	sp.		C. Eardley	19
Hymenoptera	Halictidae	Nomiinae	Pseudapis	(Pseudapis)	sp.		A. Watson	5
Hymenoptera	Halictidae				spp.		A. Watson	14
								109
Hymenoptera	Colletidae	Colletinae	Colletes		sp. 1		C. Eardley	15
Hymenoptera	Colletidae	Colletinae	Colletes		sp.3		C. Eardley	2
Hymenoptera	Colletidae	Colletinae	Colletes		sp.		A. Watson	1
Hymenoptera	Colletidae	Hylaeinae	Hylaeus		sp. 1		C. Eardley	1
Hymenoptera	Colletidae	Hylaeinae	Hylaeus		sp. 2		C. Eardley	2
Hymenoptera	Colletidae	Hylaeinae	Hylaeus		sp. 3		C. Eardley	2
Hymenoptera	Colletidae	Hylaeinae	Hylaeus		sp. 4		C. Eardley	1
								24
Hymenoptera	Megachilidae	Megachilinae	Anthidiellum	(Pycnanthidium	sp.		A. Watson	1
Hymenoptera	Megachilidae	Megachilinae	Heriades	(Heriades)	sp. 1		C. Eardley	9
Hymenoptera	Megachilidae	Megachilinae	Heriades	(Heriades)	sp. 2		C. Eardley	2
Hymenoptera	Megachilidae	Megachilinae	Heriades		sp. (damaged)		C. Eardley	1
Hymenoptera	Megachilidae	Megachilinae	Heriades		spp.		A. Watson	8
Hymenoptera	Megachilidae	Megachilinae	Megachile		discolor	Smith	C. Eardley	1
Hymenoptera	Megachilidae	Megachilinae	Megachile		lg bl&wh		A. Watson	1
Hymenoptera	Megachilidae	Megachilinae	Megachile		gratiosa	Gerstaecker	C. Eardley	5
Hymenoptera	Megachilidae	Megachilinae	Megachile		lg rusty		A. Watson	26
Hymenoptera	Megachilidae	Megachilinae	Megachile		? lg rusty		A. Watson	1
Hymenoptera	Megachilidae	Megachilinae	Megachile	(Creightonella)	sp. 2		C. Eardley	3
Hymenoptera	Megachilidae	Megachilinae	Megachile		sp. 1		C. Eardley	7
Hymenoptera	Megachilidae	Megachilinae	Megachile		sp. 2		C. Eardley	6
Hymenoptera	Megachilidae	Megachilinae	Megachile		sp. 4		C. Eardley	9
Hymenoptera	Megachilidae	Megachilinae	Megachile		sp. 5		C. Eardley	2
Hymenoptera	Megachilidae	Megachilinae	Megachile		sp. 6		C. Eardley	9
Hymenoptera	Megachilidae	Megachilinae	Megachile		sm b&wh spp.		A. Watson	69
								160
								562

BUTTERFLIES & MOTHS

Order	Family	Subfamily	Genus	Species	Authority	Identified by	No.
Lepidoptera	Arctiidae		Amata	nr. chrysozona	Hmps.	A. Watson	4
Lepidoptera	Hesperiidae	Pyrginae	Eretis	umbra	(Trimen)	A. Watson	1
Lepidoptera	Hesperiidae	Pyrginae	Sarangesa	phidyle	(Walker)	A. Watson	1
							2
Lepidoptera	Lycaenidae	Polyommatinae	Azanus	jesous	(Guerin-Menev	A. Watson	3
Lepidoptera	Lycaenidae	Polyommatinae	Azanus	moriqua	(Wallengren)	A. Watson	3
Lepidoptera	Lycaenidae	Polyommatinae	Euchrysops	subpallida	Bethune-Baker	A. Watson	8
Lepidoptera	Lycaenidae	Polyommatinae	Leptotes	pirithous	(Linnaeus)	A. Watson	2
Lepidoptera	Lycaenidae	Polyommatinae	Zizeeria	knysna	(Trimen)	A. Watson	2
Lepidoptera	Lycaenidae	Theclinae	Axiocerses	harpax	(Fabricius)	A. Watson	2
Lepidoptera	Lycaenidae		Euchrysops	sp.	Butler	A. Watson	1
							21
Lepidoptera	Noctuidae					A. Watson	1
Lepidoptera	Nymphalidae	Acraeinae	Acraea	cabira	Hopffer	A. Watson	3
Lepidoptera	Nymphalidae	Acraeinae	Acraea	neobule	Doubleday	A. Watson	1
Lepidoptera	Nymphalidae	Acraeinae	Pardopsis	punctatissima	(Boisduval)	A. Watson	4
Lepidoptera	Nymphalidae	Satyrinae	Neocoenyra	duplex	Butler	A. Watson	6
Lepidoptera	Nymphalidae	Satyrinae	Neocoenyra	gregorii	Butler	A. Watson	1
Lepidoptera	Nymphaliidae		Danaus	sp.	Klug	A. Watson	1
							16
Lepidoptera	Papilionidae		Papilio	demodocus	Esper	A. Watson	2
Lepidoptera	Pieridae	Pierinae	Colotis	antevippe	(Boisduval)	A. Watson	1
Lepidoptera	Pieridae	Pierinae	Colotis	aurigineus	(Butler)	A. Watson	4
Lepidoptera	Pieridae	Pierinae	Colotis	celimene	(Lucas)	A. Watson	1
Lepidoptera	Pieridae	Pierinae	Colotis	danae	(Fabricius)	A. Watson	1
Lepidoptera	Pieridae	Pierinae	Colotis	evagore	(Klug)	A. Watson	1
Lepidoptera	Pieridae	Pierinae	Dixeia	orbona	(Geyer)	A. Watson	3
Lepidoptera	Pieridae	Pierinae	Eronia	leda	(Boisduval)	A. Watson	1
Lepidoptera	Pieridae	Pierinae	Eurema	brigitta	(Stoll)	A. Watson	6
Lepidoptera	Pieridae	Pierinae	?Comma	brigitta		A. Watson	1
							19
Lepidoptera	Sessiidae					A. Watson (NM	1
Lepidoptera	Scythridae					A. Watson (NM	2
Lepidoptera	Satyridae		Neocoenyra	duplex	Butler	A. Watson	1
Lepidoptera	Sphingidae		Cephonodes	hylas	(Linnaeus)	Ian Kitching	1
Lepidoptera	Sphingidae		Nephele	vau	(Walker)	Ian Kitching	3
							4
						TOTAL	73

BUGS

Order	Family	Subfamily	Genus	Species	Authority	Identified by	No.
Hemiptera	Alydidae		Tupalus	maculates	Distant	A. Watson	1
Hemiptera	Coreidae		Leptoglossus	australis	Fabricius	A. Watson	1
Hemiptera	Pyrrhochoridae		Dysdocus	cardinalis	Gerstaecker	A. Watson	4
Hemiptera	Pyrrhochoridae		Dysdocus	nigrofasciatus	Stal	A. Watson	25
Hemiptera	Pyrrhochoridae		Dysdocus	pretiosus	Distant	A. Watson	1
Hemiptera	Pyrrhochoridae		?Dysdocus	spp. (juv)		A. Watson	11
Hemiptera	Reduviidae		Hediocorus	tibialis	Stal	A. Watson	1
						TOTAL	44

Appendix 2 – Pollen methods

Acetolysis

1) Select 10 samples of polleniferous material and number in sequence.
2) Label centrifuge tubes 1-10 (with permanent marker at top and bottom of tube).

For each sample

3) Pipette off excess ethanol.
4) Pour into watch glass and tease matter apart with dissecting needles.
5) Set up equipment according to the Fig.1.

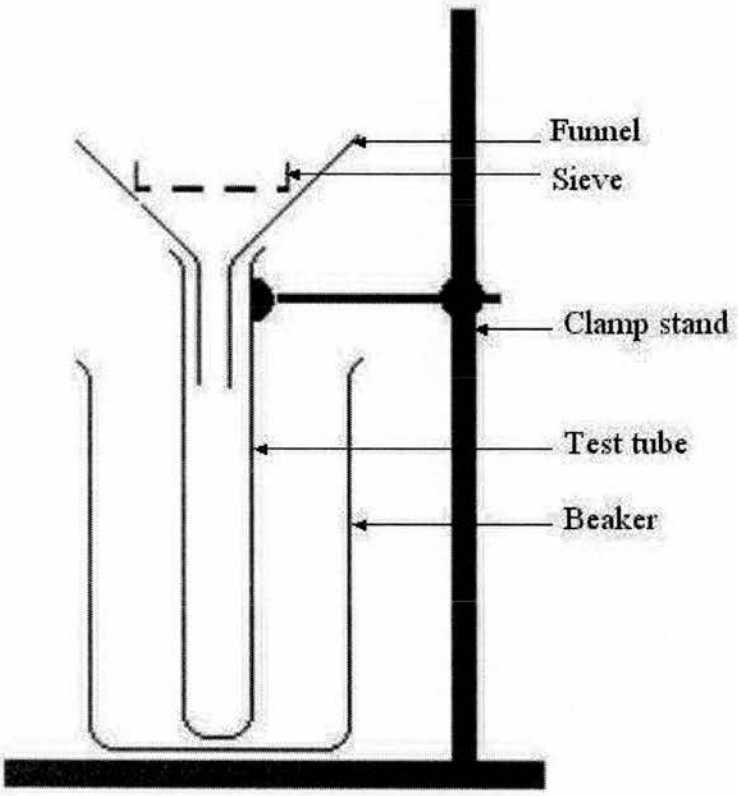


Fig. 1 - Equipment for stages 6 – 7.

6) Pour contents of watch glass into 125 micron sieve and macerate with glass rod.
7) Wash through with distilled water until the tube is almost full.
8) Centrifuge tubes for 5 minutes at 3000 rpm.
9) Pour liquid out of tube to leave pollen in tube bottom.
10) Pipette in Glacial Acetic Acid (GAA) until the tube is at least 1/3 full.

11) Centrifuge the tube for 4 minutes at 3000 rpm.

12) Decant GAA in sink.

13) Prepare acid mixture (in fume cupboard), enough for 10-12 samples.

   A) Mark 13.5ml on a dry 30ml measuring cylinder.

   B) Measure 13.5ml of acetic anhydride then pour into a dry beaker.

   C) Add a pipette-full of the sulphuric acid, drop by drop and stirring.

14) Pipette acid mixture into pollen tube until at least 1/3 full.

15) Transfer tubes into heat block set at 110°.

16) After 2 minutes, stir the tube to agitate if pollen has settled to bottom.

17) After a further 3 minutes remove from heat block.

18) Centrifuge at 3000 rpm for 5 min.

19) Decant liquid into dry beaker and dispose of into a water-filled sink.

20) Fill the tube with distilled water.

21) Centrifuge at 3000rpm for 5 min.

22) Decant water.

23) Fill the tube with distilled water.

24) Centrifuge at 3000rpm for 5 min.

25) Decant water.

26) Mix 20ml distilled water with 20ml glycerol in a 50ml measuring cylinder, pour into sealed bottle and shake well.

27) Use wide-tipped pipette to fill tubes ½ full of 50% glycerol.

28) Centrifuge at 3000 rpm for 5 min.

29) Decant most of 50% glycerol and place tubes upside down in rack, draining onto tissue paper. Leave for 45 min.

Mounting clean pollen

1) Scratch identification code onto microscope slide using a diamond tip pen rod.
2) Dip small piece of paper towel in ethanol and clean the slide thoroughly, wiping away shards of glass.
3) Cut off a small piece of glycerine jelly and place on a microscope slide. From this cut several smaller pieces approximately 1-2mm square.
4) Roll a 1mm diameter string of plastacine and place on another slide.
5) Pick up a small piece of jelly on the point of a mounted needle. Wipe it around the inside of the pollen tube until some pollen has been picked up. Place the jelly on the microscope slide (on guide line) and clean mounted needle thoroughly before using with next sample.
6) Use a scalpel to cut two small pieces of plastacine (approx 1mm x 1mm) from the string. Place these on either side of the jelly piece, about 5mm from it.
7) If there is plenty of pollen available then repeat stages 5 and 6, using the other end of the microscope slide.
8) Place on hotplate to melt.
9) Stir sample to distribute pollen grains evenly, and then gently place circular 19ml coverslip on top.
10) Press down cover slipe on the plasticine pieces with a wooden cocktail stick.
11) Melt paraffin wax (60° melting point) and use a pipette to place a drop alongside the coverslip; capillary action will draw the wax under it. Add a second drop on the other side of the slide if necessary.
12) Wipe off excess wax or scalpel away once set.