Computer Aided Identification of Biological Specimens Using Self-Organizing Maps

by

Eileen J. Dean

Submitted in fulfillment of the requirements for the degree

Magister Scientiae (Computer Science)

in the Faculty of Engineering, Built Environment and Information Technology

University of Pretoria,

Pretoria, South Africa

April 2010

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Computer Aided Identification of Biological Specimens Using Self-Organizing Maps

by

Eileen J. Dean

Abstract

For scientific or socio-economic reasons it is often necessary or desirable that biological material be identified. Given that there are an estimated 10 million living organisms on Earth, the identification of biological material can be problematic. Consequently the services of taxonomist specialists are often required. However, if such expertise is not readily available it is necessary to attempt an identification using an alternative method. Some of these alternative methods are unsatisfactory or can lead to a wrong identification. One of the most common problems encountered when identifying specimens is that important diagnostic features are often not easily observed, or may even be completely absent. A number of techniques can be used to try to overcome this problem, one of which, the Self Organizing Map (or SOM), is a particularly appealing technique because of its ability to handle missing data. This thesis explores the use of SOMs as a technique for the identification of indigenous trees of the *Acacia* species in KwaZulu-Natal, South Africa. The ability of the SOM technique to perform exploratory data analysis through data clustering is utilized and assessed, as is its usefulness for visualizing the results of the analysis of numerical, multivariate botanical data sets. The SOM's ability to investigate, discover and

interpret relationships within these data sets is examined, and the technique's ability to identify tree species successfully is tested. These data sets are also tested using the C5 and CN2 classification techniques. Results from both these techniques are compared with the results obtained by using a SOM commercial package. These results indicate that the application of the SOM to the problem of biological identification could provide the start of the long-awaited breakthrough in computerized identification that biologists have eagerly been seeking.

Keywords: Self-Organizing Map (SOM), Unsupervised Learning Algorithm, Artificial Neural Network (ANN), Artificial Intelligence (AI), Clustering and Visualization, Biological Identification, Tree Identification, Biological Keys, Botanical Identification, *Acacia* species

Supervisor: Professor A. P. Engelbrecht, Department of Computer Science, University of Pretoria

Co-supervisor: Professor A. Nicholas, Department of Botany, University of KwaZulu-Natal

Degree: *Magister Scientiae*

Department of Computer Science, University of Pretoria
Submitted in fulfillment of the requirements for the degree of
Magister Scientiae (Computer Science)

Acknowledgements

I would like to thank the following people for their assistance during the production of this thesis:

- Professor A.P. Engelbrecht, my thesis supervisor, for his insight and motivation;
- Professor A. Nicholas, my thesis co-supervisor, for his enthusiasm, support, understanding and guidance and for always finding a minute that often turned into longer invaluable discussions;
- Professor Eyono Obono, a DUT colleague, for the hours of discussions given without reserve which gave me the courage to continue when I felt over-whelmed;
- Allan for the many hours of support and help tirelessly reading, correcting and commenting;
- Peter, my brother, for all his comments and suggestions, particularly on the mathematical aspect of the thesis;
- My colleagues: Rose Quilling for sparing her invaluable time for encouragement, advice and support; Nareen Gonsalves for reading my work and patiently listening to the difficulties I experienced;
- The NRF for their financial support and understanding – Cheryl and Shirley, even though you have moved on, a big thank you – yes I have eventually finished;
- DUT for their financial support.

"If the names of things are neglected

the knowledge of them will also perish"

— *Linnaeus 1751*

Dedication

To the memory of my parents:

My father, Edmund (Ted) Scott Dean, for whom an education and knowledge was all important, and who encouraged and supported all his children in their attempts to gain academic qualifications.

AND

My mother, Mary (Molly) Vere Dean, who throughout her life gave unending love, support and understanding to me - thank you for your loyalty.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

As in all of the biological sciences, botany is extremely rich in information. In the last 100 years knowledge about plants has expanded exponentially and access to much of this information is obtained through scientific plant names. Consequently, in order to utilize this information, field researchers, ecologists, conservationists, ethnobotanists, students, para-botanists and even interested laypersons need to know the names of the plants they encounter in the wild or in their gardens.

The process by which plant names are obtained in the field is known as identification. Within this context, identification involves recognizing, selecting or associating closely the characteristics of an unknown object with another object of known identity. By far the greatest diagnostic information which a biologist uses when identifying biological specimens is based on gross morphology. This is because macromorphology is more practical to use than micromorphology or molecular data, especially when in the field [72, 160]. The macroscopic identification of a botanical tree is often done by taking material (leaf, bark, flowers, seed or root) from the unknown tree and comparing its characteristics with a list of correlated characteristics of known trees. In addition, the location, climatic conditions and general form of a tree should also be taken into consideration. Each tree Species (which consists of varied individuals) displays a range of characteristics which form a pattern specific to that species. Sometimes these species' patterns are distinct, making identification easy, but sometimes they overlap, making identification difficult. In addition, biological recognition is an elementary but fundamental type of identification [174].

By far the best solution for identifying botanical material is to have an expert taxonomist who can perform the identification drawing on intuitive and sensory recognition as well as professional expertise. However, it is often the case that no expert is available and so alternative aids to identification have to be found. This

research thesis investigates the problems involved in identifying botanical samples and attempts to find a computerized solution which could increase a novice's chance of performing a successful identification. Many techniques are available; the most sophisticated being those available through the fields of Artificial Intelligence (AI) [44, 154, 182] and DNA barcoding [61, 62, 81, 88, 89, 93, 94, 149].

Section 1.1 discusses why it is thought that AI techniques and, in particular, the Self-organizing map (SOM) could prove useful in overcoming some of the difficulties experienced with biological identification. The objectives of this research are presented in Section 1.2, followed by the scope of the study in Section 1.3. The contributions of this work are discussed in Section 1.4 and finally this introduction concludes with Section 1.5 which outlines the organization of this thesis.

1.1 Motivation

Solving the biological identification problem requires finding a system or technique that can handle a multi-attribute data set which is often incomplete and sometimes very sparsely populated with values. It also requires that the system is able to mimic decisions of a human expert and is able to discriminate between data patterns that are often fuzzy and overlapping. Human experts often base identification on the 'giss' (general identification based on shape and size), or 'gestalt' of the object being identified, or on whether the identification 'seems' right. These types of decisions are typical of humans and are extremely difficult to represent when a machine is used to make the same sort of decisions. Humans also learn from their experience. What is needed is a system that is capable of mimicking the thinking of a human expert.

Artificial neural networks (ANN) [67, 86, 91, 92, 177, 190, 258] are used for modelling biological neural systems and, in particular, their ability to learn from their environments. In addition, studies of the human cerebral cortex have shown that several areas of the brain are represented by topologically ordered maps [65, 92, 111, 114]. These studies inspired Kohonen to develop the SOM [114] which is a form of a neural network that can provide an objective way of clustering data through self-organizing networks of artificial neurons.

The typical application of SOMs is as a clustering and visualization tool for portraying the central and inter-dependencies or correlated relationships within data on a map [57, 105, 114, 157]. If the input data involve multi-attribute patterns, the use of clustering techniques is essential, while the projection of the input pattern onto a topology-preserving lower-dimensional grid provides the means for efficient and effective visualization of the projected data. Thus the SOM has features which will help group and organize multi-attribute biological data and at the same time represent data in a format that is easy to understand and analyse.

Undoubtedly one of the main advantages of using the SOM for tackling biological identification is the SOM's ability to produce results even if the data set is incomplete. This factor is very significant for samples where there is often an incomplete data set, as is the case with botanical tree identification.

Given the nature of the input data, it is felt that utilizing a SOM could provide a novel and effective technique for plant identification, and at the same time might reveal previously hidden phenetic relationships (or correlations) between the data. Also, by using a SOM some of the problems associated with finding and defining the appropriate underlying reasoning of an expert disappear.

Although DNA barcoding has recently grown rapidly worldwide and is fairly sophisticated, it is felt that it is important that one first gets a good understanding of the morphology of botanical material before progressing to DNA barcoding. The DNA results would make more sense once one has a good understand of the morphological details of plants. For this reason, for this research project it was preferred to proceed with AI techniques rather than DNA barcoding. However, the latter technique is discussed in detail in Section 2.3.

1.2 Objectives

The main objective of this thesis is to evaluate the applicability of the SOM for the problem of biological identification, as compared with other AI methods. In reaching this objective, the following sub-objectives are identified:

- To provide an overview of existing biological identification tools and of some AI methods such as C5, CN2 and SOM.

- To compare the results obtained by using the SOM algorithm with the results obtained by using the C5 algorithm to build decision trees, and the CN2 rule induction algorithm.
- To analyze the results produced by the SOM and to try to identify which characteristics are the principal components used in identifying tree samples; and to see if any previously unrecognized identification patterns can be detected.

1.3 Scope

Collection of data on biological material is a long and tedious process: the amount of material potentially available is overwhelmingly large and obtaining the data is extremely time-consuming.

For this reason, the biological data used have been drawn from botany samples, and the data sources have been further restricted to data obtained from *Acacia* species found in KwaZulu-Natal (KZN), South Africa. This genus was chosen because in KZN it is a relatively small group of indigenous trees that provided a manageable data source that was relatively easy to obtain. *Acacia* are also a prominent component of the flora of this province.

For the purposes of this study only macroscopic characteristics of data samples were considered. In the field, these are the characteristics that are most commonly employed for identification purposes.

1.4 Contributions

This thesis shows that the SOM technique is very useful for identifying biological samples, particularly because of its ability to provide results even when the data sets are incomplete.

It also shows that the SOM technique can be used for predicting the likely identity of a sample even when the set of data presented contains attribute values that have not been presented in the training set.

Finally this thesis shows that what is lacking in other algorithms used for biological identification can be provided by using neural network techniques that are capable of learning from previously presented data.

1.5 Thesis Organization

A brief outline of the difficulties of biological identification has been presented in this chapter. The remainder of the thesis is organized as follows:

- **Chapter 2:** The biological identification process is examined, and an outline of the problems and requirements for easy identification of macroscopic botanical material is given. A literature review relating to the research objectives is given.
- **Chapter 3:** Several algorithmic solutions to the problem are explored.
- **Chapter 4:** The SOM algorithm is described.
- **Chapter 5:** The methods used for collection, preparation and representation of the data are presented and the research methodology used for constructing the SOM models developed for this study is described.
- **Chapter 6:** The SOM models are presented and are analyzed, and the results obtained from testing the data using the SOM models are examined. The data are also presented to the C5 decision tree algorithm, and to the CN2 rule extraction algorithm, and the results are discussed.
- **Chapter 7:** A validation of the research is given by summarizing the main conclusions and the research goals of the thesis. Any shortcomings and possible improvements of the research are discussed and ideas for future work are suggested.
- **Bibliography** is a list of publications consulted in the compilation of the present work.
- **Appendix A** presents a list of acronyms, abbreviations and glossary of terms used in the thesis.
- **Appendix B** presents a batch SOM algorithm.

Chapter 2

Background to Biological Identification

The understanding of biodiversity, in particular organismal diversity, is central to all biology and it is frequently necessary to identify organisms. It is important for scientific and economic reasons to identify and thus name organisms. Identification is important in conservation, environmental management and in jurisprudence. Accessing the name of an organism (identification) can unlock a wealth of information that has been gathered throughout history on that organism by giving access to the warehouse of collective botanical knowledge. Names of economically important plants (especially if one thinks of fields such as genetics, agriculture and biomedicine) are by association of economic significance. As Janzen writes [98], being able to access the name of any plant would be "to plants what the printing press was to stories, education, science, law, medicine and communication". In addition, identification is more important now than in any other point of human history as biodiversity experiences global-wide extinction [10, 97, 129] and taxonomic expertise decreases [16, 21, 208, 243]. Of great concern is the fact that many groups, especially of animals, protists and fungi presently have no experts at all.

Samples of biological specimens brought in for identification are often incomplete, or are poor examples of the source material from which they have been taken, or are fragmentary (for example, museum specimens). This makes identification even more difficult. Although by far the best solution for identifying botanical material is to have an expert taxonomist who can perform the identification, these experts are often not available and due to the increasing shortage of these experts and/or cost of providing specialist expertise this option is often neither feasible nor possible. According to Hebert [93] 15,000 taxonomists will be required in perpetuity to identify life if reliance on just morphological diagnosis continues. Thus alternative aids to identification have to be found. This chapter discusses the problems involved in identifying botanical samples, and some of the techniques used historically and currently.

Section 2.1 introduces the problems involved in identification, and in Section 2.2 the historical solutions are discussed. In particular, Section 2.2.1 describes some of the manual botanical methods that have been used in the past and are still used today. Section 2.2.2 discusses some state-of-the-art computerized systems. Future trends are discussed in Section 2.3 and the last section concludes the chapter.

2.1 The Problem

The greatest problem in identifying any biological specimens is that in nature species can exhibit great variation, known as polymorphism. Entities may be discrete, but conversely there may also be cases where no absolute boundaries are available to delimit taxa. The normal range of morphological variation for biological material can be affected by climatic, environmental and geographical conditions. All of these factors can have an affect on the growth of organisms and thus influence an entity's attribute values and hence its identification pattern. The age of a plant can have a huge influence on some characteristics, for instance leaves are often larger on young plants, heavily shaded branches and coppice growth. Some trees even start off with simple leaves on the young plants and develop compound leaves as the tree matures. Also many leaf characteristics deteriorate in certain seasons with hairs falling off, latex drying up or smells fading. In addition, seldom are all possible characteristics present or observable on a specimen at any one time; for example flowers are often absent when seeds are present.

It is estimated that biodiversity (excluding microbes) is 10 million species, and of these approximately 1.7 million are named species of animals and plants (i.e. form the known portion of biodiversity). The current methods of classification and identification of biodiversity are clearly not coping and there is great need for new and improved identification techniques. According to Weeks and Gaston [240]:

> *"The single greatest impediment to biodiversity research is taxonomic. The resources available are inadequate to meet the demands for the discrimination, description and routine identification of specimens of most taxa in most areas of the world."*

With any identification system the descriptions of specimens depend on how the observer views the objects being described. This complicates the identification process when the descriptions depend on 'fuzzy' terms, such as 'short' or 'pale'. What is meant by the term 'short' can be interpreted differently by each person involved and even differently at various times by the same person.

An identification system needs to able to cope with missing data, multi-variable data and inexact descriptions and still be able to perform identification when previously unseen data are presented. Some of the identification aids that have been used in the past are discussed in the following section.

2.2 Historical Solutions

The identification of biological specimens is a fundamental human activity. For a botanist (layman or professional) identification usually means finding the name for a specimen of a plant. Irrespective of the type of material in question a specimen cannot be identified unless a classification of like-specimens already exists with which the new specimen can be compared. In this sense classification means a way of grouping specimens on the basis of some relationship between them. The groups formed are given names, and when a new specimen is examined and it is decided that it belongs to one of the existing groups it has then been identified. By far the greatest part of the information which a botanist uses when identifying specimens in the field is based on macromorphology: that is, the features that can be detected by human observation and interpreted with ease and speed.

2.2.1 Manual Botanical Methods

Traditional methods of identification include expert determination, recognition, comparison, and the use of keys and similar devices. As mentioned above, by far the best method of identification is the human expert, but such an individual is often not available when needed. When the expert's input cannot be obtained, the problem becomes one of finding a good alternative for the expert's knowledge, judgment, intuitive experience and reliability.

When time is not an issue, a method of identification is to compare an unknown specimen with a stack of previously classified specimens, using any reliable available

records (such as textual descriptions, photographs, illustrations, etc.), until a likely "match" or identification is found. This method can be successful and result in a reliable identification; however, it is not always possible or feasible to sift through available data owing to time constraints and availability of suitable material for identification. In addition, some groups of organisms are more difficult to identify than others, and the lack of specific expertise on a group of organisms can also add to the problem.

The most frequently used identification method is the diagnostic key. The use of a key for identification is several centuries old. In 1736 Linnaeus, often mistakenly referred to as the 'father of modern taxonomy', used a key which he called a *clavis* (Latin for key); however he never applied this to plants [160] and the key was not strictly dichotomous. According to Voss [237] it was not until 1778 that Lamarck, in his *Flore Françoise*, produced a dichotomous key as it is known today.

Since the 18th century most manual macroscopic identification of a botanical nature has been done by means of sequential single-entry keys (e.g. dichotomous keys) in books and articles; such as is found in Leistner [131]. This method does depend to a certain extent on time, material and experience, but is generally the most successful method when an expert is not available. These keys are made up of contrasting, mutually exclusive characteristic statements (called couplets) that require the identifier to make a comparison with the specimen being keyed, and then choose the most appropriate statement from a recognized text [174, 210]. When using a key, each time a choice is made one or more taxa are eliminated by using deductive logic, and the number of possible results (taxa) remaining on the identification list is reduced.
When designing a key macroscopic, morphological and non-variable characteristic states which are generally available to the user of the key are preferable; particularly if they are relatively easy to determine.

Specialized tests on microscopic characteristics often cannot be carried out, especially in the field. Keys have certain formats (indented or bracketed) and conventions (statements started with the same word, as demonstrated in Table 2-1 with the use of the word 'Plant'; and statements started with the name of the plant part, as demonstrated in Table 2-1 with the use of the word 'Spines') are employed to make the use of the key simpler.

Table 2-1 : Extract from Botanical Vegetative Key

(Adapted from Key by Johnson in [168])

1. Plant spinescent……2
Plant without spines……..….26
2. Spines are sharply-tipped dwarf branches……..….....3
Spines are separate structures, curved or straight……..…….....4
….
4. Spines (at least some of them) curved……..5
Spines straight……..…… 16
5. Spines scattered, or single in rows……..…..6
Spines in pairs or trios……11
6. Spines randomly scattered……....7
Spines in rows……...9
7. Tendrils present on some stems…….....*Acacia kraussiana* p132
Tendrils lacking…….....……...8
….
16. Tree always small……...17
Not so.....19
17. Bark corky; spines face towards branch tips.....*Acacia davyi* p128
Bark smooth; twigs sticky.....18
18. Twigs blackish.....*Acacia borleae* p124
Twigs pale..... *Acacia swazica* p138
....

If a key is well written, suitable specimens are available, and the person using the key is careful, a specimen can be successfully identified. However, keys do not necessarily group related species, and in fact the reverse is often the case. In addition, it sometimes happens that the use of specific characteristics is required, but these are not always evident on the specimen to be identified. Also, single access or sequential keys start at a certain point and this can result in their being unusable or difficult to use [100]. If a characteristic is absent or is misinterpreted the user cannot proceed or may end up identifying the specimen incorrectly.

Most of the botanical keys for the flora of southern Africa, besides being based on historical keys, are based on, or have similarities to, one originally developed by Phillips in the 1920s [164] and further developed by Dyer [60]. The key presented by Johnson [168] is also derived from these earlier keys. Table 2-1 gives an example of an adapted portion of Johnson's key that was developed as an aid to identification of indigenous trees found in the former Zululand, Natal and Transkei areas.

Statement numbers are given on the left side of Table 2-1 for each of the couplets presented. Each statement pair represents a choice available to the user. When the user selects one statement from a pair then the next statement pair to be considered is indicated by the number on the right side of the chosen statement. Using Table 2-1, the user reads the first statement pair then looks at the specimen to be identified and decides whether or not it has spines. If it is spinescent the next statement pair to be considered is 2. If it is not spinescent, the next statement pair to be considered would be 26. After considering statement pair 2 the user would go to statement pair 4 if the specimen has separate spine structures (all southern African *Acacia* species have thorns and not spine-tipped branches). Statement pair 4 requires the user to select whether the spines on the specimen are straight or hooked (recurved). If the spines are recurved the user goes to statement pair 5, if the spines are straight then the user jumps to statement pair 16. This process continues until an identification is made. The identification is indicated, on the right hand side of the table, by giving the name of the identified specimen and page number where the description and other information on the species can be found. An example of a successful identification is shown in the first statement in the 7th pair.

Although keys may provide an answer to an expert, a layman cannot easily or conveniently apply the key successfully, and, on occasions, even trained botanists fail in their use of the keys. Many keys only start to 'work properly' when the user begins to understand the way in which the key's compiler has defined any subjective characteristics involved. To use a key successfully, there is usually a need to know the terminology and concepts that are unique to a group or key. For example, "Twigs pale" can mean different things to different users and can lead to the choice of the wrong option in a key. It is because of this, and of the use of specialized terminology,

that the use of sequential keys is one of the main difficulties experienced by students and amateur botanists alike [210].

Although frequent use of a key increases the likelihood of successful identification as the user begins to understand the idiosyncrasies of the key, if an incorrect step is taken along the path through the key (whether by error, misinterpretation, ambiguity or aberration of the specimen itself) the identification process is likely to fail. The chance of going wrong increases as the number of steps in the key increases. Also, the ease of use of a key can vary depending on the difficulty of the plant group involved, and on the whims of the compiler who developed the key. In addition, keys may contain errors due to poor construction.

The strict sequential nature of these keys, working on a strict order of characteristic elimination, does not allow for easy backtracking or lateral progression. Nor does it allow for the free selection of a number of different starting points for identifying/retrieving data. With these types of key, options for identifying an unknown specimen are limited; and the keys make no allowance for ambiguous or atypical data, or for the absence of characteristics in the material being identified.

In order to increase the chance of successful identification some sequential keys employ reticulations. Attempts have been made to produce multiple-entry access keys or polyclaves (e.g. punch or clip card keys and tabular keys) which allow the user, rather than the author of the key, to select the characteristics to be used in the identification process as well as the order in which the characteristics are used. This provides a great advantage especially when the material to be identified is fragmentary. In addition the route taken to identify a specimen may be different from one specimen to the next. Despite these variations, the general logic of identification using a polyclave is the same as that used in a dichotomous key. Although polyclaves increase the chance of a successful identification they are not infallible and can only handle small data sets. They do have a major advantage in that they allow the user freedom to choose any characteristic in any sequence, thus avoiding the rigid format of sequential keys [100] and allowing for the possible identification of incomplete specimens with missing data.

Computerized polyclaves which employ characteristic elimination or probabilistic techniques have been developed, and some of these can list the taxa that have been

eliminated and the taxa that remain as possible identification candidates. Polyclaves and other computerized systems are discussed in Section 2.2.2.

2.2.2 Computerized Systems

More recently, computer keys have been produced which allow for identification and information retrieval; and several interactive systems have been used for plant identification. There have been several major approaches as far as computerized biological identification is concerned. These include computer-stored dichotomous keys, computer-constructed keys, simultaneous characteristic-set methods, and automated pattern recognition.

Many of the applications of these approaches provide no real advantage over printed keys, or are so time consuming that they are only practical for application to groups of taxa which require specialized attention. In addition some of these approaches require specialized equipment.

Despite this there can be several advantages of computerized keys over conventional keys which include the following:

- provide a variable starting point, i.e. make it possible to start with almost any characteristic, in any order so that characteristics not present on the specimen may be avoided,
- provide the means by which it is still possible to arrive at the correct identification even if an error is made by the user,
- provide the means for easy backtracking if a mistake is made[1], and
- provide the means for easy updating and modification.

Before data processing (for example, the formation of a key) can take place it is necessary to describe data in a systematic and unified manner, and some identification systems make use of keys generated in DELTA format Language for TAxonomy (DELTA) [47], or use keys that are at least compatible with the DELTA format. The DELTA format was first defined in 1973 by Dallwitz [48] and became the official international taxonomic standard, endorsed by the Taxonomic Data Working Group (TDWG), for defining taxonomic keys [25]. This has had the effect that many

[1] Note: there is no backtracking in Self-organizing maps (SOM)

taxonomic identification systems developed are capable of interfacing with DELTA files (which are ASCII coded files). Other formats have been used, but until recently DELTA was the accepted standard. The TDWG is at present working on a new standard called Structure of Descriptive Data (SDD) [30, 216] which is an upgrade/replacement for DELTA, which is now over 30 years old.

Several existing interactive identification programs have been investigated and compared [50, 55]. One of these is Intkey [49, 51, 52], which is a matrix-based interactive program for identifying a specimen by comparing its attributes with stored descriptions of taxa. Intkey, when used in combination with the DELTA system, offers the means to generate dichotomous, open access keys, identify unknown specimens and retrieve information.

Linnaeus II [189] software was designed and built by the Expert Center for Taxonomic Identification, and can be used for identification allowing for output in the form of text, pictures, sound track or video.

LucID [155] is another state-of-the-art computerized commercial and research system available for creation of dichotomous, open access keys and for identification of specimens and retrieval of information. It allows the developer to include text, sound and images in order to help the user to select taxonomic and diagnostic characteristics. As the user selects characteristic states, those taxa which do not possess the chosen characteristic states are excluded, thereby reducing the list of possible taxa. Once the specimen has been identified to a particular taxon, information, sub-keys, or links to web sites for further information can be obtained.

These existing systems tend to be successful only in cases that provide a clear “either/or” alternative and do not really cater for overlapping or vague information. Few computer systems cater for cases that are not an exact fit to the stored data, i.e. fuzzy cases that do not fit exactly into a defined category. The major problem with all of these identification programs is that they do not utilize the full potential of the computer and in some cases are little more than a computerized polyclave. Another feature of these systems is that they sometimes offer the user many facilities which, after practice, can be used effectively and accurately by expert botanists who understand fully the terms and implications of the use/exclusion/modification of such facilities. Unfortunately, some of these same facilities are potentially undesirable

when the systems are used by less experienced users, and lead to incorrect identification. However, it is specifically for these inexperienced users that these systems should be catering.

2.3 Current Research and Future Possibilities

In 1977 Carl Woese [248-250] used sequence differences in ribosomal RNA (rRNA) to define a new domain of life, called Archaea which led to the redrawing of the evolutionary tree. Since this work by Woese, both DNA and RNA molecular studies have been used within the biological field, and in the past four to five years there has been a flurry of work and discussion about genomic approaches to taxon diagnosis.

Both DNA arrays (micro and macro) and barcodes have been used for the species-level identification of organisms. DNA microarrays are ordered, low density, samples of an organism's DNA placed in high density on a solid support so that each sample represents a particular gene. This can then be analyzed for changes in the expression patterns of the representative genes after different treatments or conditions [88, 153]. The array-based approach has the requirement of prior knowledge of sequences in the target species, and this is a limitation of the approach [88].

Most eukaryote cells contain mitochondria. In animals mitochondrial DNA (mtDNA) is characterized by a relatively fast mutation rate, which exhibits a significant inter-species variance but comparatively small intra-species variance. DNA barcoding utilizes this property of mtDNA to provide a taxonomic method of identification. In DNA barcoding a short standardized mtDNA sequence from an unknown organism is used to assign that organism to a known species and also to aid in the discovery of new species. In 2003 Hebert [93] proposed that a library of DNA barcodes should be compiled that would be linked to named specimens. The idea behind this is that the data bank of barcodes could then provide a means for identifying species.

DNA barcoding appears to be a promising process if researchers are able to standardize the genetic sequence/s and method/s in order to provide 'barcode's for identifying species. There have been many reports of successful identification using barcodes [87, 89, 93-95, 212]. However the question remains: is a short sample of

genetic code from a reference gene specific enough on one species to distinguish it from every other species? It is claimed by some that this has been proved to be the case and that comparisons of sequence variations in that section of the gene used can reveal evolutionary relationships among species. There are avid supporters of DNA barcoding [215]. However, both the method and its applications are being questioned by others [58, 96, 244, 246, 247].

It has been argued by Whitworth *et al.* [244] that using mtDNA can lead to misleading results as it is possible for two different species to share mtDNA, or for one species to have more than one mtDNA sequence exhibited by different individuals. These authors claim that identification at the species level based on mitochondrial sequence may not be possible for all insects.

It has also been suggested that for plants a multi-loci approach (rather than the mitochondrial cytochrome *c* oxidase I (COI) gene approach which has been widely used in animal barcoding) is necessary because plant mitochondrial genes do not differ sufficiently amongst closely related species [121, 122].

Barcodes do appear to have the potential to be an extremely useful tool for taxonomy, especially when it comes to the identification of organisms which are difficult to recognize from morphological characteristics. However, in the 'plant world' the debate over barcoding rages on and identification by traditional means still prevails. Better identification methods, therefore, still need to be developed.

2.4 Conclusion

Streamlining of procedures to identify organisms is obviously needed. At the moment it seems that taxonomists are loosing the battle to identify and name organisms as many species are becoming extinct faster than they can be named and classified. Despite the different opinions on the proposal of using DNA barcoders to barcode life, there has been a genomic revolution over the last 15 years. If the advantages of DNA sequencing can be integrated with the benefits of classical taxonomy then exciting new developments could take place. This seems unlikely to happen in the near future. Challenges have been made against the rationale, methodology and interpretation of results as used for DNA barcoding [247]. Many biologists feel that DNA barcoding should be used to augment taxonomic research rather than replacing

it. There is still a need for taxonomists who rely on morphological characteristics, and therefore the printed key and computerized biological identification systems are still necessary and will remain so for the foreseeable future.

This chapter aimed to provide background information on existing biological identification techniques and systems. It established that there is a need for improving the identification methods in order to facilitate the 'identification of life'. In the next chapter several algorithmic solutions which could be utilized are discussed. In particular, artificial intelligence approaches are investigated in order to determine whether any of these approaches could offer some solutions to the identification problem which other methods do not offer.

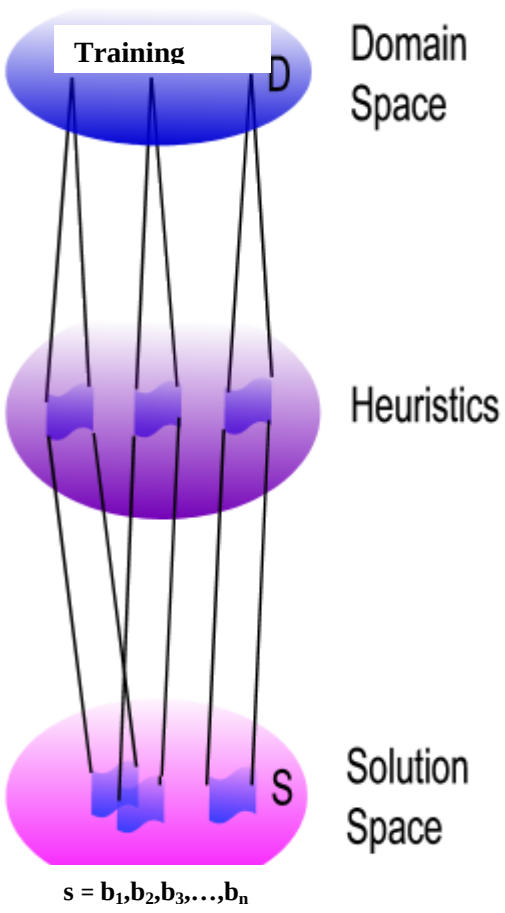
Chapter 3
Algorithmic Solutions for Biological Identification

Chapter 2 provided an insight into some of the problems which need to be addressed in biological identification, and reviewed some of the biological systems that are currently utilized for identifying biodiversity. In order to add a more 'intelligent' approach to identification in general, the employment of artificial intelligence techniques needs to be considered. This chapter aims to provide an overview of some artificial intelligence (AI) techniques and algorithms which have been or could be employed in biological identification.

3.1 Introduction

AI comprises methods, tools, and systems for solving problems that normally require the intelligence of humans. The name AI was coined in 1956 [44], and since its inception AI has undergone periods when it has been viewed with great expectations, interspersed with periods of disinterest. The main direction that the development of AI has taken in the past has been the development of methods and systems that model the way humans think or act, or by developing systems which think or act rationally [182]. Problem solving is a fundamental human activity, and humans use a lot of iteration and heuristics in their everyday life to solve problems. Heuristic methods are based on emotion, experience, intuition, rational ideas, consciousness and rules of thumb [65, 182]. One of the main objectives of AI has been to represent simple heuristics in a computer in order to get computers to 'learn' intuitive knowledge. This is exactly what is needed in biological identification: if one can reproduce heuristic learning in a computer and thereafter get a computer to implement instinctive decisions, then one can use a computer to help solve the identification problem. For example, a taxonomist looking at a tree knows what the identity of the observed tree or specimen could or could not be; this process seems to be instinctive or intuitive but in fact the taxonomist uses deductive reasoning to make the identification. The

taxonomist uses knowledge, past experiences and heuristic rules to direct the search for the identity of a specimen and place it into a likely group or class of specimens. The members of the chosen group should have similar characteristics to those of the specimen to be identified. This use of heuristic rules [102] may be represented as shown in Figure 3-1.



A problem may be defined by assuming that a set of **n** independent input attributes, $\mathbf{x_1, x_2, x_3, \ldots, x_n}$, and a set of **m** variables of the solution $\mathbf{y_1, y_2, y_3, \ldots, y_m}$, are defined. Then in the domain space **D** every possible combination of values for the input variables can be represented as a vector $\mathbf{d = (a_1, a_2, a_3, \ldots, a_n)}$, and in the solution space **S** every possible value for the set of output variables can be represented as a vector $\mathbf{s = (b_1, b_2, b_3, \ldots, b_m)}$. There is seldom a single formula for giving the optimal solution and so problem solving can be viewed as mapping **D** to **S** with heuristics providing the means of obtaining restricted projection of **D** to patches in **S**. In reality this is what an expert does.

Figure 3-1 : Model Representing Use of Heuristics
(Adapted from Kasabov [102, page 5]).

Recent years have seen a revolution in AI with researchers realizing that different areas within AI are inter-related and that these areas should not be developed and utilized in isolation [182]: according to Russel *et al.* there are, and should be, symbiotic relationships between the different areas that comprise AI.

More recently a sub-branch of AI called Computational Intelligence (CI) has evolved in which each paradigm has its origins in naturally or biologically occurring systems [65, 163, 167]. It is generally accepted that the CI paradigms consist of neural networks, fuzzy systems, evolutionary computing, and Swarm Intelligence (SI) [46, 65, 72, 163, 167]. Some of the paradigms which are reviewed in this study fall within CI, but in this document the more general and encompassing term AI will be used.

This chapter presents the algorithms of some of the AI paradigms that influenced this study. Expert systems, fuzzy expert systems and neural networks all try to represent and mimic heuristics in computers: each includes the study of mechanisms that allow intelligent behaviour in complex environments. These topics are discussed in the following sections: Section 3.2 discusses expert systems and Section 3.3 introduces fuzzy expert systems. Artificial Neural networks (ANNs) are introduced in Section 3.4 and two types of learning are discussed: supervised learning algorithms (Section 3.4.1) and unsupervised learning algorithms (Section 3.4.2). Self-organizing map (SOM) training, which is based on a competitive learning strategy, is introduced in Section 3.4.3 and in Section 3.4.4 biological applications of ANNs are discussed. Section 3.5 discusses other classification algorithms - the C5 decision tree algorithm is discussed in Section 3.5.1, and the CN2 rule induction method is discussed in Section 3.5.2. Other algorithmic solutions that have been used in biological identification are mentioned in Section 3.5.3. Finally, Section 3.6 concludes the chapter.

3.2 Expert Systems

Plant taxonomy is a complex, meticulous and information-heavy science which allows taxa to be identified by retrieving information contained on them in a classificatory system. Although there are various ways in which this identification may be performed, the most commonly used ones employ dichotomous keys. This process requires knowledge of botanical terminology and scientific knowledge of the organs of plants and, as the process is complex, botany-related activities are not particularly automated. For many taxa considerable experience, a large literature library and a comprehensive research collection are required for authoritative identifications. In

general, as an expert's experience and expertise grow, and as the expert is better able to work on problems in classification and identification, so the expert's skills become increasingly in demand. The expert cannot always meet this demand, therefore there is a very real need for more efficient devices for identification. Reducing the time, effort, and expertise required for identification will allow experts to concentrate on other issues such as the description of new species. Existing computerized biological identification systems are basically databases which store data on the specimens and find a name via a process of filtering. AI can offer a far more dynamic approach to identification. By using AI techniques it is possible for data to be analyzed in order to determine patterns and relationships which allow for the collection of information that was not stored explicitly in the database [43].

A technique that has been employed for biological identification is Expert Systems (ESs). These are computer programs that incorporate the knowledge of one or more human experts in a narrow, knowledge specific domain, and try to solve problems in that domain by matching the expert's level of performance. Durkin provides an overview of ESs: highlighting the major characteristics, comparing conventional programs with ESs, and reviewing several systems particularly developed for application in science [59].

The reason that more widespread use of ESs has not occurred is that knowledge bases are often incomplete: the methods currently used to represent knowledge in ESs rarely capture subtleties, and sometimes fail to reflect major aspects of an expert's knowledge and understanding [63]. There are technical, psychological and sociological problems associated with ES development [66]. When developing an ES it is necessary to combine expertise derived from taxonomy, system design and uncertainty logic, but getting experts from each of these fields to understand each other and work together in a constructive and productive manner is extremely difficult. Once developed, the output of such a system would reflect the opinion of the humans involved rather than the inherent nature of the data and their inter-relationships; hence such a system would be expert driven rather than data driven. Additionally, ESs are not generally successful when applied to broad, subjective problem-solving even though they can be applied successfully to specific, contained problems.

3.2.1 Early Expert Systems

In the 1960s DENDRAL, which is a portmanteau of the term "Dendritic Algorithm", was one of the first influential pioneer projects in AI, and it included the production of an expert system. The primary aim of the ES was to automate the decision-making process and problem-solving behaviour of organic chemists in order to help with the identification of unknown organic molecules. DENDRAL did this identification by analyzing the mass spectra of chemicals and using a chemistry knowledge base as a lookup table. Development of the system was carried out at Stanford University by Edward Feigenbaum and other scientists [69, 130, 135]. DENDRAL consisted of two sub-programs, Heuristic DENDRAL and Meta-DENDRAL [68, 70, 135], and was written in the Lisp programming language. Many systems have been derived from DENDRAL, including MYCIN; and the true significance of DENDRAL was as the direct progenitor of MYCIN and today's generation of ESs [135].

Developed in the 1970s, MYCIN was an important milestone in the development of ESs. It consisted of a computer program designed to function as a consultant on problems of medical diagnosis and therapy selection [27, 34, 35, 54, 68, 196-198, 200-205, 252]. Its field of application was infectious diseases, and it was used fairly successfully in dealing with cases of bacteraemia and meningitis [199, 254, 255]. Yu *et al.* [254] found that MYCIN's therapy recommendations met Stanford's standards of acceptable practice 90.9% of the time, i.e. the system was tested by specialists in infectious diseases who judged and concurred with MYCIN's final therapy recommendations, as well as MYCIN's intermediate conclusions about the significance of the infection and the identity of the infecting organisms.

MYCIN programs were further developed to produce other medical programs [34, 35], and hence the original programs are sometimes referred to as the father of ESs.

3.2.2 Biological Applications of Expert Systems

Also within the biological environment, an ES called SYSTEX [251] (SYSTematics EXpert) was developed in the 1980s to test the application of ES technology to the general problem of taxonomic diagnosis. This system used a rule-based backward chaining system and was developed using a commercially available expert system shell (M1 ES development software package from Teknowledge Corporation). The

authors of SYSTEX [251] suggest that the ES approach is superior to the dichotomous key and other identification devices in terms of efficiency and ease of use, tolerance of missing data, explanatory capability, and the ability to provide meaningful output when an unambiguous identification is not possible. However, according to the authors, the ES does not provide new information, but rather acts as an integrator of knowledge-and-delivery devices for scarce expertise and training. The insect species group chosen for testing SYSTEX was the *Signiphora aleyrodis*-group. Diagnostic characteristics of the *aleyrodis*-group species were collected and processed and then used to test the ES. In Woolley and Stone [251] no statistics are presented on the success or failure of test results, nor are the test results compared with other systems.

EXPERT KEY [12] is another expert system to aid in biological identification. The system employs the Dempster-Shafer theory of evidence [183, 194] (a generalization of probability theory used for inexact reasoning) to combine heuristic rules. Uncertain inference is also used to allow the user to express lack of certainty about the statements in the key. The use of heuristics results in the number of key couplets in the key being significantly reduced, which has the effect of reducing and simplifying the tasks to be performed by a non-expert. In Atkenson and Gammerman [12], EXPERT KEY was illustrated by using it to identify four different species of *Umbelliferae* (= *Apiaceae*)., The authors showed that the number of key couplets needed for correct identification by the system can be reduced by as much as 80% with the use of heuristics, thus making the use of the key much simpler for non-experts.

Contreras *et al.* [43] and Fajardo *et al.* [66] give a description of an ES, called GREEN, which was developed for identification of Iberian Gymnosperms (both indigenous and cultivated) and which allows online queries. The system was developed independently of the database on which it was employed, thus making it possible for the system to be adapted for identification of other species. However, for each new species the system would have to be modified.

The group of Gymnosperms chosen for demonstrating GREEN consisted of 46 taxa. Information for the knowledge base was taken from keys and used to produce a list of diagnostic characteristics or attributes. Information gathered was further compared by observing nature and consulting documents and experts. The important

taxonomic characteristics of Gymnosperms were divided into groups such as the general aspects of the taxon, characteristics of the branches, the leaves, the shoots, the fruit, the seeds and the ecology.

After the data on the specimens to be identified is presented to it, the GREEN system gives the user a set of results ordered according to how well the result fits the query. Thus, as is the case with EXPERT KEY [12], a definitive identification is not necessarily obtained. Neither paper on GREEN [43, 66] presents statistics or comparisons of results for tests performed. However, the GREEN system was developed as part of the research for a doctorate written in Spanish, and in this dissertation the author reports an identification success rate of at least 83.33% [78].

Dallwitz [49] discusses ESs and matrix-based systems (for example, Intkey) for taxonomic identification and concludes there are advantages and disadvantages of both systems. However, the author does not include any results from an empirical comparison of the two types of systems.

Despite the fact that biological data are frequently incomplete, an expert taxonomist is able to handle uncertainty and missing data and still come up with an answer. A successful identification system must be able to do the same under similar conditions. Usually the human thinking, reasoning and perception processes cannot be expressed precisely, and these types of experiences can rarely be expressed or measured using statistical or probability theory [4]. One way of dealing with these problems is to use fuzzy logic.

3.3 Fuzzy Expert Systems

Identification of botanical specimens demands an acceptance of uncertainty (for example, the use of fragmentary and subjective information) to reach an estimate of the true identification. The theoretical basis behind fuzzy techniques allows for the handling of uncertainty and imprecision, and for fuzzy reasoning schemes to be developed [132]. The theory of usuality [257] allows for the use of common sense in ESs by providing a method of representing knowledge about events or items that are often true.

Traditional ESs, which must use knowledge engineering to acquire all relevant rules from experts, are inherently "brittle", failing catastrophically when presented with situations outside the domain for which their rules were developed. Programmers have tried to solve some of these problems by attempting to develop more flexible systems. Endeavours were made to reproduce human reasoning using imprecise, or fuzzy, linguistic terms embedded in fuzzy systems.

3.3.1 Expert Systems Vs Fuzzy Expert Systems

A Fuzzy Expert System (FES) is defined in the same way as an ordinary ES, but methods and philosophies of fuzzy logic are applied for the inference process. In addition to the standard rules implemented in an ordinary ES, a fuzzy ES may use fuzzy data, fuzzy rules, and fuzzy inference. Abraham [4] gives a good introduction to ESs and FESs. A FES can provide answers where systems demand reasoning that entails uncertainty and imprecision. Typically, FESs when compared to non-fuzzy ESs require fewer rules, need fewer variables, use a linguistic rather than a numerical description, and can relate output to input. Such a system would be closer to 'human-like' thinking and would use fuzzy rules instead of exact rules; thus representing in a straightforward way 'common sense' knowledge and skills, or knowledge that is subjective, ambiguous, vague, or contradictory. This knowledge might come from many different sources, such as from long-term experience from many people over many years.

3.3.2 Biological Applications of Fuzzy Expert Systems

Tien *et al.* [217] describe a fuzzy rule base embedded into a triple-layered network structure for nonlinear modelling of a multivariable system, and used it to demonstrate two kinds of models: one to identify/predict lettuce growth, and the other to control greenhouse climate. The authors compared the FES prediction of results with actual readings of results and found that there was no scientifically or statistically significant difference in the compared results. Tien *et al.* hence conclude that data-driven modelling using neural networks and fuzzy modelling is a more suitable method for application to multivariate botanical data than mechanistic modelling procedures. They claim that the neuro-fuzzy approach is easier and faster, and that the fuzzy rules used are self-explanatory. The authors also claim that it is possible to incorporate

human knowledge and to deduce interpretable rules that describe the systems' behaviour.

Cheung *et al.* [31] used a FES to predict the vulnerability of marine fish to extinction resulting from fishing activities. Data from 159 marine fish species from the FishBase database [77] were used to test the system. Three independent data sets were used to examine the validity of intrinsic vulnerability to extinction, i.e. when the data sets were presented to the fuzzy system the results obtained were compared with the extinction risk of the sets which were already known (by other methods). When required biological data for a particular species were absent in the original data set, the data for that species were obtained from FishBase. Cheung *et al.* [31] used goodness-of-fit of test statistics as an indicator for reporting on the accuracy of extinction risk predictions.

Cheung *et al.* also compared the results obtained using the FES with results obtained from testing the same data with an ES with classical logic sets. This ES had attributes and rules exactly the same as the fuzzy system, but classical sets were used instead of fuzzy sets. Comparisons of FES results with empirical population abundance trends showed that a fuzzy system could be used to predict intrinsic vulnerability of marine fish. The tests also suggested that the use of fuzzy logic in the ES provided a better predictor of intrinsic vulnerability than a system employing classical logic. The authors state that the fuzzy system could react to new information, and that the heuristic rules, fuzzy membership functions, and the values that defined them could be modified based on expert knowledge or newly available information. Thus the FES could be extended easily and further improved. The results obtained from testing the system were compared with results obtained from empirical studies. This comparison showed that the use of fuzzy logic provides a better predictor of intrinsic vulnerability than a system employing classical logic. However, the tests were also repeated using a reduced number of attributes, and the results from these tests showed that the performance of this FES relative to other methods decreases when data are scarce.

Pappas [161] used fuzzy measures and classification rules to analyze shape groups of the diatom *Asterionella* and found that fuzzy decision-making analytical tools could be used to produce results (i.e. tables) that could then provide taxonomists with

a useful and meaningful way to undertake initial morphological identification in the field, i.e. the tables could be used for assigning an unknown specimen to a known group or class. The identification results obtained by using the FES concurred with experts' identification in 46 out of 59 cases (with a crossover point of 0.75). The method was found to work even when information on the species and number of specimens was scant. Pappas showed that fuzzy logic, approximate reasoning and information collation are all used in taxonomy and species identification. Results of this study demonstrate that the FES could help with initial identification of the diatoms. The authors, however, do suggest that when making taxonomic decisions the tables should be supported with additional analyses.

Such studies suggest that when analyzing the problems faced with identifying trees, the use of fuzzy logic to represent uncertain, overlapping and imprecise data seems very appealing. While an ES is a good tool to develop, it has some disadvantages. A major disadvantage of expert systems is that they fail whenever a situation occurs which their rules cannot handle. Other disadvantages include the essential process of collecting knowledge from domain experts, for the success of the ES will depend on the completeness or scope of the knowledge obtained from the expert. A domain expert's knowledge is expressed in terms of one's intuition and experience, and is very dependent on how one views a particular characteristic. All this implied knowledge has to be passed on to the system developer and then embedded into the system developed. When using a fuzzy ES, the task of designing the fuzzy membership functions will also require collaboration from the domain experts in order to find the most accurate or appropriate membership functions: a process which would be difficult to implement efficiently for optimal results. For these reasons it was considered worth looking at an AI model that was data-driven rather than knowledge-driven.

3.4 Artificial Neural Networks

The human brain is made up of approximately 10^{11} neurons [139] which are interconnected and which operate in parallel to process information. The neurons communicate across a network of axons and synapses and act as the computing

elements of the biological brain [258]. Figure 3-2 shows a simplified model of biological neural information flow.

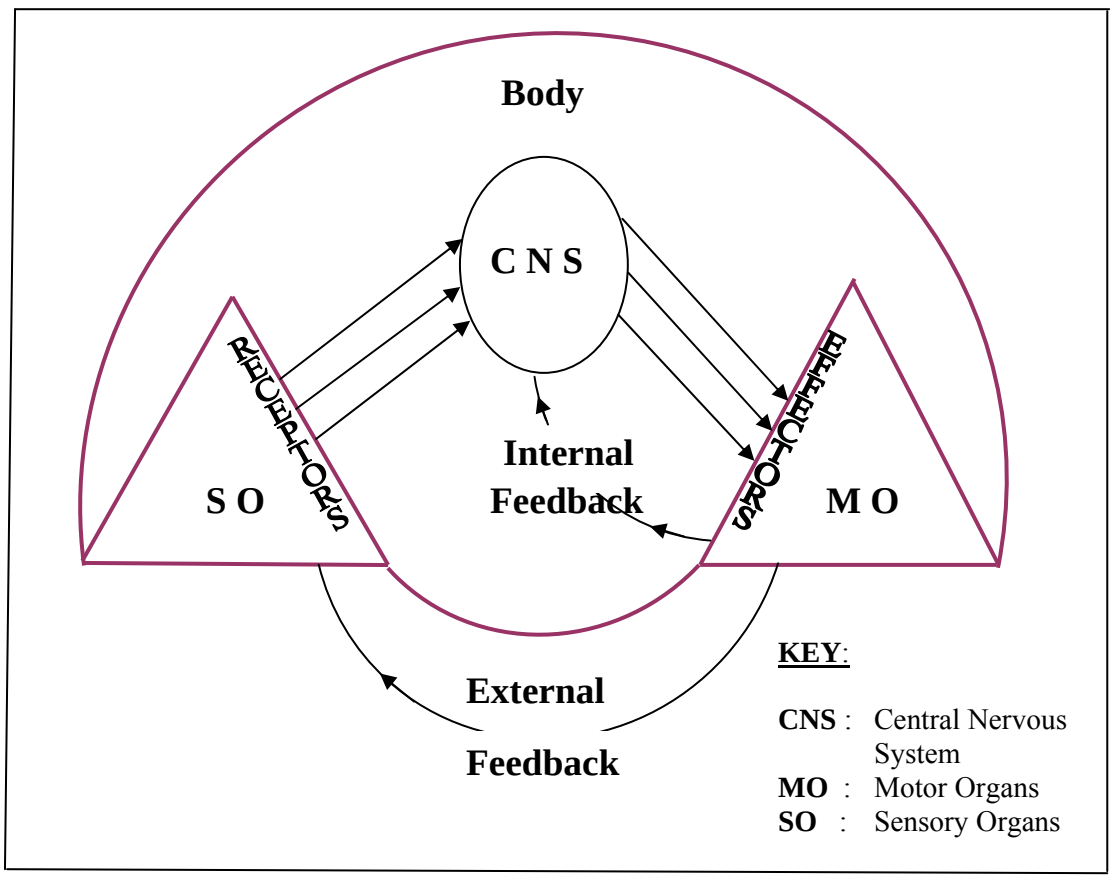


Figure 3-2 : Simplified Model of Neural (Biological) Information Flow

(Adapted from Zurada [258, page 27])

The objective of an Artificial Neural Network (ANN) is to simulate the activity of the human brain. It does this by an interconnection of neurons which mimic the structure and operating principles of the human brain. A simple model representing an artificial neural network is shown in Figure 3-3. In this figure, for clarity, only one neuron is shown, although in reality many neurons would be present. Every input set, **X**, is presented to each node (= processing neuron), each of which has an associated weight vector of values, **W**. The weights associated with the neurons are adjusted to better represent the input patterns, and when this process occurs the neurons are said to be learning. The pattern associations obtained through training may then be used to

classify appropriate test data into the correct classes. Thus an ANN can be used to model the pattern association ability of the human brain.

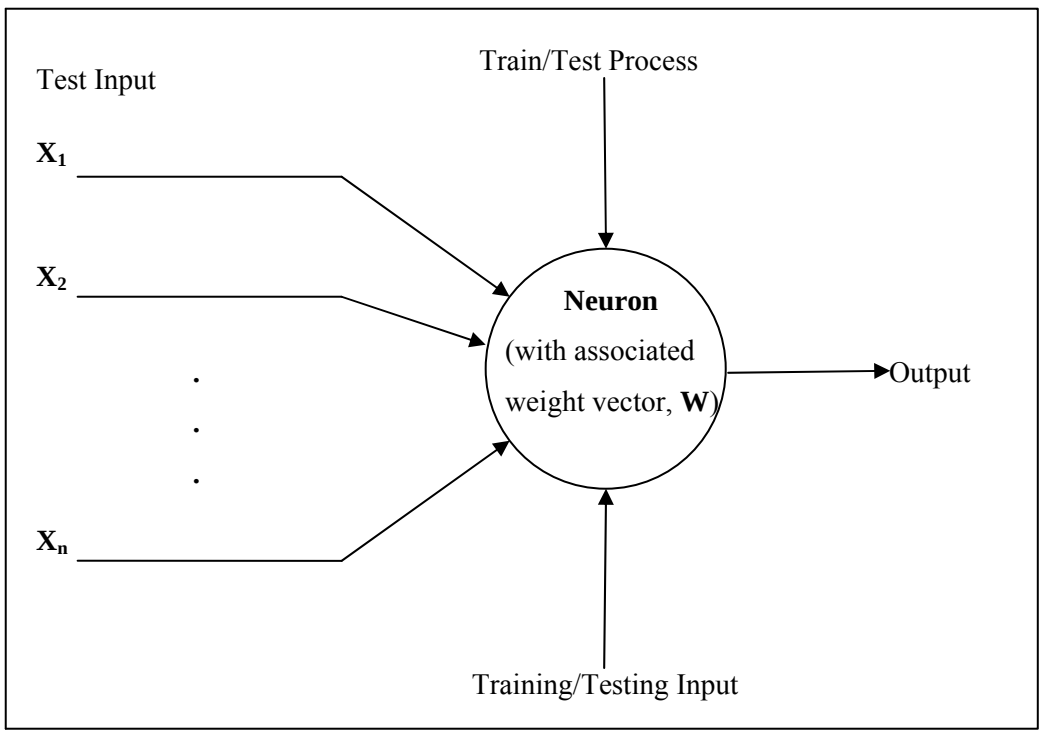


Figure 3-3 : Simplified Artificial Neural Network Model

Although ANNs are based on and attempt to mimic their biological counterparts in the human nervous system and can execute instructions at extremely fast speeds, human beings whose brains operate at much slower speeds still outperform computers at tasks such as biological identification.

There are various learning algorithms that can be used by an ANN, and two of them are discussed next.

3.4.1 Supervised Learning Algorithms

With ANNs that use a supervised learning algorithm, an input pattern and a desired response are presented to the ANN. The ANN tries to learn the functional mapping between the input and desired response vectors. Thus the learning is achieved through example. Once the ANN is trained to recognize the matching input and output patterns it can be used to predict network output accurately when presented with the previously unseen inputs.

3.4.2 Unsupervised Learning Algorithms

Unsupervised learning algorithms attempt to cluster a data set into homogeneous regions by correlating characteristics present in the data. In unsupervised learning the objective is to discover patterns or features in the input data with no help from a teacher, basically performing a clustering of the input space.

Such a data-driven model would look for inter-relationships and regularities in the input data presented to it during the training procedure, and would then use the information gathered from these training sessions to classify the test input data.

For a supervised network, the input and desired output need to be known in advance. However, in tree identification the exact input for each output is often not known: the input from one specimen need not be the same (and seldom is the same) as the data set of another input specimen even when the specimens belong to the same species. Therefore, an unsupervised network that seeks for input data relationships in order to predict the output is more likely to yield realistic results. Thus unsupervised ANN techniques are often used for classifying, organizing and visualizing data sets.

The SOM technique, which uses an unsupervised learning algorithm, is discussed next.

3.4.3 Self-Organizing Maps

A Self-organizing map (SOM) is a form of an ANN that can provide an objective way of classifying data through self-organizing networks of artificial neurons. It is a feed-forward ANN that uses an unsupervised training algorithm and can be trained to learn or find relationships between inputs, or can organize data so as to discover unknown patterns or structures.

Teuvo Kohonen [111, 114], motivated by the self-organization characteristics of the human cerebral cortex, developed the self-organizing feature map. Studies of the cerebral cortex have shown that the motor cortex, somatosensory cortex, visual cortex and auditory cortex are represented by topologically ordered maps [65]. These topological maps are formed to represent the structures sensed in the sensory input signals. Similarly, during training the SOM effectively clusters the input vectors through a competitive learning process while maintaining the topological structure of the input space.

The basic SOM algorithm involves sequential training and is outlined next. A more detailed description of the SOM algorithm will be given in chapter 4.

Each neuron in the ANN has a model (or codebook or reference) vector associated with it. This vector has the same dimension as the vectors in the input data set that are used as the training vectors. Once the codebook vectors are initialized with either random values or in some other way, the training data are presented to the unsupervised SOM algorithm. During training each input vector is assigned to the neuron with the most similar codebook vector or best-matching node (BMN).

In essence, the learning process itself gradually updates the codebook vectors to match the input vectors and, at the same time, maintains the representation of the internal properties of the input data as faithfully as possible. Thus, the input vectors which are relatively close in the input space are mapped to nodes that are relatively close in the output space.

The SOM algorithm contains elements of competitive and cooperative learning. Competitive learning is covered by selection of the BMN, the "winner", which has its vector values updated to the largest extent. Cooperative learning is applied by updating the most similar model vector as well as its closest neighbours. The closest neighbours have their associated vector values updated to a lesser extent than the winner, which results in the creation of similar areas on the output map.

The SOM algorithm has been applied to a variety of real-world problems [105, 157]. The main advantage of applying the algorithm comes from the easy visualization and interpretation of clusters formed by the map. One of the main reasons [114] for using a SOM for exploratory data analysis and data mining is that it is a numerical method and is therefore able to treat numerical statistical data naturally and to represent graded relationships. Other reasons for using the SOM algorithm are that it is a non-parametric method; no assumptions about the distribution of data need be made in advance; and it is a method that can detect unexpected pattern structures by learning without supervision. The SOM can be used deductively and is able to produce results even if the data set is incomplete, which is extremely important when dealing with biological material requiring identification.

By using a SOM some of the problems associated with finding the appropriate underlying reasoning of an expert disappear. These problems include trying to co-

ordinate the expertise of different specialists and getting them to agree on different issues involved in the identification process. In addition, the SOM method offers an easy way to visualize results. The typical applications of SOMs are as clustering and visualization tools for portraying process states by representing the central dependencies within the data on maps. The advantage of using a graphical representation is that a clear visualization of the output is given. For example, by using a SOM it is possible to 'see' the identity of a tree: by presenting an unidentified input pattern to a trained network and looking to see to which area of the map the input pattern has been allocated.

Another of the advantages of the SOM is that it can provide a probability for each species of an unidentified biological specimen belonging to a particular species. At the same time the SOM can also be used to investigate the differences between clusters of species, and in the process it is possible that some new features that discern between the species might be revealed. The SOM can also be used to determine which features or characteristics are the most important or diagnostic ones to consider when discerning between given species.

Given the nature of the input data it is felt that the SOM could provide a suitable technique for tree identification in southern Africa, and at the same time might reveal previously hidden relationships between the data items and between taxa.

3.4.4 Biological Applications of Neural Networks

ANNs were developed initially to model biological functions and have been shown to be flexible and universal function approximators for numerical data. They are powerful tools for modelling biological systems, especially when the underlying data relationships are unknown. Various types of ANNs have been used to do biological analysis, and some of these will now be reviewed briefly.

Tan and Gilbert [214] performed an empirical comparison of supervised machine learning techniques to classify data from four biological data sets obtained from the UCI machine learning repository [11]. Comparisons were made between rule-based learning systems, statistical learning systems (including ANNs) and ensemble methods (stacking, bagging and boosting). Tan and Gilbert concluded that for the task of classifying biological data combined machine learning methods perform better than

individual ones. From their study the authors also concluded that accuracy (Acc), which is the proportion of correctly identified instances, is not enough of a measure on its own when comparing systems. The authors suggest that several additional measurements, based on the sensitivity, specificity and positive predictive value of the algorithms, should also be made. Positive predictive accuracy (PPV) is the reliability of the positive predictions. Sensitivity (Sn) measures the fraction of actual positives, and specificity (Sp) measures the fraction of actual negatives. The equations defining these measurements are given as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eq. 3-1}$$

$$PPV = \frac{TP}{TP + FP} \quad \text{Eq. 3-2}$$

$$Sn = \frac{TP}{TP + FN} \quad \text{Eq. 3-3}$$

$$Sp = \frac{TN}{TN + FP} \quad \text{Eq. 3-4}$$

where

TP is the number of True Positives (correctly identified entities),
FP is the number of False Positives (incorrectly identified entities),
TN is the number of True Negatives (correctly identified entities), and
FN is the number of False Negatives (incorrectly identified entities).

When comparing different systems Podgorelec *et al.* [166] recommend using sensitivity and specificity measures in addition to using accuracy measures.

Clark and Warwick [36] used the multilayer perceptron (MLP) with the backpropagation (BP) algorithm [242] for botanical identification. The simple MLP was used with one input node for each characteristic, one layer of hidden nodes and one output node for each species to be identified. Using this ANN, the authors reported a 93.3% validation success rate when using the *Iris* data set [11], and an accuracy of 52.9% successful identification with a data set of specimens for *Lithops*,

(using 13 characteristic states for 34 species). In a separate paper, Clark [37] also reported using a supervised ANN to identify *Lithops* and a comparison was made with taxonomic keys generated by means of the DELTA system. The ANN was found to perform better than the DELTA key generator when the available data are limited and the species are relatively difficult to distinguish.

A table published by Gaston and O'Neill [79: Table 2, p662], gives examples of semi-automated and automated species identification test results based on morphological characteristics. According to this table some excellent identification results have been achieved with small data sets of biological specimen's material. For example, Gaston and O'Neill report that structures of Africanized and non-Africanized honeybee wings have been identified with 100% success rate using Lucas continuous n-tuple nearest neighbour classifier methods [80]. These tests were performed on one species of bees with two sub-specific variants. Other tests using this method are also reported as having a high success rate for identifying small data sets of biological material [80, 239].

Also reported in the table from [79] is an ANN application used to identify plant pollen (three species). This study reports an 83% successful identification rate [73]. The ANN used in this study was the Paradise (PAttern Recognition Architecture for Deformation Invariant Shape Encoding) which was designed for recognition of visual objects. Paradise used methods for feature extraction, pattern recognition and classification of image data sets. Although the techniques used were complicated, the authors reported they were able to create computer-based self-learning keys, ANNKEYs, which could be used for identification when an expert was unavailable.

Although the studies reported on above appear to have had reasonable success, they have been conducted (for the most part) on relatively small sets of input data. They cannot identify previously unseen data, and misidentification or no identification is made when data are sparse. On the other hand the SOM is able to handle data sets which exhibit such characteristics [103, 114, 128, 185, 253]. A number of studies have been undertaken specifically using SOMs on biological data sets, and some of these studies are discussed below. Many of these studies did not report on specific accuracy levels obtained or how the comparisons between different methods were assessed. Hence it is difficult to judge the true value of these studies.

Laitinen *et al.* [123] use SOMs for the visualization of the size and shape of particles. This study used model image analysis of a series of particles to obtain shape and size particle parameters. These data were then used as input for the SOM algorithm and principal component analysis (PCA). The reported results obtained using PCA were not very good as the method was unable to separate some of the clusters. Although the SOM technique was able to separate the clusters of particles, the output still needed to be analyzed and interpreted by an expert.

In another study, using a data set on the distribution of trees, Giraudel and Lek [82] compared SOM results with those obtained by using PCA, correspondence analysis (CoA), polar ordination (PO) and non-metric multidimensional scaling (NMDS). Traditional statistical methods confirmed the accuracy of the results obtained with the SOM, although the authors concluded that the comparison of methods is not a trivial task.

Blayo and Demartines [20] also compared results obtained by applying the SOM algorithm with the results obtained by using PCA and the generalized Hebbian algorithm (GHA). These authors concluded that when comparing complex non-linear data a direct comparison of results on its own is not sufficient.

Céréghino *et al.* [29] and Park *et al.* [162] applied the SOM algorithm to an environmental data set for predicting the species richness of aquatic insects in streams. The SOM was used to classify the stream sampling sites according to the environmental variables. The MLP was used in a second phase for predicting species richness. The authors found that these methods complemented each other and suggested that for ecological modelling the combination of methods could be the preferred procedure.

Samsonova *et al.* [187] used an enhanced SOM to perform cluster analysis on a protein data set. The enhancement tools were used to determine cluster confidence levels and to visualize the results as a tree. The authors felt that visualizing results as a tree structure would facilitate comparison with existing hierarchical classifiers.

Fernández *et al.* [71] used the SOM, the MLP and a network based on the adoptive resonance theory (ART) for animal science applications.

Schreer *et al.* [192] applied various algorithms (including k-means and fuzzy c-means clustering techniques, SOM and ART) to data sets of dive profiles for penguins

and seals. The authors found that although SOM, c-means and k-means performed as well as each other, the k-means technique provided results that were more logical and readable, and for dive profiles it was the method of choice.

Li [133] developed a remote login, interoperable SOM data mining system called iSOM (based on Kohonen's SOM). The system was tested using the *Iris* data set [11] as well as air pollution data. The cost function, J_{MCR}, for calculating the misidentification rate when using the *Iris* data set with nine clusters was reported by Li as 0.02 using the following formula:

$$J_{MCR} = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{m_i}{n_i} \qquad \text{Eq. 3-5}$$

where

n_c = number of clusters,
m_i = number of patterns in cluster *i* misidentified, and
n_i = number of patterns in cluster *i*.

The original Kohonen SOM network was extended by Kiang [106] to include a contiguity-constrained clustering method [150] to perform clustering based on the output map generated by the network. The results Kiang obtained with the *Iris* data set using the extended SOM method and a minimum variance criterion gave a 90.34% rate of correctness.

More recently, SOMs have been used successfully for analyzing and visualizing massive gene expression data. Tamayo *et al.* [213] tested the SOM's usefulness for analyzing yeast cell cycle data by developing a computer package called GENECLUSTER which used the SOM algorithm to cluster and display their data. These authors also tested their methods on data provided by Chu [33] and obtained results similar to those reported by Chu.

Törönen *et al.* [219] used a tree-structured SOM algorithm [120] combined with Sammon's mapping algorithm [186] to analyze published expression data from 6400 yeast genes at 7 different time points during a growth phase. Sammon's algorithm was applied so that the relationship between the individual neurons could be visualized more clearly. Törönen *et al.*'s results, as did Tamayo *et al.* [213], showed that the use of SOMs is useful for the organization and interpretation of mammalian gene

expression data. However, successful identification statistics were not directly reported in either paper.

The SOM was also used by Nikkila *et al.* [153] for performing yeast gene expression analysis and visualization, and the results obtained were compared with those obtained by using multidimensional scaling (MDS) and hierarchical clustering. The results obtained by using the SOM tool were consistent with existing knowledge of the functional classes of genes and were generally found to be more trustworthy than those obtained from using the other two clustering methods.

A comparison between the SOM and PCA was done by Brosse *et al.* [23]. The data set used consisted of 710 samples and 15 species of European freshwater fish. The authors found that the SOM was able to visualize the entire fish assemblage in a 2-dimensional space for both dominant and scarce species. They also reported that the PCA method provided irrelevant information for some scarce species.

Walley and O'Connor [238] also used ecological data to compare a non-neural self-organizing map based on information theory and Kohonen's SOM. The system developed for this work was called MIR-max (Mutual Information and Regression maximization). MIR-max separated the tasks of clustering and ordering into two different processes whereas Kohonen's SOM integrates clustering and ordering into one process using Euclidean distances and a neighbourhood function. The authors report that on average the clustering results of MIR-max were 18% and 16% higher than those produced respectively by SOMs and the generative topographical map (GTM) [18]. These improved results were gained at the expense of the MIR-max clustering phase being computationally demanding and the system being more complex. An additional disadvantage of this method is that the results showed that the SOM and GTM performed better than MIR-max with respect to ordering.

Shanmuganathan *et al.* [195] used the SOM algorithm to model environmental and economic systems.

Bernatavičienė *et al.* [17] used the SOM algorithm, combined with MDS and with Sammon's mapping, to visualize several biological data sets. These authors report that both the combined methods of SOM and MDS, and of SOM with Sammon's mapping displayed similar efficiency for clustering and visualization of data.

Goodacre *et al.* [84] used pyrolysis mass spectrometry (PyMS) to obtain high dimensional biochemical fingerprints from 4 species of plant seeds. These data were used as input for the unsupervised methods of self-organizing feature maps (SOFMs) and auto-associative ANN (a fully interconnected feedforward MLP) and the results were compared with those obtained by applying the statistical methods of PCA and the supervised method of canonical variates analysis (CVA, also referred to as discriminant analysis). The authors used the BP algorithm to train the auto-associative ANN. The auto-associative ANN and the SOFMs were both able to separate out the seed species, and the resulting groups were less subjective. The PCA and CVA methods were not able to differentiate between two types of seed species, and the CVA approach also had the disadvantage of requiring *a priori* information as to which input spectra are replicates.

Weller *et al.* [241] used SOMs to cluster images of dinoflagellate cysts, and the authors report accuracy rates up to 100% when the number of the principal characteristics presented to the SOM was increased.

Mangiameli *et al.* [140] compared the performance of the SOM and seven hierarchical clustering methods using 252 “messy” data sets (non-biological) with various levels of imperfections (including data outliers, irrelevant variables, dispersion, and non-uniform cluster densities). The authors found that the SOM results demonstrated superior accuracy and robustness when compared with the results of other cluster methods. However, the performance of each technique was only measured by the accuracy percentage of data points assigned to the correct cluster, and not by any of the other criteria that have been suggested (for example) by Tan and Gilbert [214] and Podgorelec *et al.* [166].

SOFMs have been integrated successfully with a rule-based expert system [220-222, 224]. Using a non-botanical data set, Ultsch and Vetter [223] also compared hierarchical clustering and k-means clustering with the SOFM using U-matrix techniques and found that SOFM performed better than the other two clustering techniques.

From the above mentioned examples it can be appreciated that SOMs are suited to exploratory data analysis, allowing one to impose partial structure objectively on the clusters and facilitating easy visualization and interpretation. This is in contrast to the

rigid structure imposed by hierarchical clustering, the strong prior hypotheses used in Bayesian clustering, and the non-structure of k-means clustering [213].

Also, it is felt that the nature of biological identification is such that unsupervised techniques are more suitable than supervised techniques as the data inter-relationships are important. Furthermore, it is possible that a data driven approach where a mathematically based system is left to determine the relationships might be more suited to the problem, and might even reveal new relationships previously not noticed. The SOM method has the added advantage that even when new, previously unseen data are tested against the trained map it is possible to get results without retraining the map. The SOM method is acknowledged as a method which can produce valid results even when sparse data sets are applied: in biology sparse data sets are often the norm.

According to Lisboa [136], there are two statistical methods which could add value to results. The one improvement is to add the ability to map accurately the features of the data that are difficult or expensive to find in a conventional statistical manner (for example, by providing the means for visualizing complex interactions between particular variables or attributes). The second improvement is to add substantially to the power of exploratory data analysis (for example, by raising hypotheses about unsuspected non-linear components whose explicit modelling may improve the accuracy of standard statistical methods, or by providing direct visualization of complex high-dimensional data). It is felt that the SOM may be used to fulfil both of these improvements.

In addition, even though combined methods have been shown to provide accurate and efficient results, it is felt that initially the extra complexity and time required for using combined methods might not be necessary. The SOM is simple to implement and needs to be applied on its own to the data set and tested thoroughly before applying a combined method.

Although, Kohonen's SOM has been applied successfully as a classification tool to various biological data sets, after a thorough investigation of the literature, as far as can be determined, the SOM's potential as a tool for application to large botanical data sets with a wide range of morphological characteristic states remains relatively unresearched. Certainly this appears to be the case within southern Africa.

3.5 Other Algorithmic Solutions

There is a great need for intelligent methods which can extract meaningful information from enormous amounts of data: as is found in biology. Many of the methods that have been developed have their origins in artificial intelligence and machine learning [144]. Machine learning is a diverse field linked by common goals and similar evaluation methods. The general aim of machine learning is to develop computational methods to improve the performance of a task by automating the acquisition of knowledge from experience. Since expertise requires extensive domain specific knowledge, the overall purpose of machine learning is to provide a means of releasing human experts from performing time-consuming activities which can be automated, thus leaving the experts to perform other tasks that cannot easily be performed by other persons and/or means. The general approach of machine learning involves using algorithms to find and exploit patterns in the input data. These algorithms have to be accurate and efficient. Many AI systems have been produced as potential substitutes for experts and are in regular use [125].

One of the paradigms for machine learning is rule induction (RI) which uses inductive inference to extract rules from a set of observations. The goal of inductive inference is to learn how to classify objects by analyzing a set of instances whose classes are known. Typically, instances are represented as attribute-value vectors, and learning input consists of a set of these vectors, each vector belonging to a known class. The output consists of a mapping from the attribute values to the classes, and by using this mapping one should be able to classify accurately both the given instances and any other unseen instances [165]. A decision tree is a formalism for expressing such mappings, and consists of test or attribute nodes linked to two or more sub-trees and leafs. A leaf is a decision node labelled with a class which is the decision [171].

The Quinlan family of decision tree algorithms include the ID3, C4.5 and C5 algorithms [169, 171-173] and form popular standards for RI. These algorithms will be discussed briefly in the following subsections.

3.5.1 C5 Decision Tree Algorithm

For a description and examples of decision trees there are many references in the literature [169, 182, 184, 228, 256] (and many more), and only a brief outline is given here.

Decision tree algorithms follow a top-down, divide-and-conquer induction process. The basic algorithm (based on the Quinlan model [169, 171-173]) for decision tree induction can be described as follows [85]:

- Using an information gain measure [3, 171], select an attribute to place at the root of the tree and create a branch for each possible value of the attribute. The underlying data set is thereby split up into subsets, one for each value of the attribute being investigated.
- This process is repeated recursively for each branch, using only those cases that actually reach that branch. The branches are connected by internal nodes that represent an attribute test; and each branch from that node represents an outcome of that test.
- If all instances at a node have the same classification, development stops on that part of the tree and a leaf (= terminal) node is formed which names the class.

Once induced, a decision tree can be used to classify target instances by starting at the root of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate sub-tree. The branches of the sub-tree are investigated again until a node is found that is not a test. Such a node is called a leaf. The conditions of the leaf must match those of the target instance, and the label of the leaf gives the predicted class of the target instance.

The most popular decision tree algorithm [64] was developed by Quinlan [169] and employed a greedy algorithm strategy. It was called the ID3. Changes to the ID3 algorithm resulted in the development of the C4.5 algorithm [171]. This algorithm was later further developed and refined to form the C5 algorithm [172, 173]. C5 assesses attributes and their values, and allows for tree simplification, pruning, bagging and boosting.

The output of the C5 algorithm is a decision tree, but the algorithm is also accompanied by a program to convert the decision tree into an optimal set of rules.

This program first generates a rule for each leaf node, using all conditions in the path from the root to that leaf node. The C5 rule program then applies pruning methods to get the smallest possible set of rules with the same or higher accuracy as the decision tree [181]. According to Quinlin [170], when classifying unseen samples a final set of production rules is usually both simpler and more accurate than the decision tree from which it was obtained. In addition, production rules provide a way of combining different decision trees for the same classification domain.

C5, or its variants, have been used frequently in medical informatics research [3, 85, 134, 165, 166, 245] and in non-medical biological identification [24, 151, 152, 188, 253], to cite a few references.

According to Quinlan [169], decision tree results are categorical and thus do not convey potential uncertainties in classification. This is a serious disadvantage, for minor differences in attribute values of a sample being classified may result in incorrect changes to the assigned class. Samples with missing or imprecise information may not be classified at all. Despite these limitations Quinlan's C5 algorithm is still recognized as a state-of-the-art decision tree algorithm and will therefore be used to test the validity of the SOM results.

The CN2 algorithm [22, 38-40] is another example of an algorithm that utilizes RI and will be discussed briefly in the next subsection.

3.5.2 CN2 Rule Induction Algorithm

The CN2 algorithm was designed by Clark and Niblett [39]. This algorithm inductively learns a set of propositional *if...then...* rules from a set of training examples [41]. In order to find rules the CN2 algorithm performs a general-to-specific beam search through the rule-space looking for the *best* rule and then removing the training examples covered by that rule. A control algorithm is used for re-executing this search until no more *good* rules can be found. The output from CN2 is thus a set of production rules and not a decision tree.

The original CN2 algorithm [39] defined the best rule by using a combination of entropy and a significance test. Later the CN2 algorithm was improved and the evaluation function was replaced with the Laplace estimate [40]. In addition, the original algorithm produced rules as an ordered set. This means that the rule

classifying the most records in the training data set is given first while the rule classifying the least number of records in the training data set is given last. However, the newer version of the CN2 algorithm has been improved and is able to induce unordered rule sets as well as ordered rule lists.

The original CN2 algorithm and its later versions have been used frequently for biological classification, identification and diagnosis. These references include [126, 127, 156, 206], to cite just a few.

A disadvantage of the standard CN2 algorithm [126] is that in order to allow for handling imperfect data, the algorithm may construct a set of rules which is imprecise and therefore does not classify all examples in the training set correctly. In addition, some of the induced rules are not natural as a lot of information needed for classification/identification is missing in the induced rules, making the interpretation of induced rules difficult. Also, with the CN2 algorithm, if none of the rules fire, a default rule which predicts the majority class of the uncovered training instances is invoked. This can give incorrect results and, in the context of this thesis, this means misidentification.

According to Lavrac *et al.* [127] the standard CN2 algorithm needs adjustments to improve the number of induced rules, the rule coverage and the rule significance. These authors also maintain that in the classical covering algorithm only the first few induced rules are of interest as subgroup descriptors with sufficient coverage. The authors state that subsequent induced rules are induced from biased example subsets which only include positive examples not already covered by a rule. Lavrac *et al.* claim that this bias makes it unlikely that new subgroups will be discovered. This would be a disadvantage in the current study as this thesis aims to try to discover interesting properties of subgroups of the data set.

Despite these disadvantages CN2 is a popular rule induction algorithm and will be used (along with the C5 algorithm as mentioned earlier) for comparison and confirmation of the results obtained by applying the SOM.

3.5.3 Other Techniques used for Biological Identification

It has been argued that sound statistical principles are essential if the evidence base built with any data-based methodology, including an ANN, is to be trusted. It is

argued further that these methods are best justified where they provide additional functionality to the performance of well-established statistical models [136, 193]. However, it is not within the scope of this thesis to investigate all the methods that have been tested in order to try to solve the problem of biological identification. Many of the methods investigated recently have concentrated on gene analysis. Some of the research that has come to the writer's attention includes: Ultsch's PUL information retrieval algorithm [227], Madeira's work on biclustering [138], Au *et al.*'s work on attribute clustering [14], and Eisen and Spellman's hierarchical clustering [61, 209].

It has been reported that phylogenetic trees impose strict hierarchical structure and are best suited to data that tend to have this structure naturally [213]. The tree data used for this research do not have natural hierarchical structures, and therefore any method that imposes a hierarchical structure should not be used for tree identification. Hierarchical clustering modelling methods do not show the multiple distinct ways in which the data can be similar, and this certainly would be a disadvantage when seeking relationships between different species of trees. On the other hand, SOM output results can clearly demonstrate the inter-relationships between the data even for massive and complex data sets. Hierarchical clustering has been noted [213] to suffer from lack of robustness, non-uniqueness and inversion problems that make interpreting the results difficult. In addition, hierarchical clustering may group data based on local decisions and does not allow for re-evaluation of the decisions used for performing the clustering [213]. This may result in misidentification.

Bayesian clustering is a highly structured approach requiring strong prior hypotheses, while K-means clustering is a completely unstructured approach which proceeds in a local fashion and results in an unorganized collection of clusters that is difficult to visualize and interpret [213]. The SOM imposes partial structure and therefore treads a middle path between these two extremes.

3.6 Conclusion

This chapter presented an overview of the different AI paradigms that were applied or were considered for implementation of the experimental work in this study. A brief outline of some of the AI techniques that were considered, namely expert systems and fuzzy expert systems, was given. Next, the structure of a neural network and relevant

training approaches were defined, followed by a broad overview of self-organizing maps. Thereafter, some RI techniques were described, and in particular two machine learning algorithms, the C5 decision tree algorithm and the CN2 rule extraction algorithm, were outlined.

The overall objective of this chapter was to identify advanced AI techniques that could potentially aid in the problem of botanical identification. After studying these techniques the decision was made to use the SOM algorithm for application to the problem of identifying tree data.

The results obtained from applying the SOM algorithm to the tree data set will be compared to those obtained by applying the C5 and CN2 algorithms to the same data set and analyzing the results. It is felt that the decision to use the CN2 and C5 algorithms is justified because both these algorithms are popular *bona fide* computer techniques and are state-of-the-art methods from different classes of algorithms. If either (or both) of these techniques produces meaningful results it will be possible to compare the results with those obtained from the application of the SOM. On the other hand, should the CN2 and/or C5 algorithms fail to perform adequately, the anticipated superiority of the SOM for biological identification would be highlighted.

Some of the papers discussed in this chapter report work that has been performed with neural networks using small historical data sets (for example, the *Iris* data set) and comparing different systems using accuracy measures only. It has been argued by some that accuracy measures on their own are not sufficient when comparing different systems. In this research, evaluation of results obtained from the application of the SOM will also be done using accuracy, sensitivity and specificity measures. In addition, multi-class and cluster confusion matrices will be presented. Finally ROC space graphs will be drawn to help evaluate the models.

In the next chapter the SOM technique is discussed in detail.

Chapter 4

The Self-Organizing Map: The SOM

The previous chapter discussed different artificial intelligence approaches to the problem of biological identification. It was concluded that self-organizing maps (SOMs) were likely to be a useful method for application to this problem. The SOM algorithm will be described in detail in this chapter.

The origins of the SOM are discussed in Section 4.1. Section 4.2 describes how the SOM works: first the original SOM algorithm is presented in Section 4.2.1, next the batch algorithm is described in Section 4.2.2, and some variants and related algorithms are briefly introduced in Section 4.2.3. Visualization of the SOM is discussed in Section 4.3. Problems associated with the SOM algorithm are discussed in Section 4.4: missing data are reviewed in section 4.4.1 and outliers in 4.4.2. Two measures of SOM quality are discussed next in Section 4.5. First, quantization errors are described in Section 4.5.1 and then topographic errors in Section 4.5.2. The chapter is concluded in Section 4.6.

4.1 Origin of the Self-Organizing Map Technique

The idea of using the SOM as it is applied in this thesis was conceived by Kohonen in 1981 when he suggested using ordered displays to 'illustrate' a data set [109, 110, 114]. Kohonen's idea was inspired by the work of Von der Malsburg [236]. The SOM forms a nonlinear, nonparametric regression (*viz*, methods used for describing relationships between the dependent and independent variables without specifying the form of the relationship between them *a priori*.) of an ordered set of model vectors to the distribution of input vector patterns. The model vectors form an "elastic network" that maintains the topological order of the input data and develops into specific identifiers of the respective areas in the input space. The process steps by which the "elastic network" is formed are defined by the SOM algorithm. According to Kohonen, in its basic form the SOM can be said to produce a similarity graph of the

input data. The SOM does this by taking the input patterns and compressing them onto a set of model vectors while preserving the most important topological relationships of the patterns before displaying the output, usually in the format of a two-dimensional grid. Thus the SOM is more than a clustering method: for it can be used to reduce the amount of data by clustering, while at the same time it can project the nonlinear mappings of the input data onto a lower-dimensional display. In the process the probability density function of the input space is approximated and the topological structure of the input space is maintained.

4.2 How the Self-Organizing Map Works

The essence of the SOM algorithm is that it trains the network to learn to recognize input data while preserving the topology of that data. The SOM training utilizes a competitive learning strategy during which a weighted vector associated with each neuron in a neural network is modified and is gradually developed to become sensitive to a set of patterns from a specific domain of the input space. The end result of the training process is that different neurons specialize to represent different types of input patterns. This specialization is enforced by competition among the neurons. The competition occurs when an input pattern is presented to the network, and the neuron that is best able to represent the pattern ‘wins’ the competition and is rewarded by being allowed to adjust its vector values in order to represent the input pattern even more closely. If the winner’s neighbouring neurons are allowed to learn, those neurons will also gradually specialize to represent similar patterns, and consequently the representations on the output layer will become ordered [104]. With the SOM it is crucial that the neurons doing the learning do so as a group and not independently of each other, i.e. the neurons must learn as topologically related subsets. The adjustments performed during the learning steps become smoothed out during the iterative process.

4.2.1 The Original Self-Organizing Map Algorithm

The original incremental SOM algorithm defines a special recursive regression process where only a subset of models is processed at every step [114]. The first step

in the SOM training process is to define a map structure. It is possible to create a multi-dimensional lattice structure [108], but complex structures are not generally utilized as visualization becomes difficult [234]. Usually the neurons are arranged on a regular 1- or 2-dimensional lattice type of array with a hexagonal and oblong arrangement. This type of arrangement is able to represent the data clusters better than a rectangular arrangement and fits the data input distribution more easily [57].

The number and positions of neurons on the grid are defined and fixed when the map is created and depend on the purpose for which the SOM will be used, and on the amount of input data. Sometimes the number of neurons used is determined by a heuristic formula such as $5\sqrt{N}$ or $\sqrt{N}$, where N is the number of training patterns [234, 235]. The complexity of the SOM algorithm is governed by the number of neurons used: the more neurons that are used the longer the training process takes and the greater is the memory requirement. If the natural number of clusters in the data is being investigated (as it is in this thesis) the number of neurons used in the trained SOM must be far larger than the expected number of clusters in the data [234] but less than the number of training patterns [65].

A vector of variable scalar weights is associated with each neuron. The dimension of these vectors is the same as the dimension of the input data vectors. The vectors associated with the neurons are referred to as the reference, model or prototype vectors. The values with which the model vectors are initialized can influence the final states of the map as well the learning speed (by ensuring fast convergence), and it is usually preferable that the vectors are initialized in an orderly manner rather than randomly [114]. With random initialization the self-organizing process may take a long time and a wide neighbourhood function (discussed below) may be necessary initially, resulting in heavy computation. If the model vector values are ordered the rate of convergence is faster, smoother and more reliable, and a narrower neighbourhood function can be used to obtain a more stable map [104, 114]. In addition to the model vectors being initialized in an orderly way, generally the weights in each vector are different for different neurons. Various initialization methods have been proposed and discussed in [13, 65, 114].

If a model vector is denoted by $\boldsymbol{m}_i$ a convenient measure of the match between a vector $\boldsymbol{m}$ and an input vector $\boldsymbol{x}$ can be based on the Euclidean distance between these vectors. The objective of competitive learning is to determine which neuron best represents the input $\boldsymbol{x}$. Thus, the best matching neuron (BMN) would be the neuron associated with the model vector ($\boldsymbol{m}_c$) which satisfies the following equation:

$$\left\| \mathbf{x} - \mathbf{m}_c \right\| = \min_i \{ \left\| \mathbf{x} - \mathbf{m}_i \right\| \} \qquad \text{Eq. 4-1}$$

If a pattern $\boldsymbol{x}$ has some missing values those variables are ignored in the distance calculation.

There are several measurements of similarity (or dissimilarity), but the simplest one is the Euclidean distance formula which is used to measure distances between patterns. This is the metric that is widely used with SOM [114, 233]. The minimum distance as determined by the Euclidean distance metric defines the winner ($\boldsymbol{m}_c$), and this model vector is then used to represent the input $\boldsymbol{x}$ and is rewarded by having its values updated to be closer to the values in $\boldsymbol{x}$. However, as stated above, learning must not happen in isolation from the surrounding neurons. A neighbourhood set, N_c, is defined around a neuron. At each learning step all the neurons within the neighbourhood of $\boldsymbol{m}_c$ are updated by having their values adjusted, although not to the same degree as $\boldsymbol{m}_c$. All $\boldsymbol{x}$ patterns which are best represented by $\boldsymbol{m}_c$ will select $\boldsymbol{m}_c$ as the winner and are thus mapped to it. The end result is that each model vector specializes to represent a whole domain of the input space and the "elastic net" formed takes the shape that best fits the patterns. Those neurons that are topographically close to each other in the array will activate each other to learn something from the same input pattern $\boldsymbol{x}$. This will result in a local smoothing effect on the model vectors of the neurons in that neighbourhood, and as further learning takes place this process leads to global ordering.

The size of the neighbourhood can vary: initially it should be set very wide to encourage global ordering (to give a 'zooming out' effect corresponding to a coarse global resolution showing a global view), and later it should be decreased (to give a 'zooming in' effect corresponding to a closer view with finer cluster boundaries becoming evident). The final width of the neighbourhood is important because during

the final stages of the map formation the accuracy of the map and the degree to which the map follows the local data structures is determined.

After selecting the BMN the model vectors are updated. The amount of the update is controlled by a neighbourhood function which is a decreasing function of the distance of the neighbourhood neurons from the winning neuron. The update rule [114] for the model vector *i* is:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)\left[\mathbf{x}(t) - \mathbf{m}_i(t)\right] \qquad \text{Eq. 4-2}$$

where

t denotes time,

$h_{ci}(t)$ is the neighbourhood function, which is usually a function of the distance between the locations of the neurons on the map grid such that if $\boldsymbol{r}_c$ and $\boldsymbol{r}_i$ are the locations of neurons c and i respectively then $h_{ci}(t) = h(\|\mathbf{r}_c - \mathbf{r}_i\|, t)$ and with increasing $\|\mathbf{r}_c - \mathbf{r}_i\|$, $h_{ci} \to 0$.

For convergence, it is necessary that $h_{ci}(t) \to 0$ when $t \to \infty$. The neighbourhood function has its largest value for the winning neuron and decreases monotonically with increasing distance on the map grid $\|\mathbf{r}_c - \mathbf{r}_i\|^2$.

The neighbourhood can be defined as a Gaussian kernel [114]:

$$h_{c(x),i} = \alpha(t) \exp^{-\frac{\|\mathbf{r}_i - \mathbf{r}_c\|^2}{2\sigma^2(t)}} \qquad \text{Eq. 4-3}$$

where

$\alpha(t)$ is the learning rate factor, and

$\sigma(t)$ is the width of the neighbourhood radius ($\sigma(t)$ is the same as $N_c = N_c(t)$).

Both the learning rate $\alpha(t)$ and the neighbourhood radius $\sigma(t)$ decrease monotonically during training. The learning rate decreases to zero, and the neighbourhood radius decreases to a non-zero number such as one. The exact values of N_c and α are not critical if the model vectors are initialized with ordered values. However, the final value of $\sigma(t)$ in the last training cycle influences the shape of the

SOM: a high final value of $\sigma(t)$ emphasizes the topological relationship between the model vectors at the expense of the quantization effect, whereas a low final value of $\sigma(t)$ emphasizes the quantization effect at the expense of the topological relationships. The process of updating each model vector by presenting all the data vectors once is called an epoch. In practice the number of epochs can be fairly low.

The updating process is illustrated in Figure 4-1 where the neighbouring model vectors are pulled in the same direction (because of the neighbourhood relations). The neighbouring neurons acquire similar model vectors and the map adjusts to the data by updating these model vectors.

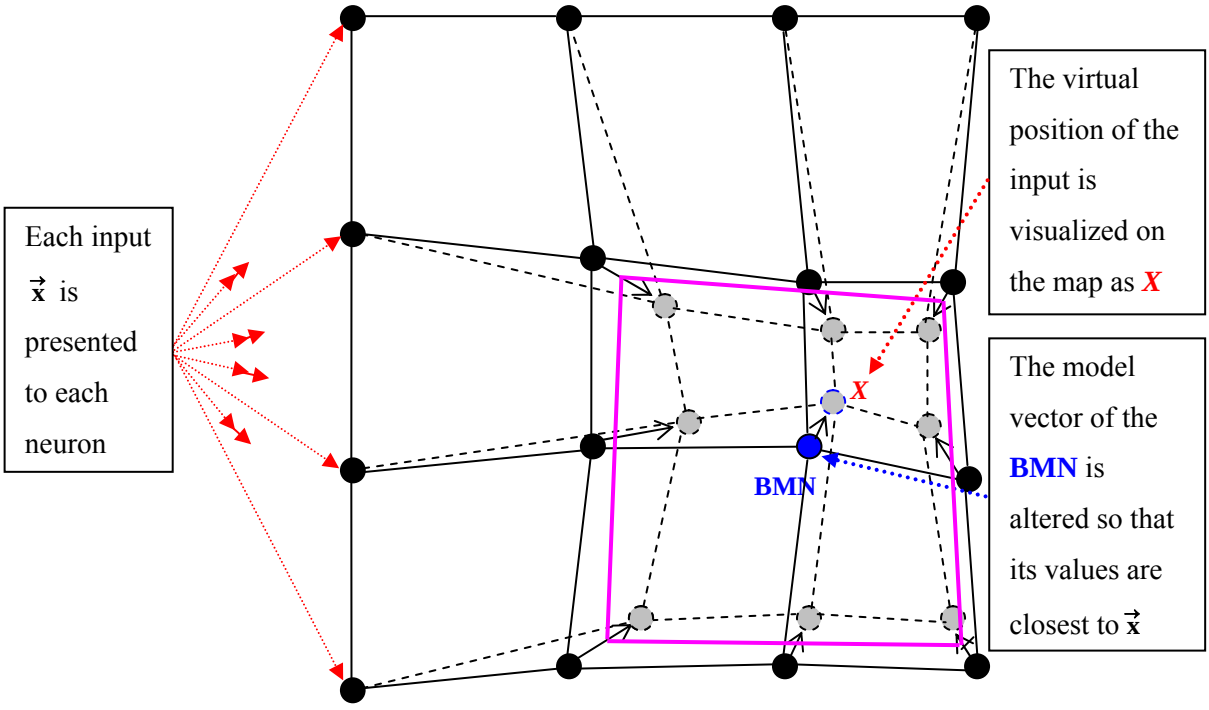


Figure 4-1 : Update Process of the Best Matching Neuron and its Neighbours
The input $\vec{x}$ is presented to each neuron in parallel and the Euclidean distance between each associated model vector and the current input is calculated in order to find the model which best matches the input. The vector values of the neurons in the neighbourhood of **BMN** are altered (◌) to represent $\vec{x}$ more closely but to a lesser extent than the vector associated with **BMN.** The neighbourhood is enclosed by purple lines (adapted from Vesanto, [234, fig. 3.2, p13]).

4.2.2 The Batch Self-Organizing Map Algorithm

The original SOM training algorithm is stochastic where the model vectors are updated after each pattern has been presented to the network. In the SOM batch training algorithm the entire data set is presented to the SOM before any updates are made.

The speed of training can be improved by approximately an order of magnitude by using the batch version of the SOM algorithm rather than the original SOM. In the batch map algorithm no learning rate parameter need be defined, but both the original and batch algorithms require the definition of a time-dependent neighbourhood function.

The batch version of the SOM may be summarized by the following steps [113]:

1. Initialize the model vectors using a suitable method.
2. Compute for each neuron the average of the data patterns for which that neuron is the BMN.
3. Denote this average for neuron j as $\bar{\mathbf{x}}_j$.
4. Calculate the new model vector ($\boldsymbol{m}_i$) values using the following formula:

$$\mathbf{m}_i = \frac{\sum_j n_j h_{ji} \bar{\mathbf{x}}_j}{\sum_j n_j h_{ji}} \qquad \text{Eq. 4-4}$$

where

j iterates over each neuron of the SOM,

h_{ij} is the neighbourhood function, and

n_j is the number of patterns for which neuron j is the BMN.

5. Test for convergence and go back to step 2 and continue training until the algorithm converges.

A version of the batch SOM algorithm is presented in Appendix B.

4.2.3 Variants and Related Algorithms

Kohonen's SOM model requires that the structure and size of the map are defined before a map is created. This places some limitations on the resulting maps. A number

of variations of the model have been proposed concerning the topology and the number of neurons.

The choice of the size of the map is important because too many neurons may cause overfitting of the training patterns with the result that many neurons have a low frequency (where the frequency of a neuron is the number of patterns for which that neuron is the BMN). In addition, too many neurons increase the computational complexity. On the other hand, too few neurons will result in clusters with a high variance among the cluster members.

The structure of a map may be made more flexible with the aim to improve the preservation of the topology, but sometimes this makes visualization more complicated than it would be when using a fixed grid [104]. There are many and various types of adaptations to the basic SOM and batch SOM algorithms which have been developed, and some of these are discussed below.

Growing Variants of the Self-Organizing Map

One approach in an attempt to optimize map size for the SOM is to allow the number of neurons used to vary. Training begins with a small number of neurons and the map is allowed to grow and shrink during training as neurons are needed.

The interpolative method described by Rodrigues and Almeida [176] allows the use of a variable number of neurons, but the structure of the map has to be predefined. The new neurons have a well defined place on the low dimension grid and so visualization is not a problem [233].

Engelbrecht [65] suggests a SOM growing algorithm, however, a square map structure is assumed.

Fritzke [74, 75] used a flexible and compact structure with a variable number of neurons. The growing cell structure method allows for the cell structure to be determined automatically from the input data and for the network size to be determined dynamically. New model vectors are added according to an error function criterion. The algorithm allows for model vectors to be removed as well as added. However, visualization is more complicated than it is with the regular SOM [104, 233].

Fritzke also developed a growing structure grid which changed dynamically and in which, according to [233], visualization is not a problem.

Blackmore and Miikkulainen [19] describe an incremental grid–growing algorithm which incorporates implicit input information directly into the structure of the map, and in the process represents it explicitly in the output. Blackmore and Miikkulainen's aim is to guide the development of structures actually present in the input distribution, and during organization to detect and correct as early as possible any false topology in the map. The authors suggest this requires an incremental approach to building and organizing the map. Initially a small number of neurons are utilized in the structure and then heuristics are used to find and remove any potentially inaccurate neurons or connections, or to add neurons where they are required. After the reorganization of the structure the process is repeated until the specified maximum number of neurons is achieved. The algorithm can yield an accurate, low-dimensional description of the structure in high-dimensional input.

Another growing variant of SOM was presented in [99] which leads to networks with rather complicated structures. Visualization, however, is not a problem with this technique.

A growing grid was presented by Bauer and Villmann in [15] where the output space topologies are adapted in an unsupervised way. This is accomplished by growing hypercubical output spaces up to a pre-specified maximum number of nodes. This Growing Self-Organizing Map (GSOM) starts with a configuration of two neurons, learns using the regular SOM algorithm, and adds neurons to the output space according to some criterion. Growing is achieved either by adding nodes in one of the directions already spanned by the output space, or by adding a new dimension. This process of learning and adding is repeated until a pre-specified maximum number of neurons is reached. The maps formed by this method deliver an output space topology adapted to the input data. The neighbourhood is thus well preserved, but the output space is constrained by being forced to maintain a hypercube shape.

The Growing Hierarchical Self-Organizing Map (GHSOM) [143, 175] has a hierarchical architecture composed of independent growing self-organizing maps. During the unsupervised training process the architecture of the model is adapted

according to the structure of the input data. This is done by allowing the size of maps and the depth of the hierarchy to adapt dynamically. The layers of the GHSOM grow in a top-down fashion. Starting at the top layer, each map grows in size to represent the data set to a specific level of detail and is then analysed. The map neurons that need expansion (because they represent an inhomogeneous set of input data) are developed into a new SOM in a lower hierarchical layer in order to represent the data better. The whole process is then repeated until a suitable level of representation is reached.

The Plastic Self-Organizing Map (PSOM) [124] is another dynamic SOM variant which adds new neurons in order to capture new information and deletes neurons that store stale information. Parameters are set without prior knowledge of the data set and require tuning by running multiple trial sessions.

Another Growing Self-Organizing Map (GSOM) was described by Alahakoon *et al.* [5-8]. This dynamically growing neural network does not require that a map size be specified, but rather can be dictated by the user or the structure of the input data. This algorithm incorporates a parameter called the spread factor (SP) which controls the map resolution. Initially, if a low SP value is used, a coarse map is obtained. After filtering out the data that belongs to a cluster a larger SP value can be used to zoom in and exam any sub-clusters.

The Cellular Probabilistic Self-Organizing Map (CPSOM) [32] is an online algorithm which has its foundation in statistical analysis. It is a learning algorithm which allows the network to adapt to new patterns and includes a forgetting factor (FF) which allows the network to forget stale information. The FF makes the algorithm better able to produce flexible dynamic maps.

The Growing Cellular Probabilistic Self-Organizing Map (GCPSOM) is a hybrid of the GSOM and the CPSOM, and is described in [8].

Tree-structured Self-Organizing Map

In the SOM the search for the BMN can be speeded up by constructing a tree-structured SOM [117-120] (TS-SOM). Each layer (level) of the tree is a complete quantization of the data set and consists of a separate, progressively larger SOM. The

size of the data set is limited by using the information gathered in the previous layer. The search for the BMN proceeds layer by layer, each time restricting the search to a subset of neurons that is dictated by the location of the BMN in the previous layer. The map is taught layer by layer, starting from the smallest layer. The tree structure offers $O(\log N)$ search complexity instead of $O(N)$, and so provides a fast search.

Minimal Spanning Tree Self-Organizing Map

Kangas *et al.* [101] proposed a minimum-spanning-tree approach (MST-SOM). In the MST-SOM the neighbourhood relations are defined using a MST which finds the shortest possible set of connections linking a set of vectors. In terms of vector quantization the dynamically changing structure of the MST-SOM is faster and more stable than the basic SOM. However, in the MST-SOM the model vectors do not have well-defined positions on a low-dimensional map. Thus the speed up of the algorithm is at the expense of neighbourhood relations, and visualization is more of a problem than it is with a regular grid.

Neural Gas

The neural gas algorithm [141] is another variant of the SOM. Here the neighbourhoods are adaptively defined during training by ranking the distance of model vectors from the given training pattern. This SOM variant preserves the neighbourhood relations but does not always reduce the dimension of the input data. Again, visualization is more complicated than with the regular SOM.

Growing Neural Gas

The Growing Neural Gas (GNG) model developed by Fritzke [76] used competitive Hebbian learning in a similar way to the neural gas model described above. The GNG model starts with a small network and allows the addition of new neurons. These neurons are added after evaluating local statistical metrics collected during prior adaptation steps. The network topology is built up in small, incremental steps. The dimension of the network depends on the input patterns and may vary, and therefore need not be pre-specified. The growth of the network continues as training progresses

until a pre-defined maximum network size or a user-defined performance criterion is reached. The advantage of this method is that it uses constant parameters and is capable of dynamic data clustering. The disadvantage is that it tends to be hard to visualize because of the topology dimensions which may vary locally.

Multiple Self-Organizing Maps

The Multi - Self-Organizing Map (M-SOM) architecture which was described by Goerke *et al.* [83]. This method consists of a set of independent SOMs which work together while covering the input space. Each SOM is topologically distinct and has its own size and dimension, and each output class is represented with a separate SOM. The small size of each map ensures that it is less likely to twist and distort than a normal map which covers the entire input space.

4.3 Visualization of the Self-Organizing Map

The SOM has properties of vector quantization, clustering and projection algorithms. Quantization of the input patterns to the model vectors reduces the data set to a smaller set. After quantization the density of the model vectors should represent the input data's density and so can be used for clustering, visualization and analysis. The reduced data set has the added benefit that the computational complexity of subsequent tasks is reduced. In addition, quantization can help reduce the effect of outliers (discussed in Section 4.4.4).

As mentioned above, the density of the model vectors of an organized map reflects the density of the input data. The model vectors are far apart in the areas between the clusters and close to each other in the clustered areas. Hence the distances between the model vectors can be used to demonstrate the cluster structure of the input data.

To obtain efficient visualization the models require vector projection. Projection methods try to find low-dimensional mappings that preserve the order of distances between the originally high-dimensional data patterns. Reliable projections can be obtained if the characteristics are highly correlated with a lot of redundant information, or if the data contains a lot of noise which can be discarded [231, 232,

234]. The models (which have well-defined positions on the low-dimensional grid) and their projections form a map of the higher-dimensional input space.

Neighbourhood relations are an essential part of the organization of a SOM, however, several potential disadvantages of a SOM can arise relating to the neighbourhood.

The neighbourhood definition is not symmetrical on the map borders, i.e. the neurons located at the edges or corners have fewer direct neighbours than the neurons located elsewhere on the map. The result of this 'border effect' is that during training the properties of border neurons are different and the density estimation is different for the border neurons than for the map's central neurons. The occurrence of the border effect increases the probability of topological errors [226]. Ultsch suggests the border effect can be avoided by embedding the grid in a finite but borderless space [225] and that the maps should be unbounded, i.e. the maps should be folded back on themselves.

The vector quantization procedure performs averaging, and this is enhanced by the neighbourhood function. This could result in extreme values (belonging to 'outliers' - see Section 4.4.4) being 'averaged' out. If these outliers contain important data which should be analyzed, their removal is a disadvantage.

Another effect that could be a disadvantage occurs when interpolating units are placed between data clusters in non-continuous data space. These interpolating units are useful for extrapolation of estimates of the data distribution. Sometimes, however, this interpolation results in inaccurate information and obscures cluster borders.

A SOM may be visualized using various techniques which include visualization of cluster structure and shape, components and data on the map. The first step of the analysis of the map output is to try to determine how clusters relate to each other as this provides an overall idea of shape of the map in input space. This means that cluster boundaries have to be found. The unified distance matrix (U-matrix) can be calculated and used for finding and displaying these boundaries. The U-matrix is a graphical representation of the SOM where shades of a colour are used to show the distances between each model vector and its adjacent neurons. The U-matrix exploits the fact that the distances between neighbouring model vectors are not uniform:

distances are small in dense areas (lightly coloured) and distances are greater in sparse areas (darker shades of colour). This means that larger distances represent dissimilar features between neighbouring nodes and separate the clusters.

Boundaries of the clusters on the map are usually found by using Ward clustering of the model vectors [65]. This clustering method, based on variance, was proposed by a statistician named Ward and is one of the most popular hierarchical agglomerative clustering (HAC) algorithms. In this approach initially each neuron forms its own cluster. The algorithm iterates over the clusters merging the closest (according Ward's distance measure), adjacent, non-empty clusters until an end criterion is reached. This stopping criterion could be either an optimal or specified number of clusters. The final clusters formed eventually contain patterns with small variance over their cluster members but larger variance over other clusters.

Beside distance matrices, some of the different ways of visualizing the cluster structure include using similarity colouring and viewing the map network in 3-dimensions.

The component maps may be analyzed to investigate correlations or partial correlations between the component variables (attributes). Each component plane represents the values of a single component in each SOM neuron and thus depicts the range of values for that component. A range of colours is used to represent different values (from minimum to maximum) for the particular attribute being considered. Blue is used to represent low values while red is used to represent high values. Pale blue, green, yellow and orange respectively represent the increasing range of values between those of blue and red.

The relationship between the SOM and the data vectors may be used for determining the accuracy of the mappings. A particular pattern's location on the map is usually its BMN, although there may be several neurons with model vectors which match the pattern almost as well as its BMN. If there is a fold in the map these neurons may be far away from its BMN. The neurons may also be far away because the data patterns are far away from the data manifold modelled by the map [232]. The various possibilities have to be investigated and the possible reasons for the results analyzed.

4.4 Problems Associated with the Self-Organizing Map

The SOM method is an excellent technique, but there are several situations where problems may occur. The software must be able to handle these, and analysis of the results must take into consideration the possibility of results being affected. These problems are summarized briefly below.

4.4.1 Border Effect

The border effect is a side-effect of the application of the neighbourhood function and is a weakness of the SOM method. This effect, and possible ways to overcome it, has already been discussed in Section 4.3.

4.4.2 Interpolating Units

When the data cloud is not continuous, interpolating units are placed between the data clusters [234]. While these interpolating units may provide useful estimates of the data distribution, they can affect the shape of the data manifold. If necessary, these units may either be left out or at least taken into consideration when the results are analyzed.

4.4.3 Missing Data

If a significant proportion of the data set consists of missing data values the software must be able to handle the missing values. This can be achieved in a number of ways including the following:

a) When comparing the input data vector with the model vectors if some of the components of the input data vectors are unavailable, not applicable or undefined then only the known components of each input vector are taken into account and the vectors are re-dimensioned accordingly, i.e. the attributes with missing values are simply ignored and the length of the vector adjusted. When the model vector values are updated, only the component values which correspond to the values available in the input vectors are modified.

b) Missing values can be replaced with the corresponding attribute value of the data records' BMNs.

c) Depending on the operation that is being performed a combination of both method a) and b) can be used.

d) The data records with missing values may be discarded or just removed from the training set and later associated and displayed on the map once it has been trained.

The way in which missing data are handled may depend on the process being performed. For example, during the process of trying to find the BMN, if the sample vector $\boldsymbol{x}_i$ has some missing values, those variables are ignored in the distance calculations. However, during map training a node's new vector values may be computed as a weighted mean for all the data records in its neighbourhood. For this purpose, the missing values in the data records may be substituted by the corresponding attribute value of the data records' BMNs.

However, as long as provision is made for their occurrence, it has been shown that missing values are not normally a problem for SOMs [185]. The SOM's ability to handle missing data was in fact one of the main reasons for selecting the SOM for this investigation.

4.4.4 Outliers

An outlier is a data pattern that differs substantially from the data distribution and hence lies far from the main body of the data. Outliers may occur because of an error in the data set, or an outlier may be data patterns that are really different to the rest of the data. In biological material this can be a fairly common occurrence and might occur due to genetic oddities. In either case, in the SOM displays each outlier affects only one map neuron and its neighbourhood, and if necessary can be discarded and the analysis performed on the rest of the data set [104].

However, sometimes these outliers are important and have qualities which need to be analyzed, in which case they should not be discarded.

4.5 Measures of SOM Quality

There are various ways of measuring the quality of the SOM, and two of these will be discussed next.

4.5.1 Quantization error

The quantization error (E_q) is the sum of distances of each data pattern to the model vector of the winning neuron. The accuracy of mapping can be measured by calculating the E_q. According to Kohonen [114] the best map is expected to yield approximately the smallest average quantization error because it is then 'fitted best' to the data. The E_q evaluates the fitting of the map to the data. The smaller the E_q the smaller the average distance from the input vectors to the model vectors, the closer the data vectors are to their models and the better the fit to the data. The lower limit of the E_q is zero, which is ideal if there is no noise in the data. An E_q of zero means that the neuron weight vector is exactly the same as the data vector, i.e. that each data vector maps exactly to one neuron weight vector. However, if there is noise then an E_q of zero may mean there is overfitting. On the other hand, the quantization error is high if there are unwanted "twists" in the map or if the configuration of the models has not reached a stable state in the learning process [114].

The average E_q (E_{qavg}) is defined as:

$$E_{qavg} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i - \mathbf{m}_c\right\| \qquad \text{Eq. 4-5}$$

where

N is the number of patterns,

$\boldsymbol{x}_i$ is the current vector pattern, and

$\boldsymbol{m}_c$ is the BMN of the corresponding $\boldsymbol{x}_i$ input vector.

Some attributes are more important, and thus statistically more significant, than others with respect to quantization. Hence leaving out, adding or rescaling attributes will affect the quantization error and how well the attributes are represented. The number of times each map node is the BMN (called the number of hits) also has an affect on the E_q. The higher the number of hits on a neuron the greater the total number of errors will be.

The E_q can be used as an indication of map accuracy and as a criterion for stopping the SOM training algorithm, i.e. the training can be stopped when the E_q is low 'enough'. The E_q may also be used as a criterion for selection of the map model,

where several maps of the same size are trained on the same data set and the one with the lowest E_q is selected. When comparing E_qs the maps have to be the same size because the E_q usually decreases as the number of neurons is increased.

The E_q does not take into account topology-preserving properties of the map and a means of assessing neighbourhood relations will be discussed next.

4.5.2 Topographic error

Topology is preserved if data patterns close to each other in the input space are mapped to areas close to each other on the map. The topographic error (E_t) measures the topology preservation and is defined as the proportion of all data vectors for which the first and second BMNs are not adjacent neurons [107]. E_t is a simple SOM-specific error measure that assesses the quality of the vector projection.

$$E_t = \frac{1}{N}\sum_{i=1}^{N} u(\mathbf{x}_i) \qquad \text{Eq. 4-6}$$

where

$u(\mathbf{x}_i) = 1$, if the 1st BMN and 2nd BMN are **not** adjacent, otherwise

$u(\mathbf{x}_i) = 0$

The lower the E_t the better the SOM, i.e. if $u(\mathbf{x}_i) = 1$ then there are similar models in different parts of the map and the mapping is not topology preserving. This happens, for example, when the map folds on itself. If $u(\mathbf{x}_i) = 0$, similar models are close to each other and the mapping is topology preserving. According to Kiviluoto [107] if the dimension of the SOM lattice is lower than the dimension of the input space then $u(\mathbf{x}_i) = 0$ is not possible as, under these circumstances, the topology can never be perfectly preserved.

4.6 Conclusion

This chapter described the SOM and some of its variants. Thereafter the visualization of the SOM was discussed together with vector quantization and vector projection. Next, the most common side effects of using SOMs were reviewed. Finally two methods for assessing the quality of SOM models were discussed.

The next chapter describes the research design used to develop the SOM models applied in this research. Reasons for selecting the application field are motivated and presented, and a description is given of the pre-processing methods applied to the data. The process steps necessary for the development, verification and testing of the models are discussed.

Chapter 5

Developing the SOM Models

The previous chapter discussed the theory and practise of the Self-Organizing Map (SOM). In order to use data for computerized analysis the data must be collected, pre-processed and represented in a format which can be input to and recognized by the selected computer program. This chapter describes how the data for this thesis were collected and treated before they could be used for biological identification using the SOM technique. It then describes how the data sets were presented to the SOM software and the SOM models were developed.

The research design used in this thesis is outlined in Section 5.1. The choice of data and the reasons for selecting these data are explained in Section 5.2. Section 5.3 explains how the data were selected and collected, including a description of any limitations. The choice of the computational algorithm and of the software tool selected is discussed in Section 5.4. A description of how the data were pre-processed and represented is given in Section 5.5. The storage of the data and any special requirements that were applied to the data before they were presented to the software tool are described in Section 5.6. A description of how the data were divided is given in Section 5.7. The process of presenting the training data set to the ANN and the formation of the SOM models is given in Section 5.7.1 while Section 5.7.2 describes the presenting of the test data set to the SOM models. Section 5.8 concludes the chapter.

5.1 Research Design Outline

The objective of this thesis is to assess the effectiveness of the SOM for biological identification with a view to improving biological identification methods. The steps taken to achieve this objective were:

- Choosing the biological application field - a discussion of what biological field was chosen and the reasons for this choice are given.

- Sampling the population of the selected biological domain.
- Indentifying the software tool to be used.
- Coding the data samples.
- Presenting these data to the software tool.
- Interpreting the software results.
- Assessing the effectiveness of the SOM models.
- Recommendations arising from using an ANN approach for biological identification.

How these steps were performed and the results obtained will be discussed in this chapter and in Chapter 6.

The choice of data is described in the following section together with the reasons for the choice.

5.2 Choice of Data

Careful consideration was given before selecting a botanical rather than zoological group. In particular, many groups of commonly encountered trees with complex evolutionary and taxonomic relationships seemed appropriate as a means of testing the ability and efficiency of the SOM algorithm as an identification tool. The choice was driven by the fact that there is a very real need to provide non-specialists with an aid to identify trees when an expert is not readily available. This is especially true in Africa where there are very few taxonomic experts while the biodiversity of the continent is extremely large. Flowering plants alone are estimated to be in the order of 70000 species. In addition, laymen often find the identification of trees difficult. There are several reasons for this: for instance, vegetative characteristics (for example leaves), although present most of the time, show variations which are not obvious to the inexperienced. To make things even more difficult, some of the most obvious macroscopic characteristics (for example flowers and fruit) although extremely useful and diagnostic in the identification process, are oftentimes not present. Although each tree species does have its own unique characteristics, these characteristics often do not

display easily identifiable macroscopic variations within its group that other biological species do. For example, most trees have green leaves and greyish/brownish/blackish bark with little variation in colour, while, on the other hand, other biological species, for instance birds, have different colours and sometimes different shapes and sizes. This variation aids enormously in differentiating the species during the identification process.

Having selected trees as the source for the data sets to test the SOM, the next step was to find a small group or genus of trees. The training and test data sets have to be limited in size so that the amount of data does not become overwhelming, but the group should still remain scientifically suitable for testing the identification tool. Consequently, in choosing a model system it is necessary to look for a relatively small genus which is taxonomically well worked out but also has some challenges. Ideally the chosen species should be identifiable by an experienced expert but should present difficulties for the layman who does not have experience with the particular taxa. To limit the size of the data set it was thought that the group chosen should be limited to a region rather than the whole of South Africa, southern Africa or the world. As the author is resident in KwaZulu-Natal (KZN) the choice of taxa was limited to this province of South Africa.

The genus *Acacia* has 23 taxa which are indigenous to KZN, and this is considered a reasonable number for a study at this level and complexity. Also the genus is found fairly commonly in KZN, thus making the gathering of data less difficult. The *Acacia* species are not easy for non-specialists to differentiate, especially as the vegetative characteristics show variation which is not obviously diagnostic to inexperienced botanists. Often the flowers and pods are not present and these are usually important macroscopic characteristics used for differentiating *Acacia* species.

Acacia is an important and widespread African genus which often dominates the landscape. In fact *Acacia* species are extremely important locally and all species make some contribution to the environment and rural economy by way of shade, shelter, soil stabilization and fertility, food for browsing animals and fuel; to mention a few uses. Many are also important ingredients in local herbal medicines called *muthi*. In

view of the above, the *Acacia* trees indigenous to KZN were selected for testing the SOM algorithm's ability to identify tree species.

In the next section the recent plans to change the name *Acacia* are discussed and the meaning of the name *Acacia* is explained.

5.2.1 Moves to Transfer African *Acacia* Species to a New Genus

The name *Acacia* is the Latin form for the Greek name *acanth/acantho* meaning thorny or spiny [211] or alternatively the Greek word *akakia* (which is derived from *ακις* (*akis*) which is Greek for a sharp point). Although irrelevant to this thesis it is interesting to note that the *Acacia* wood is reputed to be the wood used to build the Arc of the Covenant [159].

Since this thesis was started a recommendation has been made for the name *Acacia* to be conserved for the Australian species and that the African (and other) species be placed in newly proposed genera [26]. However, counter proposals are to be put forward for adoption at the 2011 Botanical Conference in Melbourne. As a result this thesis will continue to use *Acacia* for the African species until this nomenclatural problem is resolved [9]. A number of papers [137, 142, 146-148, 158] deal with the *Acacia* name change issue.

The acquisition of the data set used in this thesis is discussed next.

5.3 Data Collection

Having selected the *Acacia* species of KZN to trained and test the SOM the next step to be performed is to collect and prepare the data for use. These steps will be discussed next and an outline of the steps is depicted in Figure 5-1.

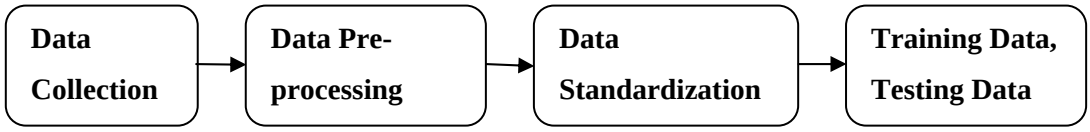


Figure 5-1 : Process Diagram of Steps in Preparing Data for SOM
(adapted from Vesanto [234, page 3])

A process of acquisition and elicitation of the data is needed for inputting into any analytical computer software and to a large extent the final function and success of such an analytical system is dependent on 'what' data are presented to it and 'how' these data are presented. Traditionally, botanical data are stored in a number of forms, but by far the most extensive source of morphological information is in the form of written descriptions. When gathering the data for computerized analysis, the first problem that has to be dealt with is the manner in which these data are structured. Morphological, anatomical and cytological botanical data are such that they cannot be presented directly to a computer system because they are often incomplete, imprecise, unstructured and dispersed [66]. In this thesis the data for the data sets were drawn from several sources. This included extracting information from authoritative botanical literature (keys and scientific descriptions) and herbarium specimens' documents, undertaking field work, and consulting experts. In particular, the data were directly derived in a manner similar to that employed by botanical experts: namely from dichotomous keys of the 'if-then' type, monograph descriptions, species description composed by experts, and herbarium specimens. These are the sources of data that are used for the classification and recognition of plant species. In particular, data sources used were the Flora written by Ross [178-180], with additional data being gathered from [28, 42, 53, 145, 168, 191, 207, 218, 229, 230]. In addition, data were obtained from observing *Acacia* species in nature, and from information gained directly from knowledgeable colleagues. This information was subsequently confirmed by consulting herbarium documents and specimens, and by taxonomic experts.

Thus, information on acacias was gathered and summarized and a list of macroscopic diagnostic attributes (descriptors or characteristics) at species level was compiled. The choice of descriptors was guided by the attributes which were presented in the consulted literature. Only macroscopic characteristics that are easily and usually observed with the naked eye, and that are normally used in the field for identification, were considered. Once a list of the most important taxonomic characteristics had been drawn up, the attributes were grouped as shown in Table 5-1.

Table 5-1 : Sub-groups of *Acacia* Characteristics

Sub-groups	Maximum No. of Characteristics
Habit & Thorns	43
Flowers	19
Pods and Seeds	25
Leaves	40
Total	**127**

The following points should be noted in conjunction with Table 5-1:

- Habit and thorn characteristics of the trees included attributes which described the tree in general (viz. the habit), for example the maximum height of the tree, colour of the trunk; and thorn qualities such as the type of thorn (straight or recurved), and the length of thorn.
- Leaf characteristics included attributes such as the length of the leaf and the number of leaflets.
- Flower characteristics included qualities such as colour and shape of inflorescences.
- Pod and seed characteristics included attributes such as shape of the pod and number of seeds.

Altogether 127 macroscopic characteristics were extracted to describe each of the *Acacia* species used in this study.

The identifying diagnostic attributes and their values were stored in a MS-Excel spreadsheet. An example of a portion of this table is shown Table 5-2. The database of information is seen as a matrix, where each row represents an instance and each column represents an attribute. In botanical science the genus name (in this case *Acacia*) can be abbreviated to its initial (in this case *A.*) when the genus name is followed by the species name. Thus, *Acacia ataxacantha* can be abbreviated to *A. ataxacantha*. This convention will be used in this thesis. In Table 5-2, as confirmed by a taxonomic expert, *A. ataxacantha* has an elongate or spike-like inflorescence with white flowers, while *A. karroo* has a capitate or head-like inflorescence with yellow

flowers. This information is indicated by the value 'Yes' in the appropriate columns in Table 5-2. For each specimen the corresponding attribute value (if available or applicable) was noted, and finally a table of values was obtained for 80 samples for each of the 23 KZN *Acacia* species. Altogether a total of 1840 specimen descriptions was obtained.

Table 5-2 : Extract from a Numerical Table Describing *Acacia* Specimens

	…	SPK[1]	HD[2]	COLY[3]	COLWH[4]	…
A. ataxacantha		**Yes**			**Yes**	
…						…
A. karroo			**Yes**	**Yes**		

where the codes:

1 **'SPK'** means spike;

2 **'HD'** means head;

3 **'COLY'** means colour yellow;

4 **'COLWH'** means colour white.

According to Lisboa [136] the sample size of the data set should be at least five times more than the number of attributes; and if the sample size is small, cross-validation should be used. Vesanto also states that for SOMs the number of samples must be considerably more than the number of attributes [234].

In this research project, half the herbarium samples were kept for training and half were kept for testing. As a result, for the whole data set the training sample size (920) is over seven times the number of attributes (127). In addition, a 30-fold cross-validation was employed.

In the next section the choice of software is discussed.

5.4 Choice of Software

According to Cottrell [45] the SOM, and its related extensions, is the most popular artificial neural algorithm for use in unsupervised learning, clustering, classification and data visualization. Kohonen [116] states that over 7000 scientific articles have been written about SOM [105, 157], and in addition many commercial projects

employ the SOM as a tool for solving hard real-world problems [45]. SOM properties include quantization, clustering and visualization, which are all useful for identifying data. The SOM technique, in particular, can still provide results for data even when many values of the characteristics are missing, thus making the method highly appropriate for the problem of botanical identification. For these reasons, and for reasons already discussed in previous chapters, the SOM algorithm was selected as an ANN technique to solve the problem of botanical identification.

Having decided to use the SOM algorithm, available SOM tools were investigated. The Helsinki University of Technology (HUT) web page [1] lists available SOM software and gives a short evaluation of each package. SOM_PAK is public domain software [112] which was developed by the SOM Programming Team of HUT and may be considered as the original SOM implementation. Deboeck [56] and Kohonen [115] also give an overview of available SOM software tools. Even though SOM_PAK appeared to be a logical choice, at the time of selecting the software, Viscovery® SOMine was the state-of-the-art SOM software, and so this package was chosen for use in this research.

The Viscovery® SOMine software is simple to operate and is reliable. The pre-processing of input data allows for scaling, priority settings and transformations. Output visualization can be in the form of cluster maps, component planes, U-matrix, iso-contours and limited statistics. Several clustering options are available and the SOM algorithm is combined with the Ward clustering method. The software can perform dependency analysis, however the final analysis/interpretation of the output is performed by the user and not by the package. The software uses the batch map algorithm and provides some accelerated computing (in the form of a growing map) which has the effect of being able to achieve high computing speed. The map array is always hexagonal, and the initialization of the model vectors is made along the plane spanned by the principal axes. The neighbourhood function is always Gaussian and missing data are handled automatically. These features combine to offer a very effective software tool.

The version of Viscovery® [2] used during this thesis is Viscovery® SOMine Plus Version 4.0 which allows up to 50,000 data records with no restriction on the number

of variables (characteristics). The Viscovery® software allows certain parameters to be set before the data are presented to the ANN. For instance, the software allows the user to select the number of nodes. When the user selects the number of nodes the software uses approximately the number requested (not necessarily exactly). The default value for the number of nodes used for SOMine is set to 2000 nodes.

With Viscovery® software the significance of a characteristic can be scaled by increasing the priority of that characteristic. If a priority of less than one is selected, the importance of that characteristic is decreased. If a priority of more than one is selected, the importance of that characteristic is increased. The default priority is 1 and this was the value used for the work in this research.

The value of the map tension may also be selected. The default value is 0.5, and this value was used as a larger tension (say above 1) would result in a rigid map.

The Viscovery® software default values were used in this research as the aim of this thesis was to see if the SOM is an effective tool for identifying biological material. More particularly, the inter-relationships of the data were important and so, as far as possible, default values were used to minimize the effect of manipulation and thereby possibly biasing the outcomes.

The next section presents a discussion on how the data were prepared for inputting into the ANN.

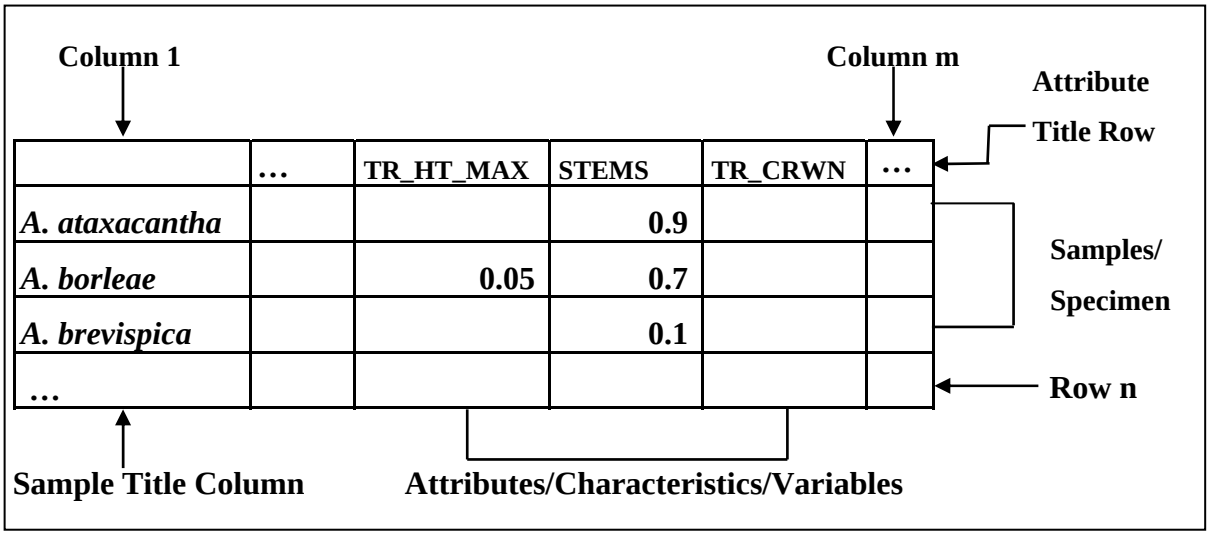
5.5 Data Pre-processing

There are two assumptions made about the data collected: firstly, the identification of all instances is unknown to the SOM, and secondly, the data are numerical and are either continuous or discrete. No formulas, calculations, macros or text are allowed within the data set except for text descriptions in the title row and title column. Consequently, in the data table the first row is a title row that labels each column of attribute values; and the first column is a title column that labels each specimen sample. The title row and title column are ignored during subsequent creation of the SOM.

Although the identity of the instances of the species is assumed unknown and not presented to the SOM, the class labels for all the instances have been determined by experts and this information is used afterwards for labelling and to aid in the interpretation of the results. The reason for not presenting species' identities to the ANN initially is so that the output of the SOM is data driven and not affected by imposing human decisions which could be biased. In this way the identities of the specimens do not affect the structures that may be found, and any pattern presented by the SOM is determined objectively and *a priori*.

Data, as extracted from botanical sources, are usually in the form of written descriptions. These textual descriptions have to be represented numerically before presentation to an ANN. This transformation was done manually and the results were stored in a table, as shown in the extract presented in Table 5-3.

Table 5-3 : Extract from Numerical Table Describing *Acacia* Species



In Table 5-3 column one, the names of the specimens are given, while in row one the names of the attributes are given. The intersection of each row [2..n] and of each column [2..m] gives the attribute value for the respective specimen. For example, in Table 5-3, *A. borleae* has a maximum tree height (TR_HT_MAX) which is represented by 0.05, and a stem (STEMS) value of 0.7 (which means the tree may have one or more stems from the base). No values are available for describing the tree

crown (TR_CRWN) so this cell in the table is left blank. By employing a numerical table-format the data can be investigated from the viewpoints of either the specimens or their attributes. When the table is investigated as a collection of specimens interesting similarities between individual specimens can be considered. When the table is investigated as a set of specimen attributes the statistical properties and dependencies of the attributes can be considered [234].

Some characteristic values were encoded with discrete values of 0 or 1 (corresponding to false or true, or present or absent) as appropriate. This encoding was used for characteristics such as colour of flowers where species with white flowers could for example have their characteristic value as 1 (true for white flowers) while species with yellow flowers could have this characteristic value stored as 0 (false for white flowers).

Other characteristic values were encoded using a real number between 0 and 1. This encoding was used for characteristics such as length of thorn. The normalization of the data is discussed in Section 5.6.

During the pre-processing stage, the original raw data were cleaned and transformed so that:

- the significant data properties were presented more clearly,
- there are fewer or no erroneous values, and
- the data are in a numerical format suitable for the SOM method.

During the next phase data were standardized, and this is discussed in the next section.

5.6 Data Standardization and Storage

Data were normalized between (0, 1) to focus attention on the pattern of the data rather than on absolute levels of the data values. Some characteristic values, such as maximum tree height, have high values while other characteristic values are much lower, such as leaflet length. In order to standardize the effect of the contributions of the different variables all data were (0,1)-normalized as was suggested by Kohonen [114].

Once the data set has been prepared it must be stored in a format that is compatible with the Viscovery® SOM software. The SOM software will accept data stored in a MS-Excel spreadsheet or in a text file. For this research Microsoft® Office Excel 2003 was used for storing the data.

The data set was compiled independently of the Viscovery® SOMine software used, and the data were presented to the SOM software only when the training and testing was done. The SOMine tool can work with any set of numerical data prepared according to the necessary criteria, and so the ANN could be used with any biological material.

The next section gives a description of the process followed when presenting the data to the SOM algorithm.

5.7 Data Utilization

Kohonen [114] reported on a technique which required that the data set be divided. This technique allows some of the data to be used for training the SOM while the remaining data are used for testing. The patterns used for training have a relatively small amount of data missing while the test data set has a higher proportion of attribute values missing. This test data set was mapped on to the trained SOM. A technique similar to this was employed for this thesis.

For this research project the data set was split to form a training set and a test set. The training set was divided further to provide subsets for cross-validation purposes, and the next three subsections describe how these sets were used. In addition, the training set was divided into subsets based on four major biological characteristics (i.e. habit and thorns; flowers; pods and seeds; and leaves), and the entire experimental procedures were repeated for each subset.

5.7.1 Training Data Set

For training the ANN data must be as complete as possible (i.e. little missing data). The training set is meant to represent an analysis of the complete knowledge of *Acacia* as known by taxonomic experts of this genus. These experts need to have a

very good overall knowledge of the attributes of each of the *Acacia* species. For this reason the training data set was made as complete as possible, and when values were missing (because they were not observed or were absent) the gaps were filled in wherever possible by referring to accredited expert sources as has been done in other biological research [31, 43, 66, 78].

In this thesis, with reference to a model or a data set, the word 'whole' has been used to refer to the data set containing all the attributes of the samples. For example, the term 'whole SOM model' is used to describe the SOM models formed using the whole data set, and the term 'whole data cross-validation sets' is used to describe the cross-validation sets consisting of data containing the whole set of attributes.

The whole data set represents 23 *Acacia* species and consists of 920 attribute patterns. These were evenly distributed among the species giving 40 patterns (or 4.7% of the training data set) for each of the species. Each pattern consists of up to 127 attributes with a small number of values missing where a particular attribute value was unknown or inappropriate for the pattern in question.

All experiments used a 30-fold cross validation process where the training set was randomly divided into thirty disjoint sets of which 29 subsets contained 30 patterns each and one subset contained the remaining 50 patterns. This gave a queue of subsets. The division of the data set is shown in Table 5-4.

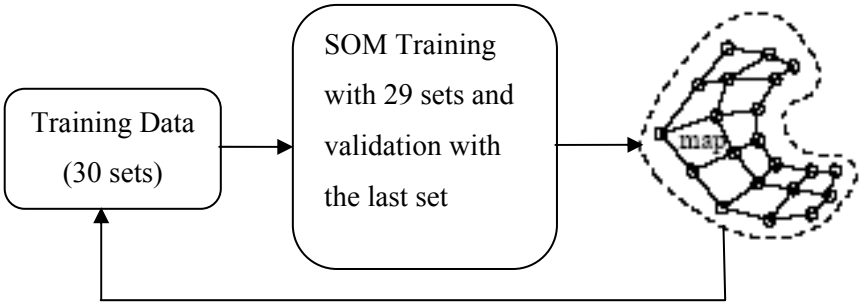
Table 5-4 : Composition of the Whole Data Cross-Validation Sets

Whole Training Set Number	Size of Training Set	Size of Verification Set
1	870	50
2 – 30	890	30

For each of 30 simulations the first 29 subsets were used as the training set and the last one as the test validation set (to measure generalization performance of the maps). After each simulation the last subset was removed from the end of the queue and reinserted at the front of the queue, and the network was trained again. Thus the verification set consisted of a unique set for each simulation (i.e. each pattern gets used exactly once for verification) but the training sets were not unique (i.e. each

pattern gets used 29 times for training). After these steps were repeated for a total of 30 simulations, 30 training-validation set pairs were obtained.

Each of the 30 training sets was presented to the ANN as depicted graphically in Figure 5-2.



(Repeated 30 times, each time after rotating the 30th set to the front of queue)

Figure 5-2 : Steps in Performing SOM Training
(adapted from Vesanto [234, page 3])

In this research the whole set of training data was presented to the software and the number of nodes selected varied from 23 (the number of *Acacia* species used in the data set) upwards. It was found that the best results were obtained by using 201 nodes, where best results were judged to be the ability to cluster the specimens accurately into 23 groups using the smallest number of nodes.

So as not to influence the output of the algorithm unduly, this research used the default parameter values (except for the number of nodes). This meant that all components (characteristics or attributes) were used without restricting, amplifying or suppressing the range of values, or removing any records.

After the network had been trained the output of each simulation was saved. Each simulation outputs a cluster map and value maps for each of the characteristics. These 30 cluster maps are the models that were used for performing the testing phase and are discussed in Chapter 6, subsection 6.1.1.

The next section describes the process of presenting the verification data to the trained map.

5.7.2 Verification Data Sets

Each of the 30 verification data sets was presented in turn to its respective trained map using Viscovery®'s recall function. This verification process was performed to demonstrate that the results were not obtained randomly and could be obtained repeatedly. The 30 verification sets were obtained as described in Section 5.7.1, and the results of the cross-validation process are discussed in Chapter 6, Section 6.1.2.

The next section discusses the test data set.

5.7.3 Test Data Set

Each of the 30 SOM models, obtained as described in subsection 5.7.1, was used to see if the model could identify an *Acacia* test data set accurately. The test data set consisted of 920 patterns consisting of up to 127 attributes. These patterns were previously unseen by the trained network and were sparsely populated, i.e. had many missing attribute values. When trees are observed in nature, many characteristics are absent or may not be observed at that particular time. For example, flowers and pods are often absent during particular seasons. Therefore, no attempt was made to fill in missing values in the test data set. Use of a sparse test data set is important to this investigation as it is essential to demonstrate that the trained ANN can identify data sets with many missing values. Table 5-5 describes the process followed for testing the SOM models.

Table 5-5 : Testing the SOM Model

SOM Models	Test Set	Function Utilized	Results
1 - 30	920 randomly sorted patterns	Viscovery® Recall Cluster Membership	See Section 6.1.3

The randomly shuffled test data set was presented to each of the 30 trained SOMs using the recall function as depicted in Figure 5-3. This process was repeated 30 times, each time using a different trained map. These results will be presented and discussed in chapter 6.

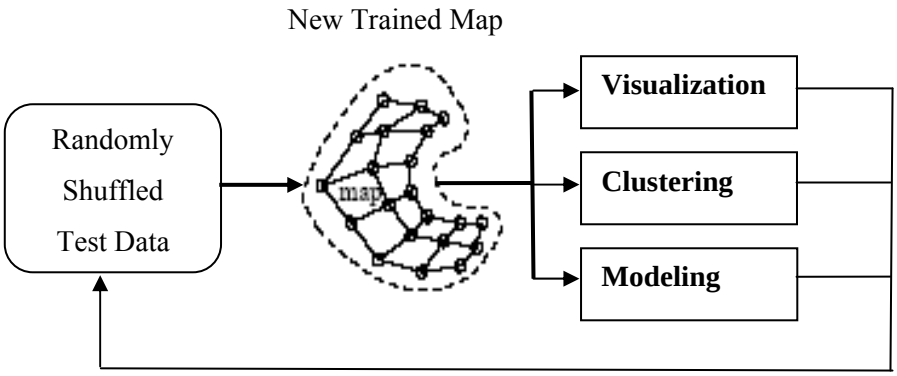


Figure 5-3 : Steps in Modelling Test Data
(adapted from Vesanto [234, page 3])

5.7.4 Data Sub-Groups

In addition to using the entire training data set as a complete entity having 127 attributes, the training data set was also split into each of the sub-groups shown in Table 5-6. This was done in order to test if a trained sub-group map was able to identify the corresponding sub-group test set. Each of these sub-groups was randomly shuffled and treated in the same way as the whole data set had been treated. The 30 training-validation data set pairs were obtained for each of the sub-groups and the network was trained with each of these training data sets. Thereafter the verification process was performed on each of the training data sets using the respective validation set.

Table 5-6 : Data Sub-Groups

Data Sub-Groups	No. of Attributes
Habit & Thorns	43
Flowers	19
Pods and Seeds	25
Leaves	40
Total	**127**

The next step involves checking and analyzing the maps and results obtained from training and testing the network. This analysis will be covered in Chapter 6. The

following section briefly summarizes the work covered in this chapter and introduces Chapter 6.

5.8 Conclusion

This chapter aimed to provide a description of the choice of software and application data used in this thesis. It discussed the process followed in choosing the biological application field and the reasons for this choice. Next, the process of how the sampling of the biological domain (in this case the acacias of KZN) was performed is described. The choice of Viscovery® SOMine software is discussed and the coding necessary for pre-processing the data is described. Finally, the presentation of these data to the SOM algorithm is described together with the training verification and testing of the SOMs.

In order to assess the SOM models as tools for biological identification, the effectiveness of the maps produced by the training data sets are discussed and analyzed in Chapter 6. Validation set results are investigated to verify that the maps are able to produce consistent and accurate results. Thereafter, in order to demonstrate the SOM's ability to identify *Acacia* trees efficiently and accurately, the results obtained from presenting the test data set to the trained maps are analyzed and discussed.

Chapter 6
Analysis of the SOM Performance

The basis of empirical science is that every problem or scientific question is investigated via the experimental collection of data. These data are then analyzed for patterns of certainty and are then interpreted in the context of the already extant body of knowledge on the subject. Good science is hypotheses driven, and these hypotheses must have deductive or predictive power. The main task of a SOM, if used in this process, would be to act as an exploration tool for acquiring and understanding the properties of these data, and for generating hypotheses about these properties.

Chapter 5 outlined the research design process that was performed for this thesis. The choice of the application field and the application software was discussed and reasons for their selection were given. The sampling process was described and the pre-possessing and the encoding of the data were explained. Finally, the presentation of data to the software and the verification and testing processes were described.

This chapter presents the results of the experiments outlined in Chapter 5 and analyzes them to assess the effectiveness of the models for identifying specimens of KZN *Acacia* species. These models consist of five sets:

1. **TreeSOM** models which were developed from the whole data set,
2. Habit and **ThornSOM** models which were developed from the habit and thorn data sets,
3. **FlowerSOM** models which were developed from the flower data set,
4. Seed and **PodSOM** models which were developed from the seed and pod data sets, and
5. **LeafSOM** models which were developed from the leaf data set.

These models are referred to as the TreeSOM, ThornSOM, FlowerSOM, PodSOM and LeafSOM models respectively.

Section 6 introduces the TreeSOM models, and the results of the training of the TreeSOM models are discussed in Section 6.1.1 followed by a discussion of the

results of the verification process in Section 6.1.2. Testing and analysis of results of the TreeSOM models are described in Sections 6.1.3 and 6.1.4 respectively. The Habit and ThornSOM models are presented in Section 6.2 and are discussed using the same structure as outlined above for the TreeSOM models. Similarly, the results of each of the remaining models are then described, but in less detail so as to avoid repetition. However, more details are given where the results showed different or interesting aspects worth noting. The FlowerSOM models are covered in Section 6.3, the PodSOM models in Section 6.4 and the LeafSOM models in Section 6.5. The C5 and CN2 results are discussed in Section 6.6, and the chapter is concluded in Section 6.7.

6.1 The TreeSOM Models

The TreeSOM models were created from the whole data sets as described in Chapter 5. The performance of these models is now investigated and analyzed by evaluating the results of each of the steps performed (i.e. the training, verification, and testing phases). These steps are discussed and evaluated in turn.

6.1.1 Evaluation of the TreeSOM Models

The whole data set contained 920 specimen patterns, 40 for each of the 23 KZN *Acacia* species. The patterns had up to 127 attributes each and were used for training and validating the neural network. Training the network using each of the 30 whole training data sets was performed as described in Chapter 5, subsection 5.7.1. The number of neurons used for training the network was 201. This number was determined by requesting the software to use 23 nodes (the number of KZN *Acacia* species comprising the data set) for training and then increasing the number of nodes used until the network could cluster the different species accurately into 23 separate ‘classes’. These 23 classes equate to the 23 species and subtaxa of KZN *Acacia* recognized by taxonomists using conventional classification rules.

The SOM cluster map obtained from the first simulation of the whole training data set is presented in Figure 6-1. The cluster maps obtained from the other 29 simulations were similar. This figure shows that 870 randomly shuffled specimens

have been assigned to their own unique clusters. The map confirms that the SOM was able to learn from the input training data and differentiated between the 23 species. The TreeSOM was able to cluster each of the specimens uniquely, i.e. to a cluster which contained only specimens from a single *Acacia* species, and thus there were no misidentifications.

After training, each of the clusters was manually labelled with the names of the species that had been mapped to that cluster of nodes. This is also illustrated in Figure 6-1. It is notable that the species that are mapped close to each other in the TreeSOM have similar characteristics in nature.

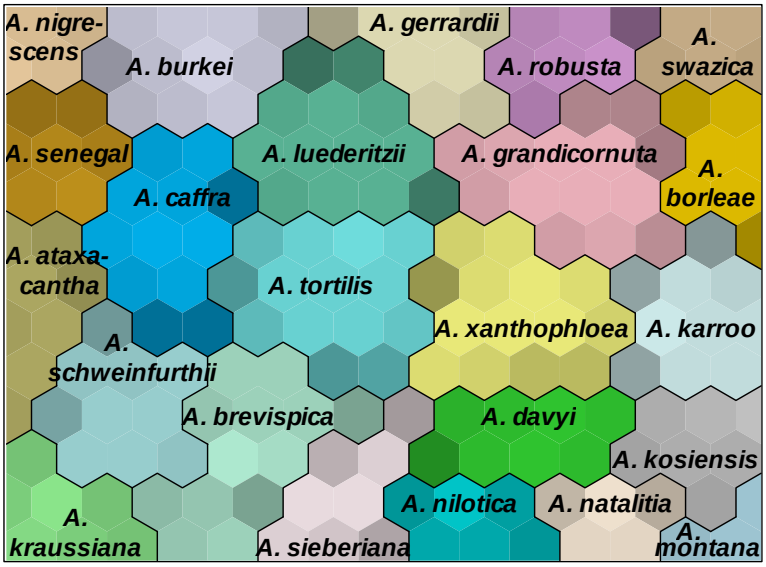
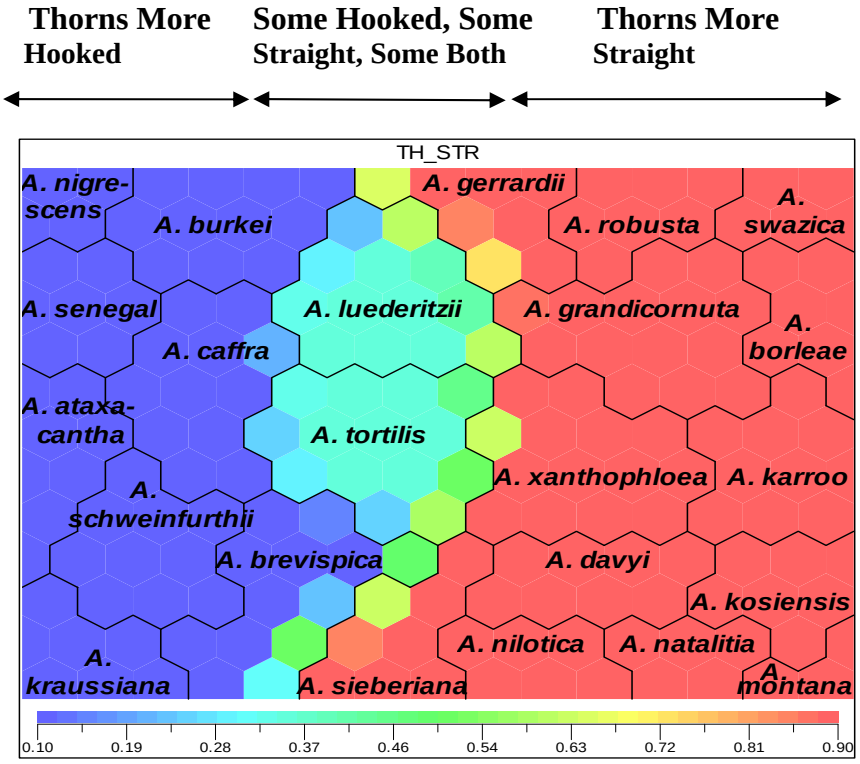


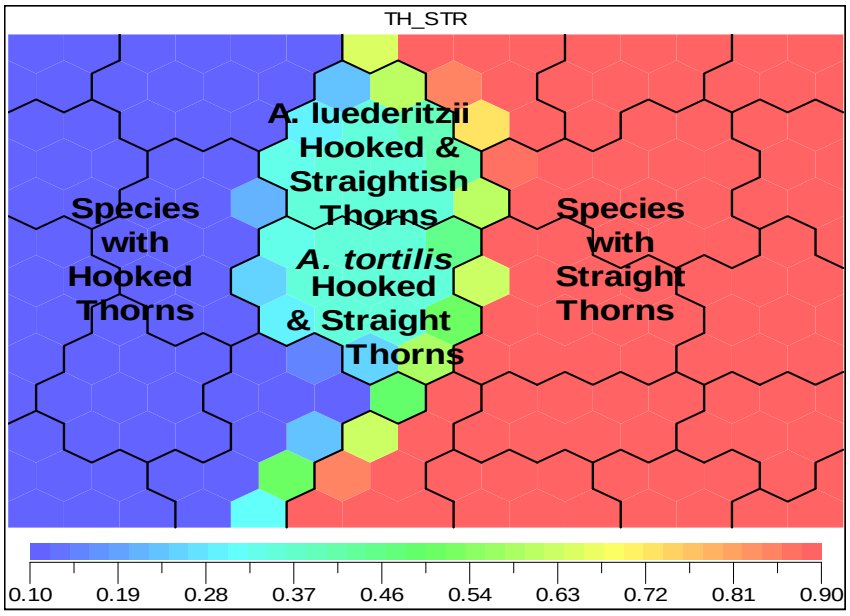
Figure 6-1 : TreeSOM Model 1

In addition to producing a cluster map, the SOM software produces a component plane map for each attribute (or morphological character) in the data set that was used in the development of the map. Each component plane of the SOM consists of the values of the same component in each model vector. The component maps are visualized by giving each neuron a colour according to the relative value of the respective component in that neuron. A blue colour coding depicts low values for an attribute, while a red colour coding depicts high values. This variation is shown in the bar chart at the bottom of each component map. By comparing the component maps with each other the correlations between variables can be seen.

A component map for the single attribute TH_STR (i.e. straight thorns) is shown in Figure 6-2, and for ease of identification each of the clusters is outlined with black lines.



(a)



(b)

Figure 6-2 : Straight Thorn Component Map for 23 *Acacia* species

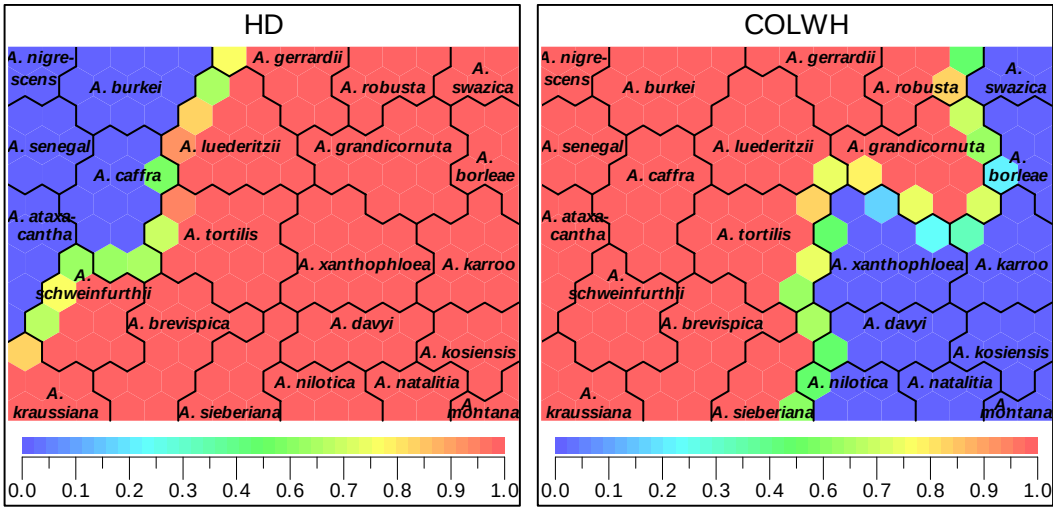
The left side of the map in Figure 6-2 (a) is blue depicting that those clusters contain patterns with low values for straight thorns (this is to be expected as the blue-coded clusters correspond to species that have hooked thorns and not straight thorns). On the right side of Figure 6-2 (a), the red colour coding depicts the clusters which have high values for straight thorns (again this is to be expected as the red-coded clusters correspond to species that have straight thorns). The green and paler blue coding coincides with clusters which have medium values for straight thorns (these clusters correspond to species which can have both hooked and straight or 'straightish' thorns). The middle clusters correspond to the specimens belonging to *A. tortilis*, which has straight and hooked thorns, or to *A. luederitzii*, which has hooked thorns with some thorns thick and 'straightish'. The position of these two species on the map is thus located between the species with low values for straight thorns (hooked thorn species) and the species with high values for straight thorn (straight thorn species).

The map displayed in Figure 6-2 (b) is essentially the same as Figure 6-2 (a) but the species labels have been removed and replaced by labels indicating the thorn types of the different sections of the map.

In Figure 6-3 the component map for the component map HD (round flower head) is displayed alongside the component map COLWH (white coloured flowers). Both maps have the clusters outlined with black lines. In the HD component map of Figure 6-3 (a), the *Acacia* species which are blue coded are clustered in the upper left corner and represent the species that have low values for capitate influorescence (HD), i.e. these species all have spicate influorescence (elongated flower). On the right side and lower portion of Figure 6-3 (a) the red coded clusters represent the species with high values for capitate influorescence, i.e. these species all have captitate influorescence.

From the COLWH component map shown in Figure 6-3 (b) it can be seen that the *Acacia* species with spicate inflorescences (top left of Figure 6-3 (a)) all have white flowers: i.e. in the left section of map Figure 6-3 (b) all specimens have high values (red colour code) for white flowers. Looking at Figure 6-2 (b) in conjunction with Figure 6-3 (a) and (b) it can be seen that the species with white flowers (top left of Figure 6-3 (b)) and spicate inflorescences (top left of Figure 6-3 (a)) also all have

hooked thorns (left side of Figure 6-2 (b)). Similarly, from these figures in can be seen that all species with capitate inflorescences (right side of Figure 6-3(a)) and yellow flowers (blue colour code on the right side of Figure 6-3(b)) have straight thorns (right side of Figure 6-2 (a)).



(a) **(b)**

Figure 6-3 : Component Maps of Species with Capitate, White Flowers

From the above observations it is apparent that the SOM is identifying species' similarities and differences as was hoped and expected.

Figure 6-4 depicts the species which have well-defined leaf cushions (*A. gerrardii* and *A. robusta*), and species which sometimes have cushions or which have poorly defined cushions. From the figure it can be seen that species with cushions (with the exception of *A. senegal*) are species which have straight thorns and are therefore located on the right side of the map. This map also clearly demonstrates the location of the species *A. gerrardii* and *A. robusta* next to each other. Botanically these species are recorded as bearing a strong resemblance to each other [180]. Similarly *A. grandicornuta* and *A. robusta* are reported to be very similar species and, according to Ross [180, p129], a case could be made to place *A. grandicornuta* under *A. robusta*. However, the former is still a very distinct taxon and Ross suggests it would be better to keep these species separate. The map in Figure 6-4 shows the closeness of

these two species to each other and to *A. gerrardii* as they are all located near to each other in the top right centre of the map.

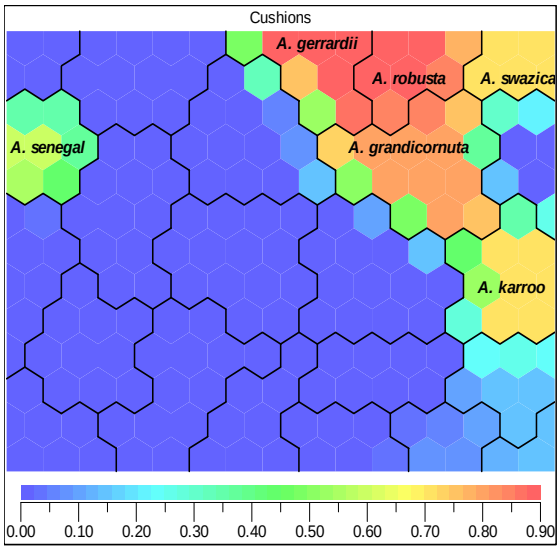


Figure 6-4 : Component Map of Species with Leaf Cushions

Another interesting exercise is to examine how the TreeSOM groups the species when the number of clusters is reduced. This is illustrated in Figure 6-5 which presents a dendrogram of how the *Acacia* species of KZN clustered when the number of clusters of the TreeSOM model was varied from 2 to 23. The height of the horizontal bars in Figure 6-5 indicates the order in which the species were split. Alternatively, Table 6-1 lists how TreeSOM splits the species' groups as the number of clusters is increased. Colours are used to emphasize the splitting of the species into subgroups, and shades of similar colours are used to emphasize species which are known to have close relationships.

Table 6-2 gives a list of some known similarities which are all demonstrated in the dendrogram in Figure 6-5. From this table it can be seen that many of the relationships demonstrated in the TreeSOM model are verified from relationships already established in the literature based on the analysis of morphology. The close locations of some species to each other are demonstrated in the TreeSOM model in Figure 6-1. For example, the species *A. gerrardii*, *A. grandicornuta* and *A. robusta* are closely related biologically, and TreeSOM verifies this by clustering the species

together at the top, central right of Figure 6-1. These species remain grouped together (group 6 on the dendrogram in Figure 6-5) and are only distinguished from each other in the latter stages of the division of the clusters.

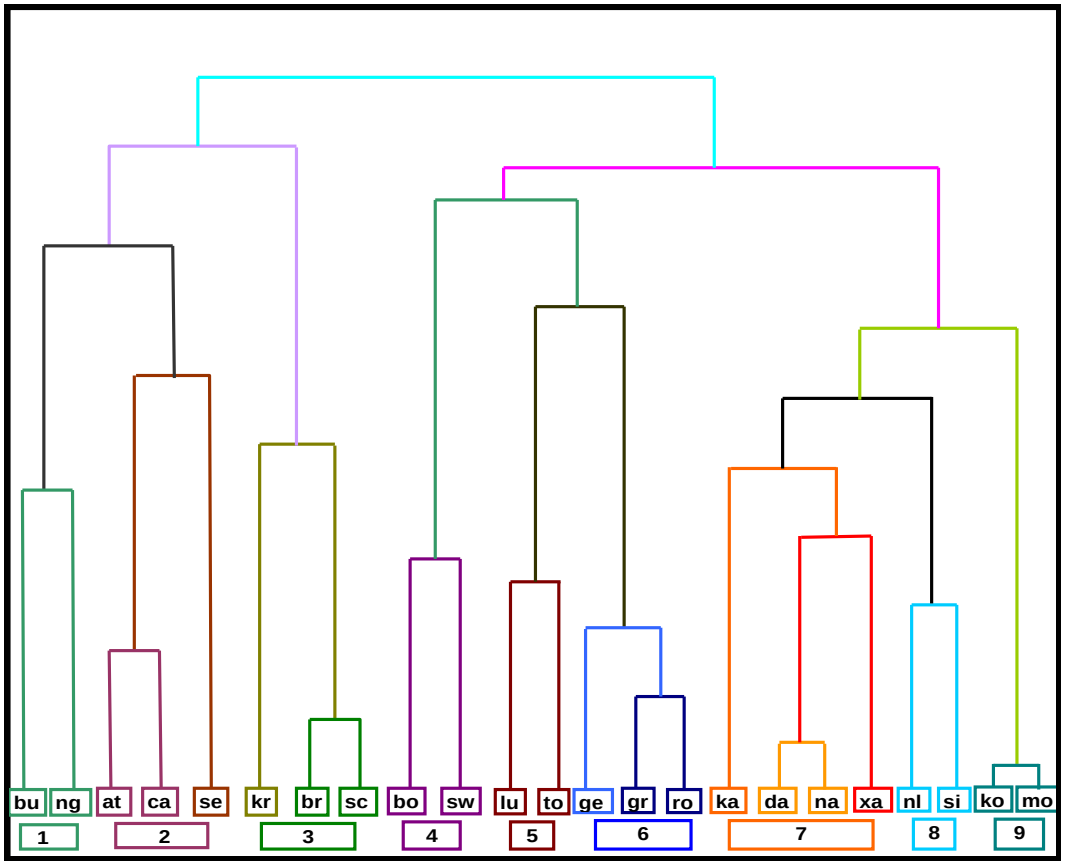


Figure 6-5 : Dendrogram of TreeSOM's Clustering of 23 KZN *Acacia* Species

(**NB:** The groups are numbered for reference purposes.)

Key:

at = *A. ataxacantha*	**ka** = *A. karroo*	**ro** = *A. robusta*
bo = *A. borleae*	**ko** = *A. kosiensis*	**sc** = *A. schweinfurthii*
br = *A. brevispica*	**kr** = *A. kraussiana*	**se** = *A. senegal*
bu = *A. burkei*	**lu** = *A. luederitzii*	**si** = *A. sieberiana*
ca = *A. caffra*	**mo** = *A. montana*	**sw** = *A. swazica*
da = *A. davyi*	**na** = *A. natalitia*	**to** = *A. tortilis*
ge = *A. gerrardii*	**ng** = *A. nigrescens*	**xa** = *A. xanthophloea*
gr = *A. grandicornuta*	**nl** = *A. nilotica*	

The abbreviations used in Figure 6-5 have been used in place of species names in some tables and figures in this chapter.

Table 6-1 : The Order in which TreeSOM Splits the Species

Number of Clusters	Division of Species (as Number of Clusters Increases) ← Group A → | : | ← Group B →		
1	(at,bo,br,bu,ca,da,ge,gr,ka,ko,kr,lu,mo,na,ng,nl,ro,sc,se,si,sw,to,xa)		
2	(at,br,bu,ca,kr,ng,sc,se)	:	(bo,da,ge,gr,ka,ko,lu,mo,na,nl,ro,si,sw,to,xa)
3	(at,bu,ca,ng,se)	:	(br,kr,sc)
4	(bo,ge,gr,lu,ro,sw,to)	:	(da,ka,ko,mo,na,nl,si,xa)
5	(bo,sw)	:	(ge,gr,lu,ro,to)
6	(bu,ng)	:	(at,ca,se)
7	(lu,to)	:	(ge,gr,ro)
8	(mo,ko)	:	(da,ka,na,nl,si,xa)
9	(se)	:	(at,ca)
10	(da,ka,na,xa)	:	(nl,si)
11	(kr)	:	(br,sc)
12	(ka)	:	(da,na,xa)
13	(ng)	:	(bu)
14	(xa)	:	(da,na)
15	(bo)	:	(sw)
16	(lu)	:	(to)
17	(nl)	:	(si)
18	(ge)	:	(gr,ro)
19	(at)	:	(ca)
20	(gr)	:	(ro)
21	(br)	:	(sc)
22	(da)	:	(na)
23	(ko)	:	(mo)

Table 6-2 : TreeSOM and Documented Similarities of KZN *Acacia* Species

Relation/Connection Shown by TreeSOM	Known Morphological Relationship
A. burkei and *A. nigrescens* are located next to each other in Figure 6-1 in the top left-hand corner of the TreeSOM model. In Figure 6-5 the species are clustered together as group **1**.	Some forms of *A. burkei* are known to be hard to distinguish from *A. nigrescens* [168, 179, 180, 229].
A. senegal, *A. ataxacantha* and *A. caffra* are located next to each other in Figure 6-1 in the left-hand centre of the TreeSOM model. In Figure 6-5 the species are clustered together as group **2.**	*A. ataxacantha* and *A. caffra* have been confused in the past [168, 179, 180, 229]. *A. ataxacantha* and *A. senegal* are classified next to each other by Ross, which is commonly thought to indicate that the author regarded the species as being close [179].
A. kraussiana, *A. brevispica* and *A. schweinfurthii* are located next to each other in Figure 6-1 in the bottom left-hand side of the TreeSOM model. In Figure 6-5 the species are clustered together as group **3.**	*A. brevispica* and *A. schweinfurthii* are sometimes difficult to differentiate [168, 179, 180, 229]. *A. kraussiana*, *A. brevispica* and *A. schweinfurthii* are classified next to each other by Ross [179].
A. borleae and *A. swazica* are located next to each other in Figure 6-1 in the right-hand top corner of the TreeSOM model. In Figure 6-5 the species are clustered together as group **4.**	*A. borleae* and *A. swazica* are classified next to each other by Ross [179].
A. luederitzii and *A. tortilis* are located next to each other in Figure 6-1 in the middle centre of the TreeSOM model. The species are also shown in Figure 6-5 to be clustered together as group **5.**	*A. luederitzii* and *A. tortilis* are classified next to each other by Ross [179].

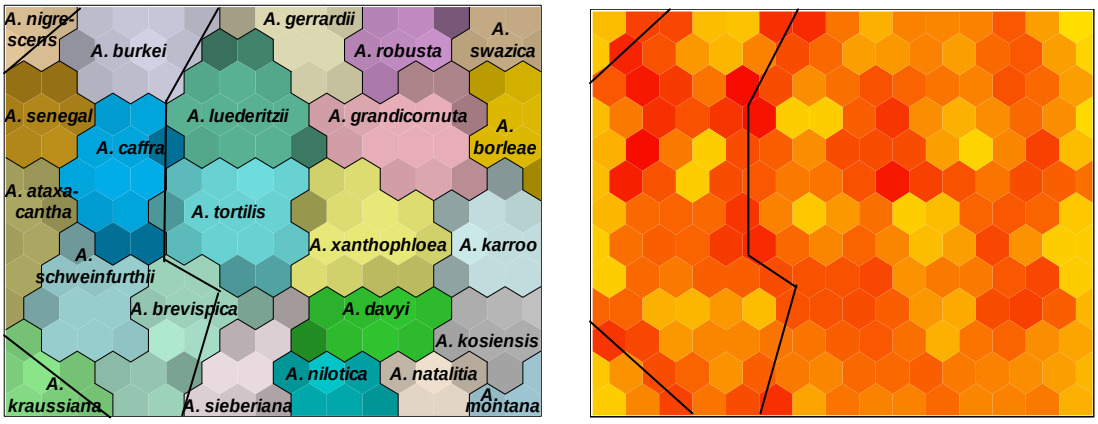
A. gerrardii, *A. grandicornuta* and *A. robusta* are shown located next to each other in Figure 6-1 in the right, top and centre of the TreeSOM model. The species are also shown in Figure 6-5 to be clustered together as group **6.**	Some *A. gerrardii* and *A. robusta* bear a strong resemblance to each other [180, p127]. *A. grandicornuta* and *A. robusta* are closely related [180, p129]. Ross classifies *A. gerrardii*, *A. grandicornuta* and *A. robusta* next to each other [179].
A. karroo, *A. davyi*, *A. natalitia* and *A. xanthophloea* are located next to each other in Figure 6-1 in the right, bottom and centre of the TreeSOM model. In Figure 6-5 The species are clustered together as group **7.**	*A. karroo* and *A. natalitia* were preciously classified together as part of the *A. karroo* complex [42].
A. nilotica and *A. sieberiana* are shown located next to each other in Figure 6-1 in the bottom centre, of the TreeSOM model. In Figure 6-5 the species are clustered together as group **8.**	These two species are not usually associated in modern published classificatory systems.
A. kosiensis and *A. montana* are shown located next to each other in Figure 6-1 in the bottom right-hand corner of the TreeSOM model. In Figure 6-5 the species are clustered together as group **9.**	*A. kosiensis* and *A. montana* were previously classified together as part of the *A. karroo* complex [42].
A. karroo, *A. borleae* and *A. swazica* are shown located next to each other in Figure 6-1 in the right-hand side of the TreeSOM model. In Figure 6-5 *A. borleae* and *A. swazica* are shown together as group **4**.	According to Ross *A. karroo* is related to the glandular-podded *Acacia* which include *A. borleae* and *A. swazica* [180, p94].

(continuation of Table 6-2)

The unified distance matrix (U-matrix) representation of the SOM visualizes the distances between the neurons. The distance between adjacent neurons is calculated and presented with different shades of colour between the adjacent nodes. A dark colour between the neurons corresponds to a large distance and therefore indicates a gap between the codebook vectors in the input space. A light colour between the

neurons signifies that the codebook vectors are close to each other in the input space. Light areas can be thought of as clusters and dark areas as cluster separators.

A U-matrix representation of the whole *Acacia* data is presented in Figure 6-6 (b) with the cluster map for the same data in Figure 6-6 (a) for comparison purposes. Lines have been superimposed on Figure 6-6 (b) to emphasize the wider spacing between some clusters. To the left of the central superimposed lines the species have hooked thorns, while to the right all the species have some straight thorns. These differences are emphasized by the separation between the clusters. Many of the smaller clusters are also evident in Figure 6-6 (b), for example, *A. nigrescens* is evident in the top left-hand corner and *A. kraussiana* is evident in the bottom left-corner.



(a) **(b)**

Figure 6-6 : U-Matrix representation of TreeSOM Model 1

Thus the cluster relations between closely related species are clearly illustrated by the U-matrix.

6.1.2 Evaluation of the TreeSOM Model Verification Results

Once the 30 whole training sets had each been used to train the network, 30 TreeSOM models were obtained. Each of these models was used in turn to see if the matching verification test set could be identified correctly. The recall function (where the cluster membership is recalled) was used for these experiments and the results were analysed. The results of these tests are shown in Table 6-3, and as recorded in the

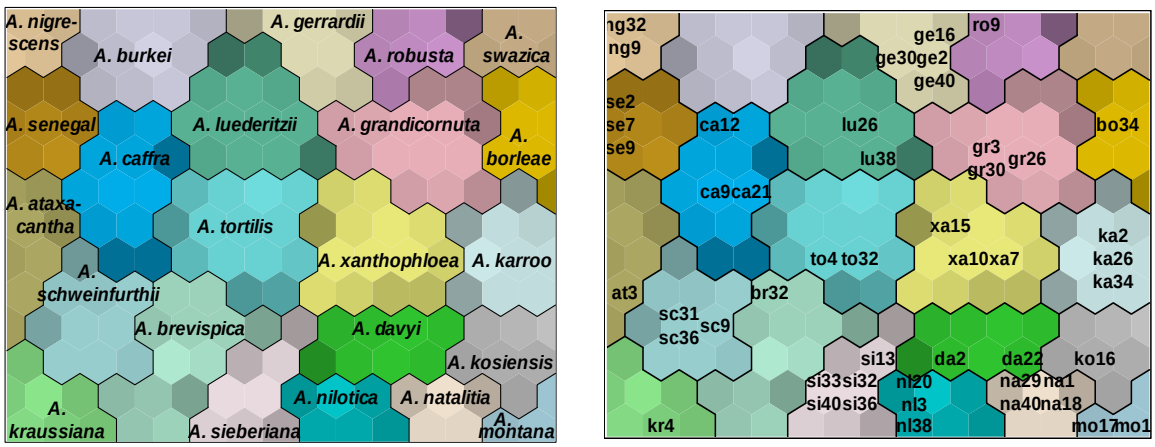
right-hand column of this table, the identification of the validation sets was successful.

Table 6-3 : 30-Fold Training and Verification Test Results

Training Set	Size	Result (No. of Unique Clusters)	Verification Test Set	Test Size	% Correctly Identified
1	870	23	1	50	100
2	890	23	2	30	100
3-30*	890	23	3-30	30	100

(In Table 6-3 in row 4 the results for training and verification testing pairs 3-30 have been condensed to one row because all the results were identical.)

For demonstration purposes one of these tests was repeated using the association function of the SOM software. In Figure 6-7 (a) a TreeSOM model is shown for training set 1 after the species labels had been added. In Figure 6-7 (b) the same model is associated with the verification test set 1. During this association the 50 verification test patterns were presented to TreeSOM to see if the model could identify the patterns correctly. The map in Figure 6-7 (b) demonstrates that TreeSOM was able to identify each of the patterns correctly.



(a) **(b)**

Figure 6-7 : TreeSOM Model 1 with Associated Verification Test Set

Table 6-4 tabulates the results of the association of verification test 1 with training map 1. Columns 1 and 3 give the label of the pattern to be identified, while columns 2

and 4 give the result of the identification. Thus, from Table 6-4 it can be seen that all 50 patterns were assigned to the correct locations on the map, i.e. that the TreeSOM had performed as was intended and had identified the specimen patterns correctly.

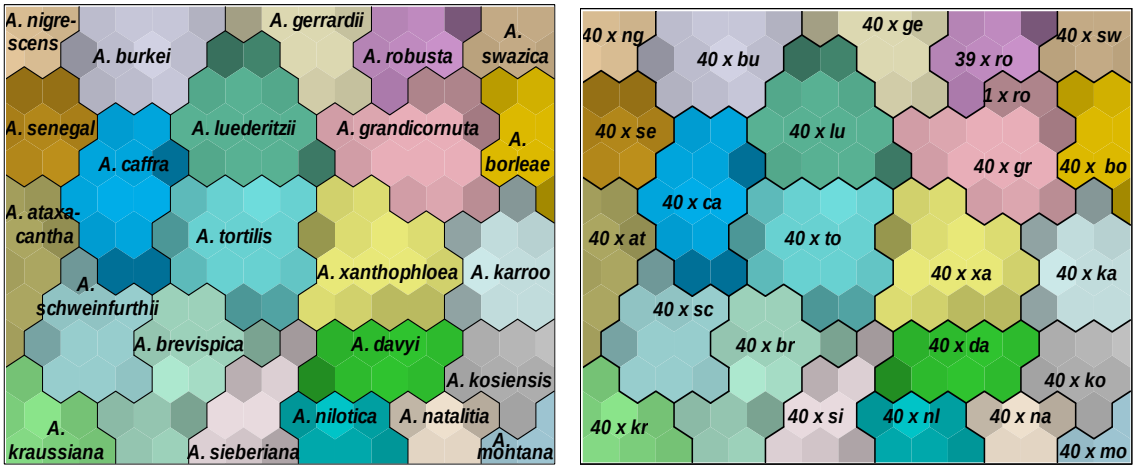
Table 6-4 : TreeSOM Associated Verification Data Set Results

Verification Specimen Code	Identity of Cluster	Verification Specimen Code	Identity of Cluster
at3	*A. ataxacantha*	**mo11, mo17**	*A. montana*
bo34	*A. borleae*	**na1, na18, na29, na40**	*A. natalitia*
br32	*A. brevispica*	**ng32, ng9**	*A. nigrescens*
ca12, ca21, ca9	*A. caffra*	**nl20, nl3, nl38**	*A. nilotica*
da2, da22	*A. davyi*	**ro9**	*A. robusta*
ge16, ge2, ge30, ge40	*A. gerrardii*	**sc31, sc36, sc9**	*A. schweinfurthii*
gr26, gr3, gr30	*A. grandicornuta*	**se2, se7, se9**	*A. senegal*
ka2, ka26, ka34	*A. karroo*	**si13, si32, si33, si36, si40**	*A. sieberiana*
ko16	*A. kosiensis*	**to32, to4**	*A. tortilis*
kr4	*A. kraussiana*	**xa10, xa15, xa7**	*A. xanthophloea*
lu26, lu38	*A. luederitzii*		

6.1.3 Evaluation of the TreeSOM Model Test Results

Each of the 30 TreeSOM models obtained from training the network was used to see if the model could identify an *Acacia* test data set accurately. The recall function was utilized for these tests.

In order to demonstrate the process visually, one map was selected and the whole test set was associated with the map. The output of that association is presented in Figure 6-8 (b). For clarity the labels on the maps show the total number of patterns that were associated with the clusters rather than displaying the individual pattern code names.



(a) **(b)**

Figure 6-8 : TreeSOM Model 1 with Associated Test Set

The TreeSOM model is presented Figure 6-8 (a) for comparison purposes. In Figure 6-8 (b) the map presented is the one obtained after 920 test specimens were associated with the selected model. In Figure 6-8 (b) it can be seen that TreeSOM was able to identify the 919 test specimens correctly, the only exception being that one *A. robusta* has been classified as an *A. grandicornuta* (second cluster from the top right corner). This misidentication is not unexpected as these species are very similar, and suggestions have been made that *A. grandicornuta* be placed under *A. robusta* because of their morphological similarity [179].

TreeSOM test error results are summarized in Table 6-5.

Table 6-5 : TreeSOM Test Error Results

Map No. X	Cluster No.	ID of Cluster	No. Correctly IDed in Cluster	Type of Error	No. of Errors	Errors in Cluster
1	2	**gr**	40	**ro** as **gr**	1	1
2 – 30					0	0
Total Errors						**1**
Error Rate						**0.004%**
Correct Rate						**99.996%**

As can be seen from Table 6-5, the only error that occurred was in map 1. Consequently, the TreeSOM had an error rate of 0.004% and a correct rate of 99.996% when performing these tests. The test data set had an average of 19.52 attribute values per specimen.

In the next section the test results are analyzed statistically.

6.1.4 Statistical Analysis of the TreeSOM Model Test Results

Table 6-6 presents the multi-class confusion matrix for the results depicted in Figure 6-8(b). The values on the major diagonal of the confusion matrix quantify the patterns that were correctly identified by TreeSOM. These are the **t**rue **p**ositive (TP) counts for each species. The off-major-diagonal values indicate the number of misidentified patterns (consisting of the **f**alse **n**egative (FN) and the **f**alse **p**ositive (FP) results). For example, in the **gr** row of Table 6-6, the FP count is the sum of all off-major-diagonal values in row **gr**, i.e. those patterns which are not actually *A. grandicornuta* species but which are predicted as *A. grandicornuta* species. The FP count for *A. grandicornuta* is 1 in this case as an *A. robusta* (column **ro**) has been predicted as *A. grandicornuta*. Similarly, the FN count for *A. robusta* is 1, i.e. the sum of all off-major-diagonal values in column **ro**.

Table 6-6 : TreeSOM Multi-Class Confusion Matrix

PREDICTED \ ACTUAL	at	bo	br	bu	ca	da	ge	gr	ka	ko	kr	lu	mo	na	ng	nl	ro	sc	se	si	sw	to	xa	FP
at	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bo	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
br	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bu	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ca	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
da	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ge	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gr	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
ka	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ko	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kr	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0
lu	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0
mo	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0
na	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0
ng	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0
nl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0
ro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0
sc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0
se	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0
si	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0
sw	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0
to	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0
xa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0
FN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	

The results in Table 6-6 show that the TreeSOM is able to identify KZN *Acacia* species accurately when presented with a wide range and diversity of attributes. The one error – the prediction of an *A. robusta* as an *A. grandicornuta* - was the only

misidentification made in any of the 30 TreeSOM models (it also only occurred with one model: the other 29 models were able to identify all the patterns correctly).

One of the ways to evaluate the results of classifiers is to calculate true positive, false positive and other related rates (discussed in Chapter 3). Choosing multiple class performance measures can become quite complicated, but according to Hand and Till [90] one way to evaluate performance is to derive pair-wise confusion matrices. This method yields an overall measure of how well each class is separated from all of the other classes. In this thesis this technique was followed.

Figure 6.9 shows how the pair-wise confusion matrices were derived.

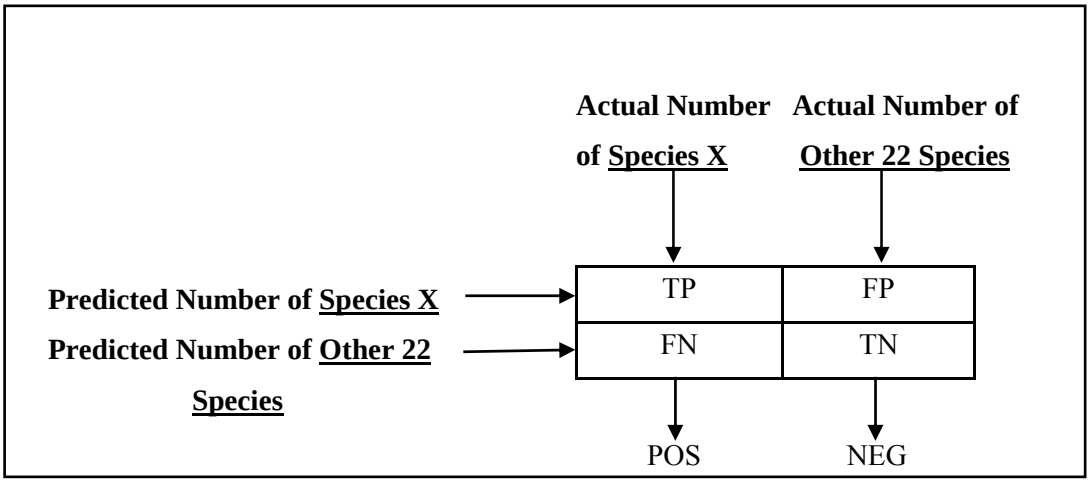


Figure 6-9 : Binary Class Confusion Matrix Template for each KZN *Acacia* Species

where TP, FP, FN and TN represent respective counts for true positive, false positive, false negative and true negative results,
POS is the total sum of actual positive patterns (TP + FN), and
NEG is the total sum of actual negative patterns (FP + TN).

With reference to Figure 6-9, the columns of the matrix display the actual/known results of the tests. In column 1 of the matrix the numbers of actual true positives and false negatives for species X are shown and in column 2 the numbers of actual false positives and true negatives for all the other 22 KZN *Acacia* species (i.e. other than the one used in column 1 and depicted as Species X) are shown. The rows of the matrix display the predicted results of the tests. In row 1 the number of predicted true

positives and false positives for species X is shown and in row 2 the number of predicted false negatives and true negatives is shown for all the other 22 KZN *Acacia* species (other than the one used in row 1). A binary confusion matrix was produced for each of the 23 KZN species.

Using the values from the binary class confusion matrices for the 30 TreeSOM test results the true positive and false positive rates were calculated for each species and are presented in Table 6-7. Averaged pair-wise comparisons of the species (classes) [90] have been used to calculate the fractions in this table. These results show that TreeSOM was able to separate each of the species with a very high degree of accuracy.

Table 6-7 : Fraction Metrics for TreeSOM Species

Species	TP	FP	Acc	Prec	Sp
at	1	0	1	1	1
bo	1	0	1	1	1
br	1	0	1	1	1
bu	1	0	1	1	1
ca	1	0	1	1	1
da	1	0	1	1	1
ge	1	0	1	1	1
gr	1	3.79E-05	0.999964	0.999187	0.999962
ka	1	0	1	1	1
ko	1	0	1	1	1
kr	1	0	1	1	1
lu	1	0	1	1	1
mo	1	0	1	1	1
na	1	0	1	1	1
ng	1	0	1	1	1
nl	1	0	1	1	1
ro	0.999167	0	0.999964	1	1
sc	1	0	1	1	1
se	1	0	1	1	1
si	1	0	1	1	1
sw	1	0	1	1	1
to	1	0	1	1	1
xa	1	0	1	1	1
Average	**0.999964**	**1.65E-06**	**0.999997**	**0.999965**	**0.999998**

Key for Table 6-7

Assessment Measure	Abbreviation	Formula
TP fraction	tp	tp = Sn = Sensitivity tp = TP / (FN+TP) (Eq. 3.3)
FP fraction	fp	fp = FP / (FP + TN) Eq. 6.1
Precision	Prec	Prec = PPV (Eq. 3.2) Prec = TP / (TP + FP)
Accuracy	Acc	Acc = (TN + TP) / TN + TP + FP + FN (Eq. 3.1)
Specificity	Sp	Sp = TN / (TN + FP) (Eq. 3.4)

Table 6-7 shows that except for the misidentification of *A. robusta* as *A. grandicornuta* the TreeSOM was able to identify all the other KZN *Acacia* species correctly. This misidentification, as previously noted, occurred on only one of the 30 TreeSOM models and was restricted to one specimen of *A. robusta* being identified as *A. grandicornuta*. In addition these species are biologically closely related so the misidentification is not unexpected. The results displayed in Table 6-7 are exceptionally good and demonstrate clearly that TreeSOM is well able to identify *Acacia* species.

The results of Table 6-7 were used to plot a ROC space diagram of the TP fractions against the FP fractions for each species. Figure 6-10 displays this graph.

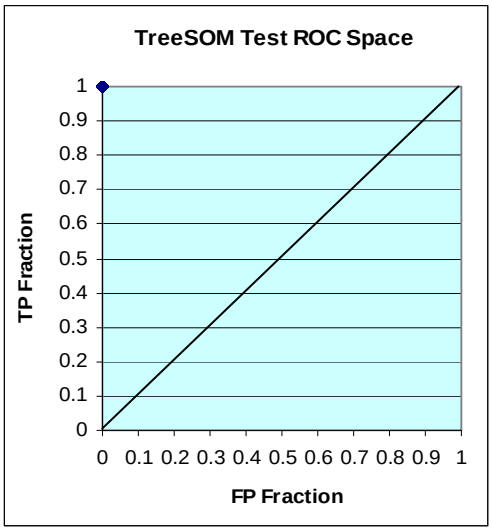


Figure 6-10 : TreeSOM Test ROC Space

6.2 The Habit and ThornSOM Models

The habit and thorn data set is a subset of the whole data set and was created by deleting/excluding all attributes other than the habit and thorn attributes. Consequently, only the attributes pertaining to habit and thorn characteristics of the KZN *Acacia* species form this data set: no flower, seed, pod or leaf characteristics are included. The training, verification, testing and analysis of results for the Habit and ThornSOM models are discussed in the following subsections. The habit and thorn data set will be referred to as the thorn data set in this chapter for the sake of brevity.

6.2.1 Evaluation of the Habit and ThornSOM Models

The patterns in the habit and thorn data set were pre-processed in the same way as was previously done for the whole data set. This resulted in 30 thorn training and verification pairs. The division of the thorn data set is shown in Table 6-8.

Table 6-8 : Makeup of Thorn Data Training and Cross-Validation Sets

Thorn Training Set Number	Size of Training Set	Maximum No of Attributes	Size of Verification Set
1	870	43	50
2 – 30	890 (each set)	43	30 (each set)

The 30 training sets were presented in turn to the SOM software using 103 neurons, and 30 ThornSOM maps were obtained. The ThornSOM cluster map obtained from one of these simulations (in this case the thorn training set 6) is presented in Figure 6-11. For convenience, each of the clusters in this figure has been labelled manually with the abbreviated name of the species that has been mapped to that cluster of nodes. The SOMine association function was used to perform the labelling. It can be seen that not all the 890 randomly shuffled patterns have been assigned to their own unique clusters. The map shows that the ThornSOM was able to learn from the input training data the differences between 21 of the 23 *Acacia* species, and it was able to cluster these 21 species uniquely in each of the 30 maps, i.e. in each of the 30 maps 21 clusters contained patterns belonging to one species only.

In the bottom right-side of Figure 6-11 it can be seen that the model did not cluster *A. robusta* and *A. grandicornuta* species separately (see the large cluster, right edge, second from bottom). These results are consistent with the close relationship of these species which has been documented in botanical literature [168, 179, 180].

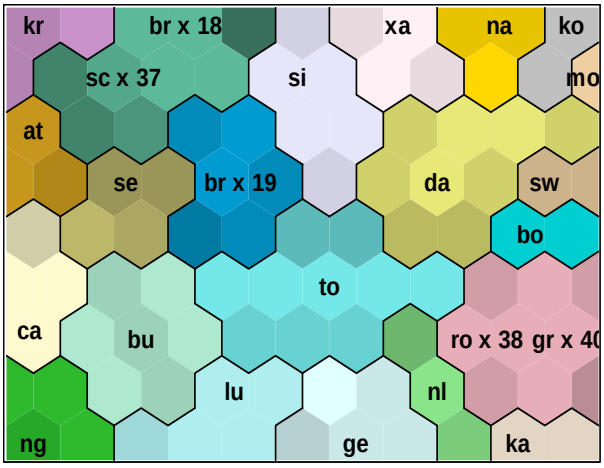


Figure 6-11 : ThornSOM Model 6

Additionally, in the top left-hand side of Figure 6-11 the model has classified 18 of the *A. brevispica* patterns together with 37 of the *A. schweinfurthii* patterns. The other 19 *A. brevispica* patterns have formed their own unique cluster (in the middle of the map, slightly left of centre).

Table 6-9 lists details of errors obtained in the 30 thorn training maps. Only the results of the training maps that had errors are presented. From this table it can be seen that the ThornSOM maps 6 and 11 were unable to separate *A. robusta* and *A. grandicornuta*. Similarly, maps 10, 21 and 27 classified some *A. grandicornuta* patterns as *A. robusta*. In maps 24 and 26 some *A. robusta* patterns have been classified as *A. grandicornuta*. As mentioned earlier in this chapter, the botanical literature and the whole data set map results already show that these species are very close. The inability of the ThornSOM to separate *A. robusta* from *A. grandicornuta*, though regrettable, is not surprising in view of the fact that the literature remarks on the similarity between these species. However, these results do show that the range and diversity of thorn attributes in this training data set is sometimes insufficient for the ThornSOM to be able to separate the *A. robusta* species from the *A. grandicornuta*

species. It must also be noted that only two of the 30 ThornSOM maps were unable to separate *A. robusta* species from the *A. grandicornuta* species.

Table 6-9 : Habit and ThornSOM Errors

Thorn Map X	No. of Clusters in Map X	ID of Cluster	Cluster No.	Type of Error	No. of Errors in Cluster	No. of Errors in Map X
6	22	**gr**	2	**ro** as **gr**	38	
		sc	4	**br** as **sc**	18	56
10	40	**ro**	30	**gr** as **ro**	4	4
11	22	**sc**	11	**br** as **sc**	17	
			10	**gr** as **ro**	38	55
21	43	**ro**	29	**gr** as **ro**	4	4
24	43	**gr**	14	**ro** as **gr**	4	
		gr	43	**ro** as **gr**	7	11
26	44	**gr**	19	**ro** as **gr**	10	10
27	40	**ro**	10	**gr** as **ro**	4	
		ro	26	**gr** as **ro**	2	6
Total errors						**146**
Error Rate						**0.55%**
Correct Rate						**99.45%**

The ThornSOM maps 6 and 11 have also misclassified some *A. brevispica* as *A. schweinfurthii*. The closeness of these two species has already been discussed in Table 6-2, and again this overlap between the species is not surprising. In fact Ross [180, p 43] reports about these species that *A. brevispica* subsp. *dregeana* is very variable and in southern Africa it bridges many of the discontinuities with *A. schweinfurthii* that exist further north in Africa. Ross then states “Consequently difficulty is sometimes experienced in southern Africa in distinguishing specimens of *A. brevispica* subsp. *dregeana* from *A. schweinfurthii*”

The results in Table 6-9 show that with the thorn training set (which used, at most, 43 attributes for training) could not always differentiate the 23 KZN *Acacia* on thorn

and habit characteristics alone. The number of attributes coupled with their similarity in some species was insufficient to classify correctly or differentiate between the species, even at the training level.

However, many of the relationships already demonstrated in the TreeSOM can also be seen in the thorn model. The component maps in Figure 6-12 depict some of these relationships.

The **BK_COL** plane (bark colour, top left map) in Figure 6-12 the relationship between *A. xanthophloea*, *A. sieberiana* and *A. davyi* is clearly shown (blue colour, top centre). These species all have pale or yellowish bark while most of the other species have darker bark (with the exception of *A. ataxacantha*).

In the **BK_TX** plane (bark texture, top centre map) the similar bark texture of several species can be seen. In the top right corner of the map: *A. xanthophloea*, *A. natalitia*, *A. kosiensis*, and *A. swazica* (blue) can have flaking bark; *A. xanthophloea*, *A. natalitia* and *A. kosiensis* (dark blue) also can have smooth bark; *A. davyi*, *A. montana* and *A. kosiensis* (bluish) can have fissured bark although *A. davyi* is usually corky. The species in the bottom half of the map (*A. caffra*, *A. nigrescens*, *A. burkei*, *A. luederitzii*, *A. tortilis*, *A. gerrardii*, *A. robusta* and *A. grandicornuta*) all have rough and fissured bark and are coloured red, yellow or orange.

Still with reference to Figure 6-12, the value map **BRLET_HR** (branchlet hair, top right map) represents the species which have hairy branchlets. Of particular note is that the young branchlets of *A. gerrardii* and *A. luederitzii* (bottom centre) are very hairy (orange colour). Also, *A. xanthophloea*, *A. davyi*, *A. natalitia A. kosiensis*, *A. montana*, *A. swazica*, *A. borleae*, *A. grandicornuta* and *A. robusta* are glabrous, or glabrous-to-pubescent, and are coloured green or blue.

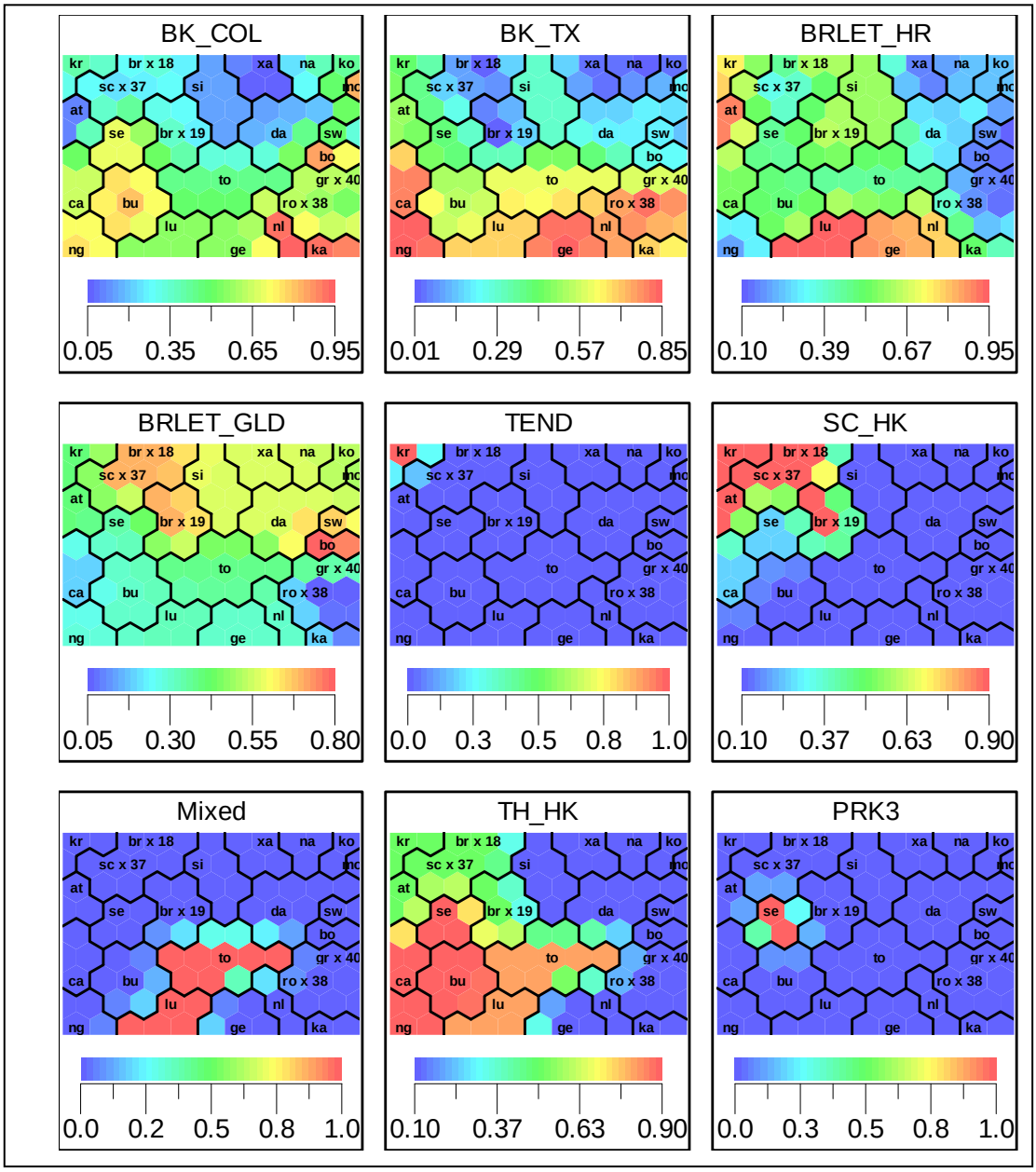


Figure 6-12 : Component Maps for Some Habit and Thorn Attributes

In the map **BRLET_GLD** (branchlet gland plane, centre row, left map) the branchlet gland attributes are shown: *A. brevispica* and *A. schweinfurthii* (red/orange colour, top left) have some glands, as does *A. swazica* (pale orange colour, right centre), while *A. borleae* has many glands (dark red colour, right, centre). The **TEND** map (tendrils, centre row, middle map) shows that the only KZN species that has tendrils is *A. kraussiana* (red colour, top left).

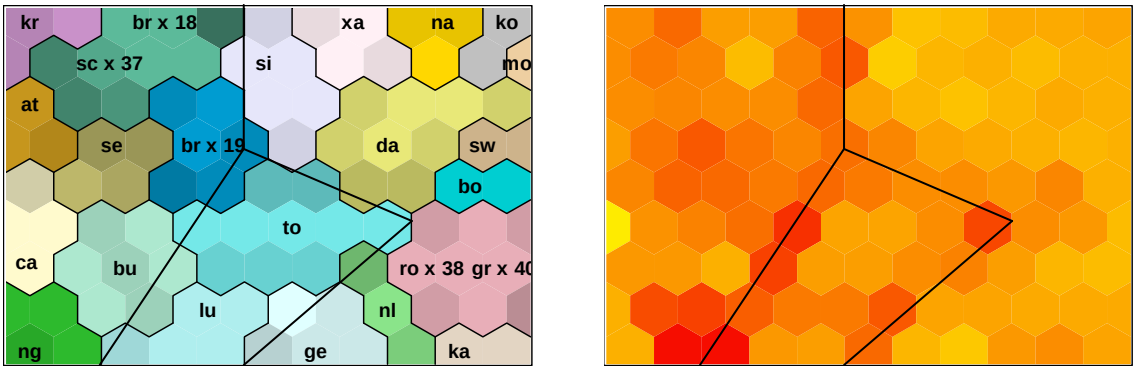
The **SC_HK** map (scattered hooks, centre row, right map) shows that *A. kraussiana*, *A. ataxacantha*, *A. schweinfurthii* and *A. brevispica* (top, left corner) are the species with thorns scattered between the nodes.

The component plane **Mixed** (mixed thorns, bottom left map) shows the clear differentiation of *A. luederitzii* and *A. tortilis* (bottom centre) from the other species. These two species each have some hooked thorns and some straight/straightish thorns.

The map labelled **TH_HK** (thorn hooked, bottom centre map) shows on the left the species that have at least some hooked thorns (*A. kraussiana*, *A. brevispica*, *A. schweinfurthii*, *A. ataxacantha*, *A. senegal*, *A. caffra*, *A. burkei*, *A. tortilis*, *A. nigrescens* and *A. luederitzii*), and on the right those with only straight thorns (*A. sieberiana*, *A. xanthophloea***,** *A. natalitia*, *A. kosiensis*, *A. montana*, *A. davyi*, *A. swazica*, *A. borleae*, *A. grandicornuta*, *A. robusta*, *A. karroo*, *A. nilotica* and *A. gerrardii*).

The last component map presented in Figure 6-12, **PRK3**, (prickles in threes, bottom right map) shows that *A. senegal* (middle, left) is the only KZN species that has thorns arranged in groups of three (rather than in pairs or singularly).

A U-matrix representation of the thorn *Acacia* data set is presented in Figure 6-13 (b) with the cluster map for the same data in Figure 6-13 (a) for comparison purposes.



(a) **(b)**

Figure 6-13 : U-Matrix Representation of Habit and ThornSOM Model

The closeness between the species with hooked thorns, the species with straight thorns and the species with some hooked and some straight or ‘straightish’ thorns is clearly demonstrated in Figure 6-13. The species with hooked thorns are clustered on

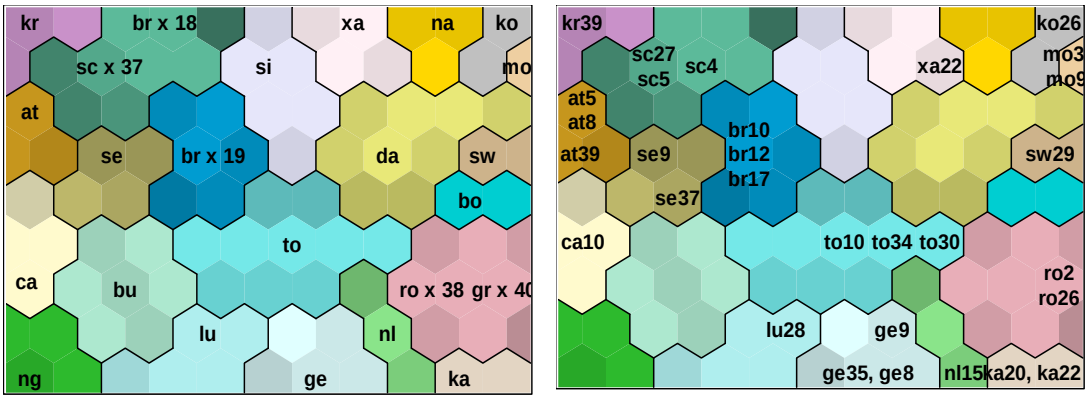
the left-most sides of the lines superimposed on the figures; species with hooked and straight/straightish thorns are clustered between the lower central lines; and the species with straight thorns (only) are clustered on the right-most sides of the lines.

Thus it can be seen that the ThornSOM models, despite being composed from less data than were used for the TreeSOMs, are still capable of recognizing patterns and relationships within the data.

6.2.2 Evaluation of the Habit and ThornSOM Model Verification Results

As was previously done for the TreeSOM models, each of the 30 ThornSOM models obtained from the thorn subset training process was used in turn to see if the corresponding verification test set could be identified correctly.

For demonstration purposes, the thorn verification test set 6 was mapped onto ThornSOM model 6 using the association function of the SOM software. In Figure 6-14 (a) the ThornSOM model 6 is shown after the species labels were added. The associated verification map is displayed in Figure 6-14 (b) together with the labels of the associated verification test patterns.



(a) **(b)**

Figure 6-14 : ThornSOM Model 6 with Associated Verification Set

The recall function was used with each of the 30 trained maps to obtain results for analysis of the verification test experiments. The results of the clusters which had misidentification errors are shown in Table 6.10. ThornSOM 6 had two errors in cluster two, where two *A. robusta* patterns were identified as *A. grandicornuta* species. In map 24 cluster 14, another *A. robusta* pattern was identified as *A.*

grandicornuta species. Map 11 had two types of errors: in cluster 10 one *A. grandicornuta* pattern was identified as *A. robusta*, and in cluster 11 two *A. brevispica* patterns were identified as *A. schweinfurthii*.

Table 6-10 : 30-Fold Thorn Verification Test Error Results

Thorn Map X	Size of Training Set	No. of Unique Clusters	Verification Test Set No.	Test Set Size	Cluster No.	Type of Error	No. of Errors in Cluster
6	890	20 (out of 22)	6	30	2	**ro** as **gr**	2
11	890	20 (out of 22)	11	30	10	**gr** as **ro**	1
	890	20 (out of 22)			11	**br** as **sc**	2
24	890	41 (out of 43)	24	30	14	**ro** as **gr**	1
Total Errors							**6**
Error Rate							**0.65%**
Correct Rate							**99.35%**

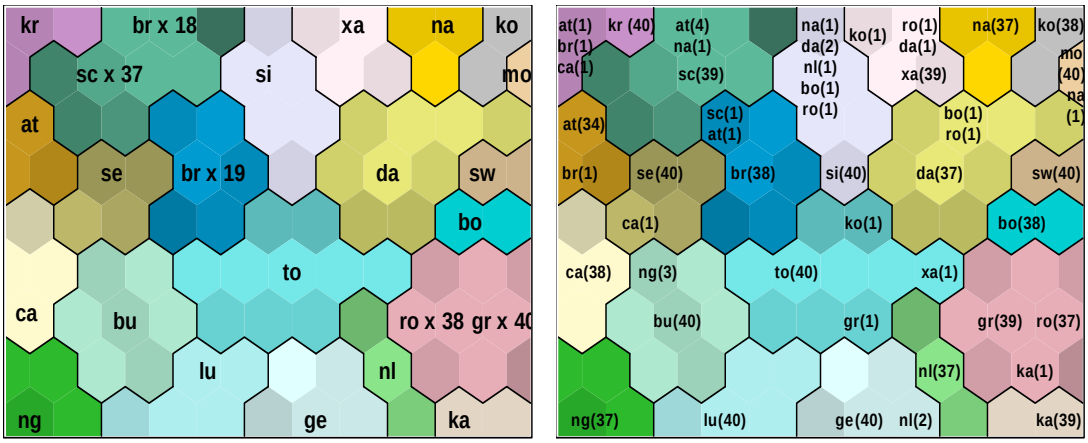
The results in Table 6-10 show that there were six errors for the 30 ThornSOMs verification tests. Consequently, the average error rate is 0.65%.

6.2.3 Evaluation of the Habit and ThornSOM Model Test Results

Each of the 30 ThornSOM models obtained from training the network was used to see if the model could identify an *Acacia* thorn test set accurately. This test set was composed of 920 unseen patterns and consisted of up to 43 possible attributes. However, the test set differed from the training data set in that many more attribute values were missing. Whereas the training data set was as complete as possible, the test set was sparsely populated. As mentioned before, this is the normal situation in nature where only a few attributes are present or are observed at one time. In the habit and thorn data set the average number of attributes per test specimen was 8.61.

The recall function was utilized for performing these tests. In addition, in order to demonstrate the results visually, a trained map was selected and the test set was associated with the map and the output is presented in Figure 6-15 (b). For clarity the labels on the maps show the number of patterns that were associated with the cluster rather than the individual pattern code names. The ThornSOM model is presented in Figure 6-15 (a) for comparison purposes.

The map presented in Figure 6-15 (b) is the one obtained after 920 test specimens were associated with the model in Figure 6-15 (a). It can be seen that ThornSOM correctly identified most of the 920 thorn test patterns. Although the trained ThornSOM was unable to separate *A. robusta* and *A. grandicornuta* species, 39 of the *A. grandicornuta* and 37 of the *A. robusta* test patterns were mapped correctly to the same cluster (see large cluster on right edge, second from bottom of Figure 6-15 (b)). An *A. karroo* pattern was also grouped in this cluster and so was incorrectly identified as *A. grandicornuta*/*A. robusta* species. Most of the other errors found are for species that are biologically closely related, and these have already been discussed. However, there are a few misidentifications for which no apparent explanations could be found.



(a) **(b)**

Figure 6-15 : ThornSOM Model 6 with Associated Test Set

The errors for each of the 30 ThornSOM models are tabulated in Table 6-11. The average error rate for the ThornSOM tests was 3.44%. Thus the ThornSOM had an average correct rate of over 96% even though the test data had many missing values. The maximum number of thorn attributes that could have been present was 43 but in the thorn test set there was an average of 8.61 attribute values present per specimen. Despite the high number of missing values the correct identification rate of 96.56% is statistically significant.

Table 6-11 : Habit and ThornSOM Test Error Results

Trained Map X	Total Errors	Total Correct	Trained Map X	Total Errors	Total Correct
1	35	885	16	32	888
2	25	895	17	33	887
3	32	888	18	30	890
4	31	889	19	28	892
5	28	892	20	29	891
6	33	887	21	29	891
7	33	887	22	29	891
8	31	889	23	32	888
9	23	897	24	41	879
10	33	887	25	28	892
11	70	850	26	34	886
12	27	893	27	33	887
13	29	891	28	24	896
14	31	889	29	32	888
15	28	892	30	27	893
Total Error					**950**
Error Rate					**3.44%**
Correct Rate					**96.56%**

For clarity, Table 6-12 tabulates the results of the ThornSOM map 6 test. Only the clusters which had errors are shown. Besides showing other errors, the table shows that one *A. natalitia* has been identified as an *A. schweinfurthii*. This is shown in cluster 4 of Table 6-12 and can also be seen in Figure 6-15 (b) (large cluster on the top edge, second from the left). As *A. natalitia* has straight thorns and *A. schweinfurthii* has hooked thorns this is an obvious error.

The misidentification of *A. kosiensis*, *A. xanthophloea* and *A. grandicornuta* as *A. tortilis* (cluster 1 in Table 6-12 and the cluster slightly below centre in Figure 6-15 (b)) also does not have an apparent biological explanation. The *A. tortilis* species, which has some hooked thorns and straight thorns, does differ from the other species which have straight thorns only. Similarly the misidentification of *A. caffra* as *A. kraussiana* does not have an obvious explanation although both have hooked thorns (see cluster 17 in Table 6-12 and in Figure 6-15 (b) top left corner).

Table 6-12 : ThornSOM Model 6 Test Error Results

Cluster No.	ID of Cluster	No. Correctly IDed in Cluster	Type of Error	No. of Errors	Errors in Cluster
1	**to**	40	**gr** as **to** **ko** as **to** **xa** as **to**	1 1 1	3
2	**gr** & **ro**	39 + 37	**ka** as **gr/ro**	1	1
3	**da**	37	**bo** as **da** **ro** as **da**	1 1	2
4	**sc**	39	**at** as **sc** **na** as **sc**	4 1	5
5	**bu**	40	**ng** as **bu**	3	3
6	**br**	38	**sc** as **br** **at** as **br**	1 1	2
7	**si**	40	**na** as **si** **da** as **si** **nl** as **si** **bo** as **si** **ro** as **si**	1 2 1 1 1	6
8	**ge**	40	**nl** as **ge**	2	2
10	**se**	40	**ca** as **se**	1	1
11	**xa**	39	**da** as **xa** **ro** as **xa** **ko** as **xa**	1 1 1	3
15	**at**	34	**br** as **at**	1	1
17	**kr**	40	**at** as **kr** **br** as **kr** **ca** as **kr**	1 1 1	3
22	**mo**	40	**na** as **mo**	1	1
Total Errors					**33**
Error Rate					**3.59%**
Correct Rate					**96.41%**

The ThornSOM test results are analyzed further in the next subsection.

6.2.4 Statistical Analysis of the Habit and ThornSOM Model Test Results

The multi-class confusion matrix for the results obtained in Figure 6-15(b) is presented in Table 6-13. As already mentioned, in the matrix the values on the major-diagonal of the matrix represent the patterns that were correctly identified. The sum of the off-major-diagonal values in each row indicates the FP for the predicted species.

Similarly, the sum of the off-major-diagonal values in each column indicates the FN for the actual species in that column.

Table 6-13: Multi-Class Confusion Matrix for ThornSOM Map 6 Test

		ACTUAL																							
		at	bo	br	bu	ca	da	ge	gr	ka	ko	kr	lu	mo	na	ng	nl	ro	sc	se	si	sw	to	xa	FP
PREDICTED	at	34	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	bo	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	br	1	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
	bu	0	0	0	40	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	3
	ca	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	da	0	1	0	0	0	37	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2
	ge	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
	gr	0	0	0	0	0	0	0	39	1	0	0	0	0	0	0	0	37	0	0	0	0	0	0	38
	ka	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ko	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	kr	1	0	1	0	1	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	3
	lu	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0
	mo	0	0	0	0	0	0	0	0	0	0	0	0	40	1	0	0	0	0	0	0	0	0	0	1
	na	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0
	ng	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0
	nl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0
	ro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	sc	4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	39	0	0	0	0	0	5
	se	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	1
	si	0	1	0	0	0	2	0	0	0	0	0	0	0	1	0	1	1	0	0	40	0	0	0	6
	sw	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0
	to	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	40	1	3
	xa	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	39	3
	FN	6	2	2	0	2	3	0	1	1	2	0	0	0	3	3	3	40	1	0	0	0	0	1	

In the confusion matrix it can be seen that all the *A. robusta* were misidentified (37 were identified as *A. grandicornuta*, one as *A. davyi*, one as *A. sieberiana* and one as *A. xanthophloea*). The identification of 37 *A. robusta* as *A. grandicornuta* meant that the ThornSOM model was unable to separate the test patterns of these two species. Other misidentifications which occurred can be seen by inspecting Table 6-13. In the table the columns represent the actual identity of the pattern, and the rows represent what the pattern was predicted to be.

Using the values from the confusion matrices for the 30 ThornSOM test results the true positive, false positive, accuracy, precision and specificity rates were calculated and are presented in Table 6-14. Averaged pair-wise comparisons of the classes [90] from the 30 ThornSOM test simulations have been used to calculate the fractions in this table. From the rates shown in Table 6-14 it can be seen that the average FP rate for ThornSOM is 0.16% and the TP rate is 96.42%. This demonstrates that the models were highly successful in identifying *Acacia* specimens using only thorn data.

Table 6-14 : Average ThornSOM Rate Fractions

Species	TP	FP	Acc	Prec	Sp
at	0.900833	0.000492	0.995217	0.988807	0.999508
bo	0.946667	0.000152	0.997536	0.996624	0.999848
br	0.982500	0.004545	0.994891	0.910202	0.995455
bu	0.958333	0.002424	0.995870	0.949884	0.997576
ca	0.973333	0.000152	0.998696	0.996706	0.999848
da	0.921667	0.000303	0.996304	0.993095	0.999697
ge	0.999167	0.002879	0.997210	0.941356	0.997121
gr	0.873333	0.002803	0.991812	0.922481	0.997197
ka	0.974167	0.000833	0.998080	0.981977	0.999167
ko	0.967500	0.000000	0.998587	1.000000	1.000000
kr	0.993333	0.000417	0.999312	0.991130	0.999583
lu	1.000000	0.000492	0.999529	0.989431	0.999508
mo	0.953333	0.000152	0.997826	0.996623	0.999848
na	0.996667	0.003333	0.996667	0.931915	0.996667
ng	0.961667	0.002235	0.996196	0.952756	0.997765
nl	0.942500	0.001477	0.996087	0.967039	0.998523
ro	0.895000	0.004053	0.991558	0.896325	0.995947
sc	0.958333	0.001326	0.996920	0.972358	0.998674
se	1.000000	0.000265	0.999746	0.994309	0.999735
si	0.991667	0.002803	0.996957	0.942429	0.997197
sw	1.000000	0.001250	0.998804	0.973738	0.998750
to	1.000000	0.000833	0.999203	0.982343	0.999167
xa	0.987500	0.004167	0.995471	0.915953	0.995833
Average	**0.964239**	**0.001625**	**0.996890**	**0.964673**	**0.998375**

Using the FP and TP rates from Table 6-14, a graph showing the ThornSOM test ROC space has been drawn up and is presented in Figure 6-16. This ROC space graph shows that the ThornSOM performed much better than the average guess, i.e. all results are above the diagonal line. The (FP, TP) points for *A. grandicornuta* (0.002803, 0.873333) and *A. robusta* (0.004053, 0.895000) are the furthest from the perfectly-performing-classifier point (which is (0, 1)). The (FP, TP) points for *A. ataxacantha* (0.000492, 0.900833) and *A. davyi* (0.000303, 0.921667) are the next furthest points from the ideal points. Plotting FP and TP values gives a good indication of the true ability of the ThornSOM to identify thorn data. Again the results

show that ThornSOM was able to identify thorn data successfully as it performed well above the diagonal line superimposed on Figure 16-16 for all the species.

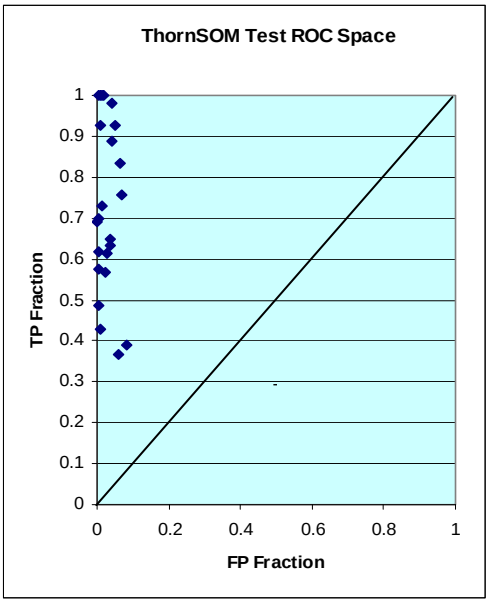


Figure 6-16 : ThornSOM Test ROC Space

In the next section the FlowerSOM models are discussed.

6.3 The FlowerSOM Models

The flower data set is a subset of the whole data and was created by deleting all attributes other than the flower attributes. The training, verification and testing procedures performed are the same as those already carried out on the whole, and on the habit and thorn data sets. These procedures will therefore only be discussed briefly for the FlowerSOM models.

6.3.1 Evaluation of the FlowerSOM Models

The patterns in the flower data set were pre-processed in the same way as was previously done for the whole data set. This resulted in 30 flower training and verification pairs. The division of the flower data set is shown in Table 6-15.

Table 6-15 : Makeup of Flower Data Cross-Validation Sets

Flower Training Set Number	Size of Training Set	Maximum No of Attributes	Size of Verification Set
1	870	19	50
2 – 30	890 (each set)	19	30 (each set)

After training SOMs using different numbers of neurons, it was found that the SOM produced the best results with 93 neurons. Map 2, which was formed with 93 neurons, is displayed in Figure 6-17.

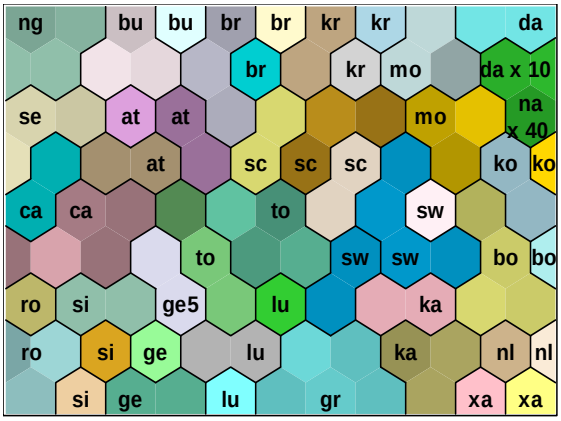


Figure 6-17 : FlowerSOM Model 2

The FlowerSOM training maps show that the training of the flower data set on its own was not as successful as the TreeSOM data and ThornSOM data. The total number of attributes possible was at most 19. This number of attributes appears to be too small to separate the flower data into unique *Acacia* clusters with no errors or just a few errors.

There was a great variation in the number of clusters formed by the flower maps. In some of the maps, 49 clusters were formed, some of which were very small and had very few specimens mapped to them. In each of two maps only four clusters were formed, while other maps had 21 clusters or more. This appears to suggest that the FlowerSOM models had not stabilized.

For clarity, Table 6-16 tabulates the error results of the FlowerSOM. Only statistics of the maps which had errors are displayed.

Table 6-16 : Misidentification Errors in FlowerSom

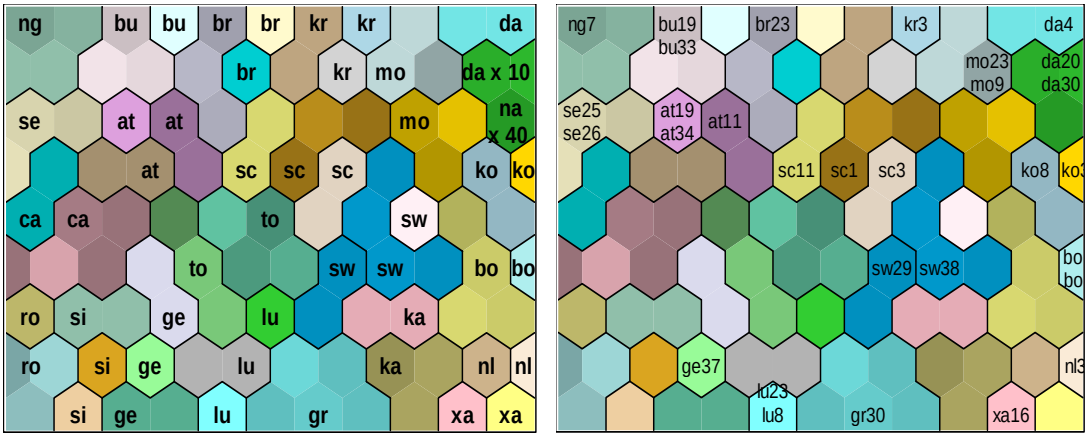
Flower Map X	No. of Clusters in Map X	ID of Cluster	Cluster No.	No. in Cluster	Type of Error	No. of Errors in Cluster	No. of Errors in Map X
1	22	na	15	76	da as na	38	38
2	49	da	12	50	na as da	10	10
3	22	da ge si	6 10 13	79 50 27	na as da si as ge ge as si	39 22 9	70
4	50	na	21	51	da as na	12	12
5	20	ge to na nl	1 2 3 19	77 46 76 79	si as ge lu as to da as na bo as nl	37 8 36 39	120
6	22	da	10	77	na as da	38	38
7	49	na	9	49	da as na	12	12
8	21	si na	1 10	76 76	ge as si da as na	38 37	75
9	49	na	10	51	da as na	12	12
10	22	to na	4 8	49 80	lu as to da as na	10 40	50
11	21	si na	2 15	76 80	ge as si da as na	38 40	78
12	21	ge na	1 10	79 78	si as ge da as na	39 39	78
13	21	si na	2 11	76 78	ge as si da as na	36 39	75
14	22	na	8	76	da as na	38	38
15	21	si to na	1 3 14	76 56 80	ge as si lu as to da as na	38 17 40	95
16	21	si to da	1 6 12	78 41 79	ge as si lu as to na as da	39 5 39	83
17	21	si da	1 9	77 78	ge as si na as da	37 38	75
20	22	na	9	79	da as na	39	39
21	21	ge lu na xa	1 3 11 15	78 49 79 52	si as ge to as lu da as na nl as xa	38 10 39 14	101

22	**4**	**si** **ko** **se** **sc**	**1** **2** **3** **4**	**230** **345** **196** **119**	**ge, gr, lu, ro** & **to** as **si** **bo, da, ka, mo, na, nl** & **sw** as **xa** as **ko** **at, bu, ca,** & **ng** as **se** **br** & **kr** as **sc**	**39, 37, 38,** **37, 39** **39, 38,** **39, 38,** **36, 38,** **38, 39** **40, 38,** **40, 38** **39, 40**	**730**
23	**21**	**ge** **da**	**3** **9**	**79**	**si** as **ge** **na** as **da**	**39** **37**	**76**
24	**21**	**si** **to** **na**	**3** **5** **8**	**76** **41** **80**	**ge** as **si** **lu** as **to** **da** as **na**	**37** **1** **40**	**78**
25	**48**	**na**	**12**	**50**	**da** as **na**	**12**	**12**
26	**4**	**gr** **za** **se** **sc**	**1** **2** **3** **4**	**232** **348** **191** **119**	**ge, lu, ro, si** & **to** as **gr** **bo, da, ka, ko, mo, na, nl** & **sw** as **xa** **at, bu, ca,** & **ng** as **se** **br** & **kr** as **sc**	**40, 37, 39, 38,** **38** **39, 39, 38, 39,** **39, 39, 38, 38** **40, 37, 38, 36,** **39, 40**	**731**
27	**22**	**da**	**8**	**79**	**na** as **da**	**39**	**39**
Total Errors							**2765**
Error Rate							**17.26%**
Correct Rate							**82.74%**

(Table 6-16 continued)

6.3.2 Evaluation of the FlowerSOM Model Verification Results

The verification sets presented to the trained flower maps showed few errors other than those to be expected considering the training errors. The trained FlowerSOM using data set 2 is shown in Figure 6-18 (a). The results of presenting the verification set to this map are displayed in Figure 6-18 (b). From this map it can be seen that two *A. davyi* patterns (da20 and da30) were mapped to the *A. natalitia* cluster (second cluster from the top, right side). These were the only errors that occurred on this map, although it is notable that the *A. natalitia* cluster had ten *A. davyi* patterns mapped to it during the training stage.



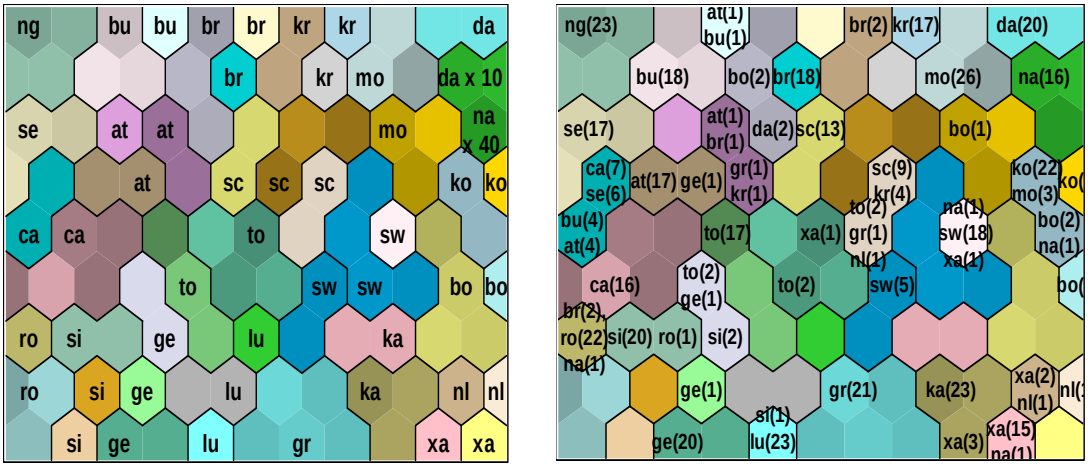
(a) **(b)**

Figure 6-18 : FlowerSOM Model 2 with AssociatedVerification Test Set

6.3.3 Evaluation of the FlowerSOM Model Test Results

Although there were originally 920 test patterns some of them did not have any flower data in them. The flower test data set was randomly shuffled and presented to each of the 30 FlowerSOM models in turn. SOM ignored the empty patterns entirely, hence there were effectively only 534 testing flower specimens in total.

In order to demonstrate the results visually, a trained map was selected (in this case map 2) and the test set was associated with this map. The output of the association is presented in Figure 6-19 (b).



(a) **(b)**

Figure 6-19 : FlowerSOM Model 2 with Associated Test Set

Figure 6-19 (b) shows that there were 69 misidentifications. For clarity the labels on the maps show the number of patterns that were associated with the cluster rather than the individual pattern code names. The FlowerSOM model is presented in Figure 6-19 (a) for comparison purposes. The map presented in Figure 6-19 (b) is the one obtained after the flower test specimens were associated with the map in Figure 6-19 (a). Again, considering the problems already existing with the trained map, the FlowerSOM was able to identify some patterns accurately. However, there were numerous misidentifications which are evident from studying Figure 6-19 (b).

The overall results obtained from presenting the test patterns to the 30 FlowerSOM models are summarized in Table 6-17.

Table 6-17 : FlowerSOM Test Results

Trained Map X	Total Errors	Total Correct	Trained Map X	Total Errors	Total Correct
1	104	430	16	93	441
2	69	465	17	106	428
3	103	431	18	80	454
4	82	452	19	75	459
5	135	399	20	97	437
6	87	447	21	111	423
7	68	466	22	449	85
8	116	418	23	106	428
9	76	458	24	109	425
10	88	446	25	75	459
11	112	422	26	451	83
12	109	425	27	79	455
13	105	429	28	80	454
14	99	435	29	78	456
15	110	424	30	61	473
Total Error					**3513**
Error Rate					**21.93%**
Correct Rate					**78.07%**

Table 6-18 breaks down the errors that occurred when FlowerSOM model 2 was presented with the test data set. Many of the misidentifications do not appear to make sense. For example, in cluster 6 *A. kraussiana* and *A. schweinfurthii* have both been identified as *A. borleae*. Also, in cluster 17 *A. grandicornuta* has been identified as *A. ataxacantha*. However, results like this might have some basis which could be

revealed if further studies were performed using, for instance, data for *Acacia* from the rest of southern Africa, or if studies were performed to look at the ancestral relationships between the *Acacia* species.

Table 6-18 : FlowerSOM Model 2 Test Results

Cluster No.	ID of Cluster	No. of Tests per Cluster	Type of Error	No. of Errors	Errors in Cluster
4	**to**	3	**xa** as **to**	1	1
6	**bo**	11	**da** as **bo** **kr** as **bo** **sc** as **bo** **xa** as **bo**	2 2 2 2	8
10	**ka**	26	**xa** as **ka**	4	4
12	**na**	17	**nl** as **na**	3	3
17	**at**	4	**br** as **at** **gr** as **at** **kr** as **at**	1 1 1	3
19	**ca**	21	**se** as **ca** **bu** as **ca** **at** as **ca**	6 4 4	14
20	**at**	18	**ge** as **at**	1	1
21	**sc**	17	**kr** as **sc** **to** as **sc** **nl** as **sc** **gr** as **sc**	4 2 1 1	8
23	**ko**	28	**na** as **ko** **bo** as **ko** **mo** as **ko**	1 2 3	6
28	**si**	21	**ro** as **si**	3	3
30	**si**	5	**ge** as **si** **to** as **si**	1 2	3
31	**at**	3	**bu** as **at**	2	2
33	**ro**	25	**br** as **ro** **na** as **ro**	2 1	3
34	**sw**	20	**na** as **sw** **xa** as **sw**	1 1	2
43	**xa**	3	**nl** as **xa**	1	1
47	**lu**	24	**si** as **lu**	1	1
48	**xa**	21	**na** as **xa**	6	6
Total Errors					**69**
Error Rate					**12.92%**
Correct Rate					**87.08%**

6.3.4 Statistical Analysis of the FlowerSOM Test Results

Confusion matrices were drawn up for the results of each FlowerSOM model and the metrics obtained are shown in Table 6-19.

Table 6-19 : Fraction Metrics for FlowerSOM Species

Species	TP	FP	Acc	Prec	Sp
at	0.747826	0.002283	0.986954	0.880152	0.997717
bo	0.678261	0.000457	0.985705	0.887524	0.999543
br	0.769565	0.000457	0.989638	0.922754	0.999543
bu	0.802899	0.004044	0.98764	0.848285	0.995956
ca	0.957971	0.023418	0.97578	0.762332	0.976582
da	0.469697	0.006901	0.971536	0.428903	0.993099
ge	0.688406	0.012785	0.974345	0.578878	0.987215
gr	0.855072	0.003066	0.990824	0.874309	0.996934
ka	0.915942	0.017352	0.979775	0.672654	0.982648
ko	0.984058	0.032746	0.967978	0.794114	0.967254
kr	0.717391	0.002348	0.985581	0.877066	0.997652
lu	0.928986	0.006458	0.990762	0.82834	0.993542
mo	0.783908	0.002574	0.98583	0.885889	0.997426
na	0.36087	0.01696	0.956242	0.454418	0.98304
ng	0.933333	0.005153	0.992197	0.841955	0.994847
nl	0.785507	0.00287	0.988015	0.878864	0.99713
ro	0.921739	0.006132	0.990762	0.817658	0.993868
sc	0.795652	0.015329	0.976529	0.758373	0.984671
se	0.744928	0.003979	0.985206	0.855612	0.996021
si	0.595652	0.010241	0.972784	0.580869	0.989759
sw	0.933333	0.020939	0.977091	0.632831	0.979061
to	0.881159	0.025114	0.970849	0.802134	0.974886
xa	0.689855	0.007567	0.979401	0.775348	0.992433
Average	**0.780087**	**0.009964**	**0.980931**	**0.766924**	**0.990036**

The TP and FP fractions displayed in Table 6-19 were used to draw up a graph of the FlowerSOM ROC space. This graph is displayed in Figure 6-20.

The rates shown in Table 6-19, and the ROC graph displayed in Figure 6-20, show that the TP rate for *A. natalitia* is low (0.36087) because of the misidentification of *A. natalitia* which occurred with nearly every map. Similarly, the identification of some *A. davyi* as *A. natalitia* (and as other species) resulted in a low TP rate (0.469697) for *A. davyi*. Low TP rates were also demonstrated by *A. sieberiana* and *A. gerrardii*. Despite these relatively low results the FlowerSOM models demonstrated that, even

with data containing an average of 4.62 attribute values per specimen, the FlowerSOMs could still differentiate many of the *Acacia* test specimens.

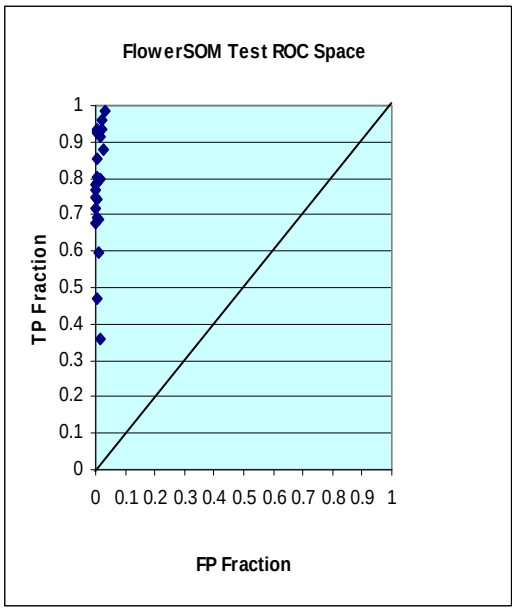


Figure 6-20 : FlowerSOM Test ROC Space

The Seed and PodSOM and the LeafSOM will be discussed in the next two sections.

6.4 The Seed and PodSOM Models

The same procedure as was followed for the other models was performed using the seed and pod data set. The maps produced and results obtained are discussed in the following sections, but in less detail than for TreeSOM and ThornSOM.

6.4.1 Evaluation of the Seed and PodSOM Models

The SOM was trained with 870/890-pattern seed and pod data sets and 587 neurons. Each specimen pattern contained values for up to 25 attributes. The Seed and PodSOM model obtained using training data set 23 is presented in Figure 6-21. As shown in this figure, 23 clusters were formed with two clusters having two misidentifications. These misidentifications are shown in the top right-hand corner of

the map where two *A. burkei* were mapped to an *A. caffra* cluster. Also near the top right-hand corner, the second cluster from the top shows that eight *A. ataxacantha* were mapped to an *A. burkei* cluster. The other 21 clusters on the map formed unique species clusters.

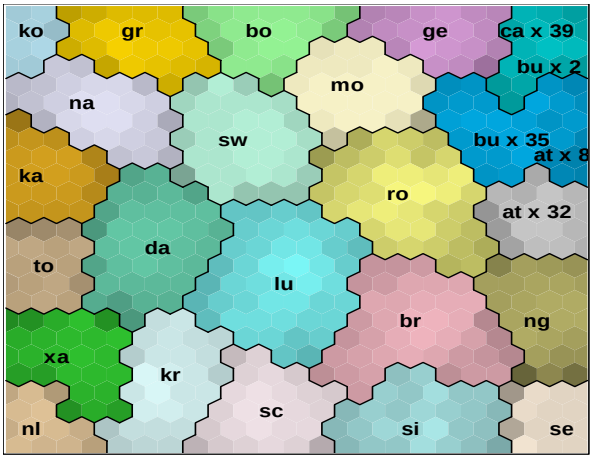


Figure 6-21 : Seed and PodSOM Model 23

The errors obtained in the 30 Seed and PodSOM maps are shown in Table 6-20.

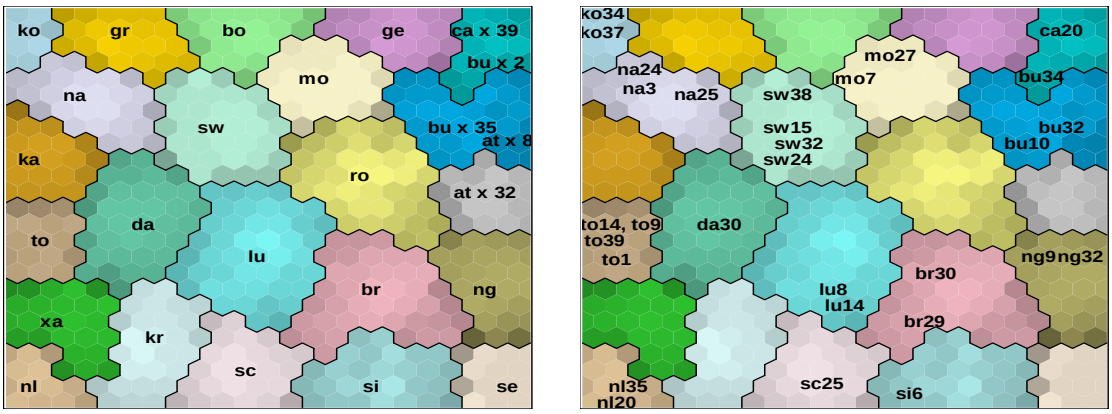
Table 6-20 : Misidentification Errors in Seed and PodSOMs

Trained Map X	No. of Clusters	ID of Cluster	Cluster No.	No. in Cluster	Type of Error	No. of Errors in Cluster	No. of Errors
2	23	bu	2	47	at as bu	9	9
21	23	ca	21	38	bu as ca	2	2
23	23	bu	6	43	at as bu	8	
		ca	19	41	bu as ca	2	10
30	23	bu	5	46	at as bu	8	8
Total Errors							29
Error Rate							0.11%
Correct Rate							99.89%

6.4.2 Evaluation of the Seed and PodSOM Model Verification Results

Once the Seed and PodSOM models had been formed they were verified using the remaining data sets left out during the training sessions. Figure 6-22 shows the results of mapping the verification test to map 23. Figure 6.22 (a) presents the PodSOM

obtained from using training map 23, and the map in Figure 6.22 (b) is obtained from associating verification set 23 with this map. The map in Figure 6.22 (b) shows there was only one type of error, with an *A. burkei* being identified as an *A. caffra*.



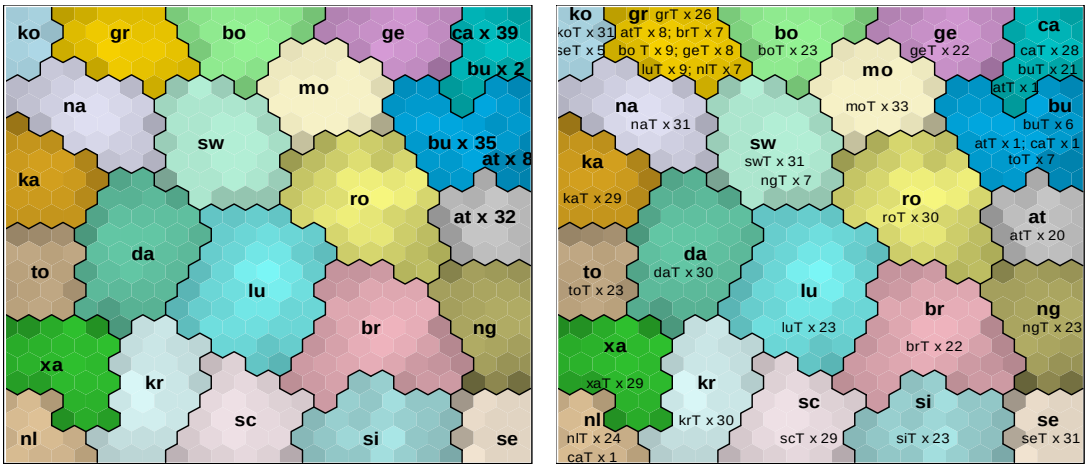
(a) **(b)**

Figure 6-22 : Seed and PodSOM Verification for Map 23

6.4.3 Evaluation of the Seed and PodSOM Model Test Results

The seed and pod test data set, as was the case with the flower test data set, included specimen patterns which were empty, i.e. they did not have data on seeds and pods. This is what happens in nature, as in different seasons the seeds and pods are not always present (or observed or collected) on the tree specimens. Thus the test data set consisted of only 688 unseen seed and pod patterns.

Figure 6-23 presents the Seed and PodSOM model 23 and the associated test set.



(a) **(b)**

Figure 6-23 : Seed and PodSOM Model 23 with Associated Test Set

The map presented in Figure 6-23 (a) is the trained PodSOM map 23, and Figure 6-23 (b) presents the same map after 688 test specimens were associated with it.

Table 6-21 lists the errors which occurred when the seed and pod test set was presented to each of the 30 Seed and PodSOM models. The average error rate for these models was 10.16%. This error rate is relatively high compared with the error rates obtained for TreeSOM and ThornSOM. However, it must be remembered that the seed and pod test set only had an average of 4.23 attribute values per specimen. When this low number of attribute values is taken into consideration the error rate is surprisingly good.

Table 6-21 : Seed and PodSOM Test Results

Trained Map X	Total Errors	Total Correct	Trained Map X	Total Errors	Total Correct
1	68	620	16	70	618
2	72	616	17	68	620
3	68	620	18	68	620
4	68	620	19	68	620
5	70	618	20	71	617
6	70	618	21	66	622
7	69	619	22	68	620
8	70	618	23	92	596
9	72	616	24	69	619
10	72	616	25	69	619
11	66	622	26	68	620
12	67	621	27	72	616
13	67	621	28	68	620
14	70	618	29	69	619
15	69	619	30	74	614
Total Error					**2098**
Error Rate					**10.16%**
Correct Rate					**89.84%**

Table 6-22 tabulates the results of presenting the test set to PodSOM Model 23. Only the results for clusters which had errors are presented.

Table 6-22 : Seed and PodSOM Model 23 Test Error Results

Cluster No.	ID of Cluster	No. of Tests per Cluster	Type of Error	No. of Errors	Errors in Cluster
5	**sw**	**38**	**ng** as **sw**	**7**	**7**
6	**bu**	**15**	**at** as **bu** **to** as **bu** **ca** as **bu**	**1** **7** **1**	**9**
14	**gr**	**74**	**at** as **gr** **bo** as **gr** **br** as **gr** **ge** as **gr** **lu** as **gr** **nl** as **gr**	**8** **9** **7** **8** **9** **7**	**48**
19	**ca**	**50**	**bu** as **ca** **at** as **ca**	**21** **1**	**22**
22	**nl**	**24**	**ca** as **nl**	**1**	**1**
23	**ko**	**36**	**si** as **ko**	**5**	**5**
Total Errors					**92**
Error Rate					**13.37%**
Correct Rate					**86.83%**

6.4.4 Statistical Analysis of the Seed and PodSOM Model Test Results

For each of the tests performed, a confusion matrix was drawn up and a metrics table was produced. This table is shown in Table 6-23. The (TP, FP) point for *A. ataxacantha* species (0, 0.687778) is the point furthest from the perfect classifier point (0, 1). The (TP, FP) point for *A. luederitzii* species (0, 0.695833) is the next furthest point from the point (0, 1). These results occurred because these two species were sometimes misidentified as other species. Conversely, the FP rate for *A. grandicornuta* was 0.020141, which was relatively high because other species were misidentified as this species.

Table 6-23 : Fraction Metrics for Seed and PodSOM Species

Species	TP	FP	Acc	Prec	Sp
at	0.687778	0.000000	0.986822	1.000000	1.000000
bo	0.756250	0.005691	0.974128	0.924324	0.991463
br	0.758621	0.000000	0.989826	1.000000	1.000000
bu	0.902469	0.012405	0.985368	0.739286	0.987342
ca	0.997778	0.005235	0.995833	0.906924	0.994985
da	1.000000	0.000456	0.999225	0.989449	0.999493
ge	0.884444	0.043617	0.967878	0.643452	0.960740
gr	0.866667	0.020141	0.969477	0.804517	0.978248
ka	1.000000	0.000000	1.000000	1.000000	1.000000
ko	0.997849	0.007610	0.992684	0.860847	0.992390
kr	1.000000	0.000000	1.000000	1.000000	1.000000
lu	0.695833	0.000000	0.985998	1.000000	1.000000
mo	1.000000	0.000000	1.000000	1.000000	1.000000
na	1.000000	0.000169	0.999952	0.997917	0.999899
ng	0.766667	0.000000	0.989826	1.000000	1.000000
nl	0.766667	0.000068	0.989826	0.998611	0.999949
ro	0.988889	0.000000	0.999225	1.000000	1.000000
sc	1.000000	0.000000	1.000000	1.000000	1.000000
se	1.000000	0.000000	1.000000	1.000000	1.000000
si	0.821429	0.000000	0.992733	1.000000	1.000000
sw	1.000000	0.010654	0.989826	0.815789	0.989346
to	0.766667	0.000000	0.989826	1.000000	1.000000
xa	1.000000	0.000000	1.000000	1.000000	1.000000
Average	**0.898174**	**0.004611**	**0.991237**	**0.942657**	**0.995385**

The TP and FP rates from Table 6-23 were used to produce a graph of the Seed and PodSOM ROC space which is displayed in Figure 6-24. This graph shows that the Seed and PodSOM models were able to identify many test specimens even though there were problems in the identification of some species. The point (0.687778, 0) for *A. ataxacantha* is the further point from the perfect classifier point (0, 1). *A. luederitzii* (0.687778, 0) is the next furthest point from the point (0, 1). These results highlight that these two species were sometimes identified as other species. Despite this, the Seed and PodSOM test results demonstrate that the models were able to identify some species correctly even though there were many missing values.

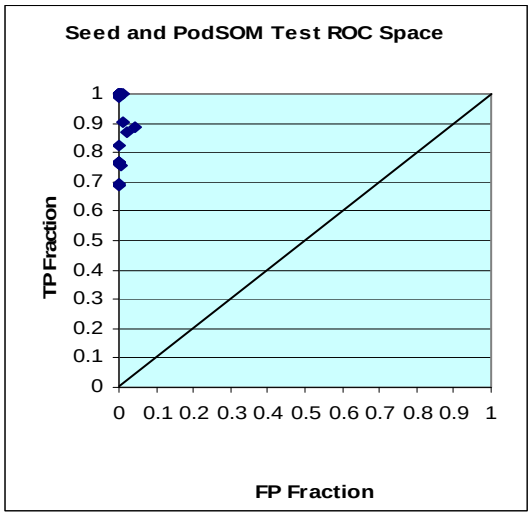


Figure 6-24 : Seed and PodSOM Test ROC Space

The LeafSOM model is discussed in the next section.

6.5 The LeafSOM Models

The leaf data set consisted of 920 patterns with up to 40 characteristics. The 30 sets of training data were presented to the SOM software as described in Chapter 5 using 162 neurons.

6.5.1 Evaluation of the Trained LeafSOM Models

The LeafSOM model obtained using the leaf data set 21 is shown in Figure 6-25.

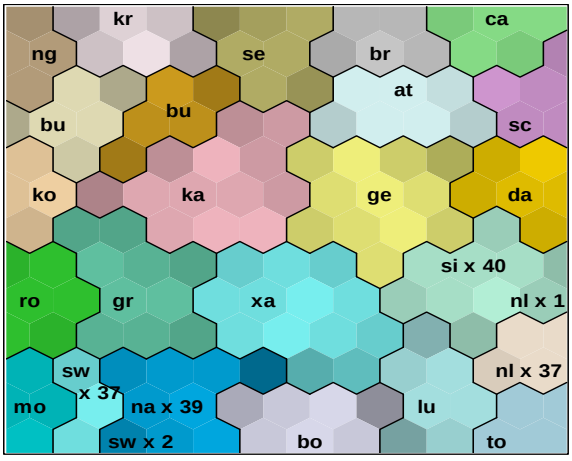


Figure 6-25 : LeafSOM Model 21

Several errors are shown on the map in Figure 6-25. One *A. nilotica* was identified as *A. sieberiana* as shown in the cluster on the right edge, just below centre. Another error occurred and is shown in the cluster at the bottom edge of the figure, third cluster from the left. Here two *A. swazica* were identified as *A. natalitia*.

What is interesting to note in the map displayed in Figure 6-25 is that although only leaf attributes were used in the training data set, it can be seen that LeafSOM has clustered all the hooked-thorns species (i.e. those with no straight thorns) in the top half of the map. Going from left to right and down the map these species are: *A. nigrescens*, *A. kraussiana*, *A. senegal*, *A. brevispica*, *A. caffra*, *A. burke*i, *A. ataxacantha* and *A. schweinfurthii*. The species *A. tortilis* and *A. luederitzii* which both have some hooked thorns and some straight/straightish thorns are clustered next to each other in the bottom right corner of the map.

Some of the component maps that were output using data set 21 are shown in Figure 6-26.

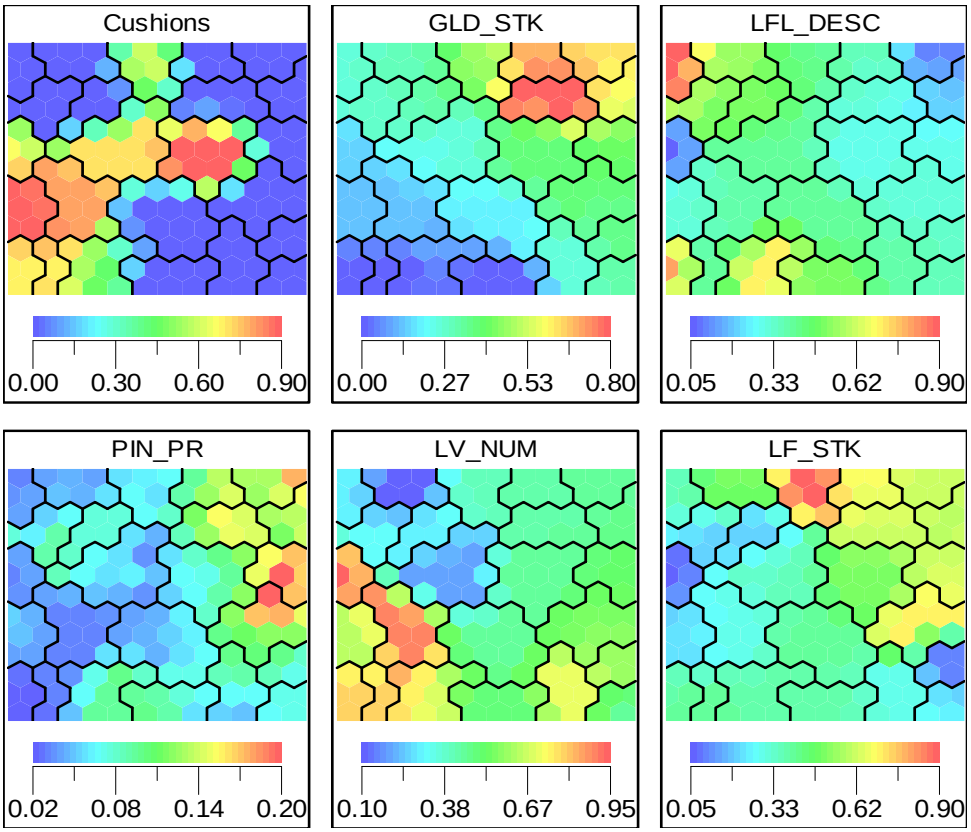
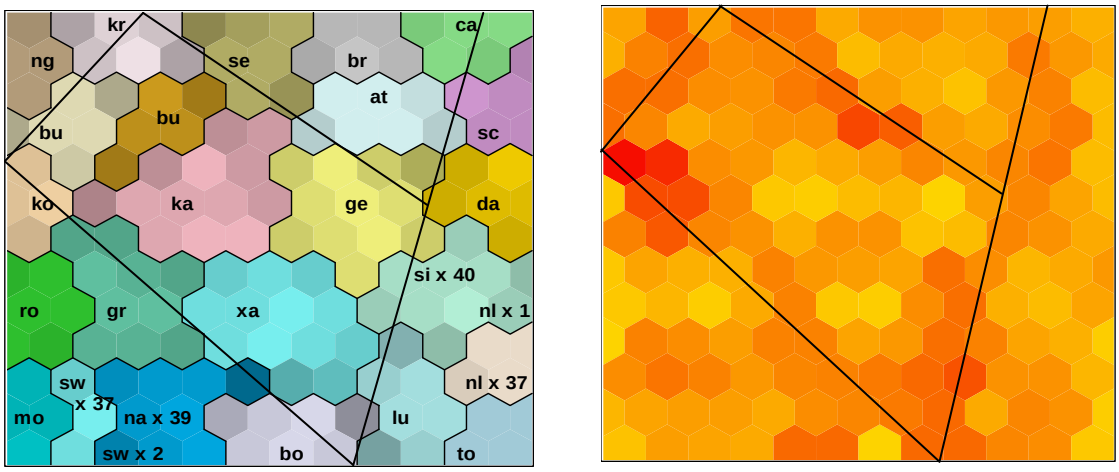


Figure 6-26 : LeafSOM Component Maps

These maps in Figure 6-26 show clearly some of the attributes for the different species. For instance, the Cushions component map clearly shows that *A. robusta*, *A. gerrardii* and *A. grandicornuta* all have high values for the cushion attribute.

Another example of similar species which are depicted in a component map is the species which have stalked glands which are shown grouped in the top right-hand corner of the GLD_STK component map. Similarly, the species which have numerous pinna pairs are shown as a distinct group at the right side and top of the PIN_PR component map. While the species with numerous leaves are shown in the bottom, left-hand corner of the LV_NUM component map.

The U-Matrix for the LeafSOM model 21 is shown in Figure 6-27 (b), and the trained LeafSOM is shown in Figure 6-27 (a). If the map in Figure 6-27 (b) is analyzed in conjunction with the maps shown in Figure 6-26 then certain traits become evident. Lines have been superimposed on the maps for reference purposes. From Figure 6-27 (b) some of the boundaries between the different species can be deduced. For example, the groups at the bottom left of Figure 6-27 (b) seem to be influenced by the LV_NUM attribute, and these groups consist of the species with larger numbers of leaves. The top centre group bears a correlation to the species with leaf stalks (LF_STK). The top left group seems to be influenced by the component leaflet description (LFT_DESC). The right-most group of clusters seems to be correlated to the number of pinna pairs (PIN_PR) attribute.



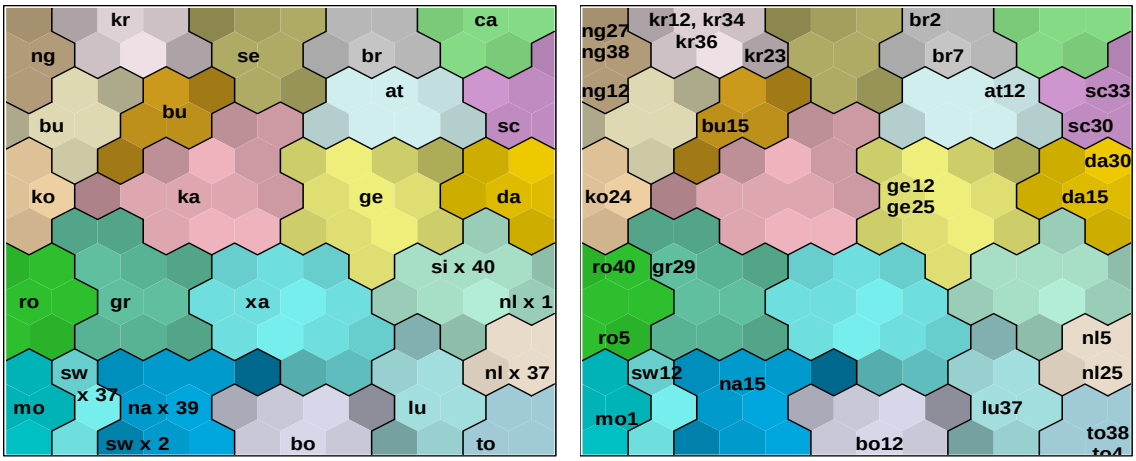
(a) **(b)**

Figure 6-27 : U-Matrix representation of LeafSOM Model 21

6.5.2 Evaluation of the LeafSOM Model Verification Results

The verification data sets were used to validate the LeafSOM models as was described in Chapter 5. Figure 6-28 shows the results for one of these verification tests. In this figure the leaf verification set 21 has been mapped to the LeafSOM model 21.

The results of the verification test in Figure 6-28 (b) show that the 30 verification patterns were all mapped to their correct clusters. The LeafSOM model 21 is presented in Figure 6-28 (a) for comparison purposes.

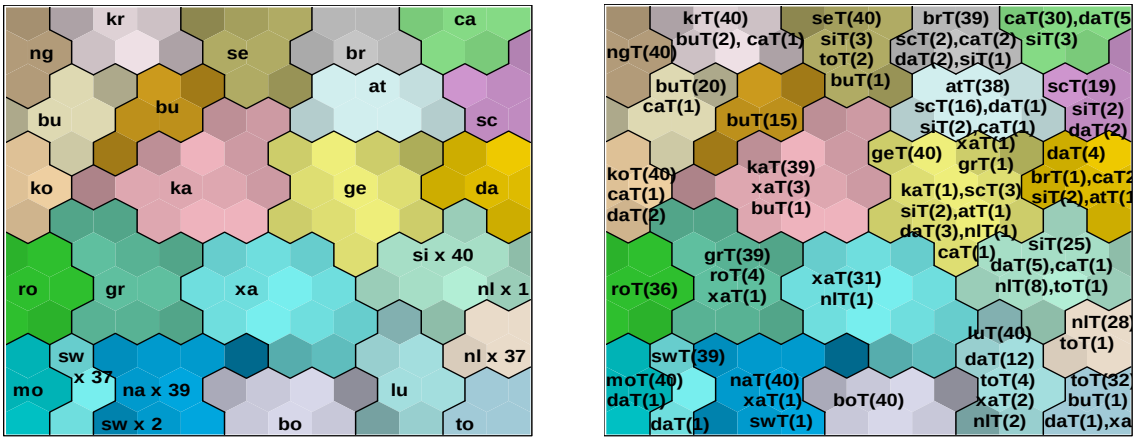


(a) **(b)**

Figure 6-28 : LeafSOM Model with Associated Verification Set 21

6.5.3 Evaluation of the LeafSOM Model Test Results

The leaf test data set consisting of 918 unseen patterns was associated with the LeafSOM models. The results obtained for model 21 are presented in Figure 6-29.



(a) **(b)**

Figure 6-29 : LeafSOM Model 21 with Associated Test Set

The errors that occurred when the leaf test data set was associated with each of the 30 LeafSOM models are tabulated in Table 6-24. From this table it can be seen that the average error rate is 12.64%. Considering that the leaf data set consisted of a possible 40 attributes these results were not as good as one expected. However, it must be noted that there was on average 5.55 attribute values per test specimen in the leaf test data set. When taking this into consideration the error rate was relatively good.

The LeafSOM models were unable to identify the *A. davyi* species satisfactory. Only four *A. davyi* specimens were correctly identified by LeafSOM model 21. This can be seen in Figure 6-29 (b) in the cluster on the right edge, 3rd cluster from the top.

Table 6-24 : LeafSOM Test Results

Trained Map X	Total Errors	Total Correct	Trained Map X	Total Errors	Total Correct
1	115	803	16	116	802
2	115	803	17	105	813
3	107	811	18	102	816
4	112	806	19	159	759
5	111	807	20	126	792
6	118	800	21	124	794
7	114	804	22	112	806
8	116	802	23	110	808
9	111	807	24	114	804
10	123	795	25	119	799
11	113	805	26	124	794
12	122	796	27	111	807
13	121	797	28	115	803
14	120	798	29	102	816
15	120	798	30	109	809
Total Errors					3480
Error Rate					12.64%
Correct Rate					87.36%

The errors that were shown in Figure 6-29 (b) are tabulated in Table 6-25. From cluster 14 in this table it can be seen that although the cluster was identified as *A. davyi* only four of the ten specimens grouped on this cluster were *A. davyi*.

Table 6-25 : LeafSOM Model 21 Test Error Results

Cluster No.	ID of Cluster	No. of Tests per Cluster	Type of Error	No. of Errors	Errors in Cluster
1	**xa**	32	**nl** as **xa**	1	1
2	**ka**	43	**xa** as **ka** **bu** as **ka**	3 1	4
3	**ge**	54	**ka** as **ge** **sc** as **ge** **si** as **ge** **at** as **ge** **da** as **ge** **nl** as **ge** **ca** as **ge** **xa** as **ge** **gr** as **ge**	1 3 2 1 3 1 1 1 1	14
4	**gr**	44	**ro** as **gr** **xa** as **gr**	4 1	5
5	**si**	40	**da** as **si** **ca** as **si** **nl** as **si** **to** as **si**	5 1 8 1	15
6	**na**	42	**xa** as **na** **sw** as **na**	1 1	2
8	**at**	58	**sc** as **at** **da** as **at** **si** as **at** **ca** as **at**	16 1 2 1	20
9	**lu**	60	**da** as **lu** **to** as **lu** **xa** as **lu** **nl** as **lu**	12 4 2 2	20
10	**se**	46	**si** as **se** **to** as **se** **bu** as **se**	3 2 1	6
11	**kr**	43	**bu** as **kr** **ca** as **kr**	2 1	3
13	**bu**	21	**ca** as **bu**		1
14	**da**	10	**br** as **da** **ca** as **da** **si** as **da** **at** as **da**	1 2 2 1	6
16	**ca**	38	**da** as **ca** **si** as **ca**	5 3	8
17	**sc**	22	**si** as **sc** **da** as **sc**	2 1	3
18	**br**	46	**sc** as **br** **ca** as **br** **da** as **br** **si** as **br**	2 2 2 1	7
19	**mo**	41	**da** as **mo**	1	1
21	**nl**	29	**to** as **nl**	1	1

22	**ko**	43	**ca** as **ko** **da** as **ko**	1 2	3
23	**to**	35	**bu** as **to** **da** as **to** **xa** as **to**	1 1 1	3
24	**sw**	40	**da** as **sw**	1	1
Total Errors					**124**
Error Rate					**13.51%**
Correct Rate					**86.49%**

(Table 6-25 continued)

6.5.4 Statistical Analysis of the LeafSOM Model Test Results

The rates in Table 6-26 were drawn from the confusion matrices created to display the results of the LeafSOM tests. In this table, the TP rate for *A. davyi* is 0.2, which is extremely low. This low rate resulted because very few of the *A. davyi* specimens were correctly associated with the *A. davyi* cluster (also shown in Table 6-25, cluster 14).

Table 6-26 shows that the TP rate for *A. schweinfurthii* (0.521667) was also low. This result can be confirmed by looking at cluster 17 of Table 6-25, which shows that for map 21 only 22 *A. schweinfurthii* test specimens were correctly identified.

The TP rate for *A. sieberiana* is 0.524167, which is low and resulted because this species was frequently identified as other species. From Figure 6-29 it can be seen that map 21 only had 25 of the test *A. sieberiana* specimens correctly identified.

Also of note in Table 6-26 are the rates for *A. nigrescens*. These show that no errors were made in identifying this species. This result is not surprising as the leaves of *A. nigrescens* are very distinctive and distinguish the species from other KZN *Acacia* species (except for some *A. burkei* which can also have large leaflets).

Table 6-26 : Fraction Metrics for LeafSOM Species

Species	TP	FP	Acc	Prec	Sp
at	0.965833	0.025209	0.974401	0.644432	0.974791
bo	1.000000	0.001025	0.999020	0.978353	0.998975
br	0.954167	0.011883	0.986638	0.762696	0.988117
bu	0.883333	0.000645	0.994299	0.984515	0.999355
ca	0.730833	0.006112	0.982426	0.848845	0.993888
da	0.200000	0.006780	0.960385	0.553142	0.993220
ge	0.984167	0.013174	0.986710	0.773739	0.986826
gr	0.945000	0.003834	0.993936	0.918823	0.996166
ka	0.974167	0.003986	0.995062	0.918964	0.996014
ko	1.000000	0.003037	0.997095	0.938250	0.996963
kr	1.000000	0.004708	0.995497	0.908028	0.995292
lu	0.995833	0.013743	0.986674	0.769423	0.986257
mo	1.000000	0.001746	0.998330	0.963517	0.998254
na	1.000000	0.000835	0.999201	0.982382	0.999165
ng	1.000000	0.000000	1.000000	1.000000	1.000000
nl	0.810000	0.004480	0.987436	0.895555	0.995520
ro	0.915000	0.001898	0.994481	0.957915	0.998102
sc	0.521667	0.002923	0.976362	0.895737	0.997077
se	1.000000	0.007365	0.992956	0.862544	0.992635
si	0.524167	0.011314	0.968446	0.687485	0.988686
sw	0.994167	0.002354	0.997495	0.951302	0.997646
to	0.854167	0.003531	0.990269	0.918729	0.996469
xa	0.806667	0.001557	0.990087	0.959525	0.998443
Average	**0.872138**	**0.005745**	**0.989009**	**0.872778**	**0.994255**

Figure 6-30 displays the FP and TP rates plotted on a graph in ROC space. These rates are obtained from Table 6-26. The poor identification rates for *A. davyi, A. schweinfurthii* and *A. sieberiana* are again demonstrated on this graph.

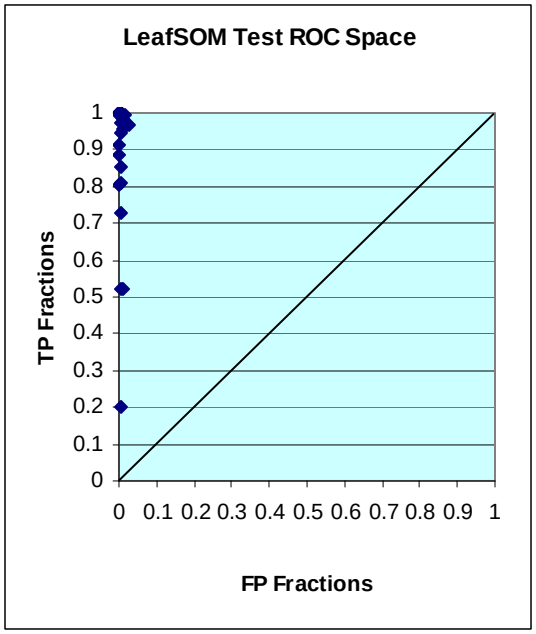


Figure 6-30 : LeafSOM Test ROC Space

In the next section the C5 and CN2 results are discussed.

6.6 C5 and CN2 Results

The *Acacia* data sets were tested using the C5 and CN2 algorithms. However, the results from these tests were not meaningful. The C5 algorithm chose the shortest route to determining the identity of the species, so the training results showed 100% success, as did the TreeSOM results. However, as soon as attribute values were missing from the test data the C5 algorithm was unable to produce meaningful results.

The extracted rules for CN2 include default cases, which meant that if a characteristic was absent then the specimen could fall through to the default argument and be incorrectly identified as a totally different species.

Hence it was concluded that the use of these algorithms with the data sets used in this research was totally inappropriate. The C5 and CN2 algorithms could not identify the test data specimens meaningfully because they could not handle the large number of missing values. The C5 and CN2 results did not make sense and it was therefore not possible to make a comparison between these results and the results obtained

from the SOM as had been intended at the start of this investigation. Consequently these results are not presented here.

6.7 Summary of SOM Results

A summary of the test results obtained using SOM is presented in Table 6-27

Table 6-27 : Summary of SOM Test Data Set Results

SOM Model	Size[1] of Training Data Set	No. of Neurons used for Training	Max.[2] No. of Attributes	Size[3] of Test Data Set	Average[4] No. of Attribute Values	Correct Rate
TreeSOM	890/870	201	127	920	19.52	99.996%
Habit & ThornSOM	890/870	103	43	920	8.61	96.56%
FlowerSOM	890/870	93	19	534	4.62	78.07%
Seed & PodSOM	890/870	587	25	688	4.23	89.86%
LeafSOM	890/870	162	40	918	5.55	87.36%

1 For training the size of each data set was either 890 for the first set, or 870 for each of the other 29 data sets.

2 This is the maximum number of attributes that were used for describing the data in the relevant data set.

3 In the test data sets some test data specimens had no data on flowers, seed and pods, or on leaves, consequently the size of the set was smaller.

4 This is the average number of attribute values per test specimen.

From Table 6-27 it can be seen that if there is an average of 19.52 attribute values present in the test data set per specimen, as there is for testing TreeSOM, then identification results are very good. The correct rate for identification using TreeSOM was 99.996%.

For Habit and ThornSOM the average number of attribute values per specimen present in the test data set was 8.61 and the correct rate for identification was 96.56%.

Any correct rate over 96% is statistically significant, so these identification results are also good.

FlowerSOM had an average number of attributes values of 4.64 per specimen in the test data set, and had a correct rate for identification of 78.07%. This rate is too low and would need to be improved. However, it is felt that the reduction in the correct rate was affected by two factors: the low average number of attribute values per specimen present in the test data set, and the low number of attributes available for describing the training data (there were at most 19 attributes available for describing the flower data sets).

Similarly, Seed and PodSOM had an average number of attributes values of 4.23 per specimen in the test data set, and had a correct rate for identification of 89.84%. These identification results were higher than for FlowerSOM, and it is felt that this improvement was probably due to the fact that PodSOM had up to 25 attributes available for describing the seed and pod training data. Again, it is thought that if the number of attributes for describing the training data, as well as the average number of attribute values per test specimen, were increased the identification correct rate would be higher. The number of neurons needed to get these results was very high being over half the total number of training data patterns presented to the networks.

LeafSOM had up to 40 attributes for describing the leaf training data, but there was only an average of 5.55 attribute values per specimen used to describe the leaf test data set. The correct identification rate for LeafSOM was 87.36%. This rate was disappointing in view of the relatively high number of attributes available for describing the leaf data. However, the relatively low correct rate is again thought to be as a result of the low average number of attribute values per specimen used in the test data set.

What is also worth noting is that none of the SOM models had difficulty identifying and subdividing what was until recently known as the *A. karroo* complex which included the species *A. karroo, A .montana, A. natalitia* and *A. kosiensis* [42]. The newly described species, *A. montana, A. natalitia* and *A. kosiensis*, were previously included under the umbrella of *A. karroo*.

Overall the results of the SOMs were considered good, and the TreeSOM models showed what an excellent identification technique the SOM has proved to be. Therefore the SOM promises to be of immense use in the identification and classification of biodiversity.

6.8 Conclusion

The maps produced by using the whole, habit and thorn, flower, seed and pod, and leaf data sets were presented and discussed in turn in this chapter. First, the training data sets used to produce the SOM maps were discussed and the results analyzed. Next the results of verifying the maps using a 30-fold cross-validation method were presented and discussed. Thereafter, the results of testing the maps to see if they could identify unseen *Acacia* material from their data sets were presented. The results of the identification test sets were analyzed, and confusion matrices and/or ROC space graphs used to rate the performance of the models.

The test results obtained from presenting the unseen data to the C5 and CN2 algorithms were not meaningful and consequently were not presented here. The C5 and CN2 techniques have been used in the other research for identification (see references given on pp42-43) and for this reason were selected to be used for comparison with the SOM. However, the C5 and CN2 methods were clearly unable to identify the biological data used in this research and it is believed that this was because these methods could not cope with the large number of missing values in the unseen data sets.

In the next chapter the final conclusions will be presented together with recommendations for future work.

Chapter 7

Future Development and Conclusion

The primary objective of this thesis was to evaluate the suitability of the SOM for use as an identification tool for biological species. The first step towards this was to discuss traditional identification methods that have been used so as to develop an understanding of the shortcomings of procedures in current use. These traditional methods were discussed in Chapter 2, and the need for improved identification techniques was highlighted.

In Chapter 3, various attempts to utilise AI techniques to identify biological specimens were reviewed, but it is apparent that the full potential of modern AI technology has not been achieved. This has been part of the motivation for conducting this experimental research project which uses Viscovery® SOMine software to attempt to find a more effective identification tool.

Chapter 4 described the SOM algorithm in detail. The basic theory and concepts behind this algorithm were presented together with the problems which can occur with the technique. Some variants and extensions of the SOM were also described.

Chapter 5 outlined the research design that was followed for this thesis. The choice of KZN *Acacia* as the application field and the reasons for their selection were given. Any limitations imposed on the extent of the data set were also described. The choice of Viscovery® SOMine software was also motivated. The sampling process was described and the pre-processing and the encoding of the data were explained. Finally, the presentation of data to the software and the verification and testing processes were described.

The models obtained from performing the experiments described in Chapter 5 were presented in Chapter 6. These models were discussed and analyzed. It was concluded that the SOM had been applied very successfully to the *Acacia* data sets. The models were able to predict the species of the test specimens with accuracy ranging from 78% to 99%.

The experiments conducted with the SOM algorithm had two distinct objectives.

1. The teaching or inductive phase. The first objective was to examine whether the SOM was able to structure and correlate the data and identify statistically distinct biological patterns contained within that data. This the SOM proved well able to do; as well as being able to recognize distinctive patterns within species. Additionally, in performing this function the SOM confirmed that the conversion of the botanical data from a descriptive form into a normalized numerical form was effective.
2. The testing or deductive phase. The second objective of the experiments was to determine the accuracy with which the SOM would be able to identify "unknown" biological specimens. This accuracy proved to be high (up to 99%) when an extensive range of attribute values was used to describe the data set for developing a model. For example, the TreeSOM models, which were formed from large data sets, were able to identify test data accurately. However, accuracy was reduced when only a few attributes were used to form the models and/or to describe the test data set. For example, the FlowerSOM models, which were formed from limited data sets, had a correct identification rate of only 78.07%.

At the start of this thesis it was hoped that the SOM would be able to help with identifying biological specimens. The TreeSOM models have successfully demonstrated that they are able to differentiate the different *Acacia* species occurring in KZN. These models, created with training data sets which had up to 127 attribute values, obtained an average correct rate of 99.996% with test data sets which had an average of 19.52 attribute values per test specimen.

The Habit and ThornSOMs were also used successfully to identify unseen habit and thorn test data specimens. The average correct rate for these models was 96.56%. Even though the number of attributes used for training these models was at most 43 and the average number of attribute values in the test data set was 8.61 per test specimen.

The Seed and PodSOM models were trained with data sets which had at most 25 characteristics and obtained an average correct rate of 89.84% when presented with unseen data. In view of the fact that the average number of attribute values per test specimen was 4.23 the correct rate for Seed and PodSOM was surprisingly good.

The LeafSOM models were trained using up to 40 attributes. The average correct rate achieved when these models were presented with unseen test specimens was 87.36%. The lower rate of success of LeafSOM was disappointing because of the relatively high number of attributes were available for describing the leaf training data. However, it must be noted that the average number of attribute values used for the leaf test specimens was only 5.55 per specimen. This paucity of test data probably contributed to the correct rate being lower than expected.

The FlowerSOM correct rate was 78.07%, the lowest rate obtained during the experiments. However, the number of attributes used for training the FlowerSOM was at most 19 and because of the correlation between some of the attributes the FlowerSOM models were not sufficiently diagnostic. Additionally, the flower test specimens themselves were not well defined as the average number of attribute values per test specimen was only 4.62. Taking these last two factors into account the correct rate of 78.07% is understandable, and it is felt that the lower identification rate was more to do with the sparse data sets, rather than with a shortcoming of the technique utilized.

Although the Habit and ThornSOM models were the only subset models to obtain an average correct rate over 96%, all the subset models show promise of being able to identify the species.

Also, at the start of this thesis it was hoped that the C5 and CN2 algorithms could be used to provide meaningful results for comparison with those obtained by using the SOM for identifying unseen biological specimens. The selection of the C5 and CN2 algorithms for identification of incomplete biological data sets turned out to be a poor choice for comparison purposes. However, the SOM results showed that the SOM technique was suitable for biological identification and was able to succeed where other techniques had failed and that the SOM algorithm’s superiority for identification was clearly established.

7.1 Future Work

Some of the relationships between species and the correlations between characteristics, as were illustrated by the SOM models, have no obvious or known biological explanation. It is felt that an important continuation of the work started in this thesis would be to investigate the taxonomic relationships suggested by the TreeSOMs. These may have evolutionary or genetic significance which could be tested. For example, in Figure 6-5 TreeSOM's clustering of the species splits the KZN *Acacia* species into four major groups (at the level where 4 clusters were requested). At this level TreeSOM splits the species into two subgroups for the hooked thorn species and two subgroups for the species with some straight thorns. It would be interesting to see if there are valid phylogenetic reasons behind this split into four groups.

Also, some of the associations between species are well documented and based on morphological evidence [168, 179, 180], but other relationships depicted by TreeSOM appear not to be documented. The authors of the above references discuss similarities between *A. robusta, A. grandicornuta* and *A. gerrardii,* and between *A. luederitzii* and *A. tortilis*, but not between *A. nilotica* and *A. sieberiana*. It would be very interesting to see if other evidence (morphological or genetic) supports the relationship suggested by TreeSOM between *A. nilotica* and *A. sieberiana*.

The unsupervised SOM techniques have been demonstrated on a botanical data set and its four subsets. The subset models were not as successful as the whole model in identifying unknown specimens. The apparent loss of accuracy demonstrated by the FlowerSOM needs to be investigated further. It is felt that the current flower data set, for instance, only offers two different types of data on what are very important characteristics: flowers are either white or yellow, while inflorescences are either capitate or spicate. These characteristics are highly correlated, with all the spicate inflorescences being white. This means that the data are easy to separate into two groups, but these important flower characteristics on their own are not sufficient to differentiate species. More data (i.e. more characteristics) need to be collected to make the differentiation between species and attributes more successful. In addition,

because of time restraints and lack of resources, important reproductive characteristics were omitted from the data set and should be included in future work.

The experimental work in this thesis was performed on data collected on KZN *Acacia* species. The SOM results would change if data on *Acacia* species from the rest of southern Africa were included. *Acacia* data from species found in South Africa or even southern Africa should be extracted and used to see what relationships the SOM could detect with these extended data sets. If this wider range of species were included then some of the relationships detected by SOM in the KZN data sets might also make more sense. For example, the relationship between *A. nilotica* and *A. sieberiana* might change or be clarified.

One of the main strengths of SOMs is the successful preservation of neighbourhood relations. This strength should be further utilized to discover structures within other botanical and biological species. It would also be interesting to see if the SOM could identify infraspecific taxa such as subspecies and varieties. Again, this would be an interesting future development of this research.

Another extension of this research project would be to use other species that are more difficult to identify, and see if SOM models can be developed which could help with the identification of problem groups. The SOM could be used in conjunction with DNA sequence data to trace evolutionary histories and to determine whether species in a genus originate from a common ancestor.

From the foregoing it can be seen that the SOM promises to be an exciting tool for biological identification, and with further work could even help solve the backlog dilemma regarding the barcoding of biodiversity.

7.2 Conclusion

AI technology offers non-experts the advantage of identification of botanical material, particularly when experts are not available. This is of special relevance in Africa where the demand for experts far exceeds what is currently available.

When identifying species, botanical data by their nature are often sparse. A technique that can succeed in getting a species name when information is missing is

particularly valuable. In this respect, the field guides that are most commonly utilized fail dismally in the hands of a layperson. Thus the SOM technique, which is a well known technique which has been tried and tested for its ability to handle sparse data sets, is extremely important.

In this thesis the use of SOM as a tool for analysis of data sets based on 23 KZN *Acacia* species has been demonstrated. The results have been shown to be consistent with existing knowledge of the species, and new relationships have been demonstrated within the data set.

It is therefore believed that the SOM is indeed a very real and viable alternative tool for the identification of biological specimens, and because of its capability of handling fuzzy and sparsely populated data sets it must be recognized as an essential tool to help with the identification and classification of biodiversity.

Bibliography

[1] "SOM Toolbox for Matlab: Software implementations of the SOM," 2001, http://www.cis.hut.fi/projects/somtoolbox/links/somsoftware.shtml.

[2] "Viscovery Software GmbH - Viscovery SOMine 5.0," 2008, http://www.viscovery.net/somine.

[3] H. A. Abbass, M. Towsey, and G. Finn, "C_Net: Generating Multivariate Decision Trees From Artificial Neural Networks Using C5," 1999, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.6514.

[4] A. Abraham, "Rule-based Expert System," in *Handbook of Measuring System Design*, P. H. Sydenham and R. Thorn, Eds.: John Wiley & Sons, Ltd, 2005.

[5] D. Alahakoon and S. K. Halgamuge, "Knowledge Discovery with Supervised and Unsupervised Self Evolving Neural Networks," in *Proceedings of 5th International Conference on Soft Computing and Information/Intelligent Systems*, Fukuoka, Japan, 1998, pp. 907-910.

[6] D. Alahakoon, S. K. Halgamuge, and B. Sirinivasan, "A Self Growing Cluster Development Approach to Data Mining " in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, San Diego, USA, 1998, pp. 2901-2906.

[7] D. Alahakoon, S. K. Halgamuge, and B. Sirinivasan, "A Structure Adapting Feature Map for Optimal Cluster Representation," in *Proceedings of The 5th International Conference on Neural Information Processing (ICONIP 98)*, Kitakyushu, Japan, 1998, pp. 809-812.

[8] D. Alahakoon, S. K. Halgamuge, and B. Strinivasan, "Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery," in *Foundations of Computational Intelligence: Volume 4: Bio-Inspired Data Mining*, A. Abraham, A. E. Hassanien, and A. Ponce de Leon F. de Carvalho, Eds. Berlin, Heidelberg: Springer-Verlag, 2005.

[9] N. Allsopp, "*Acacia* Name Change," *Grassroots: Newsletter of the Grassland Society of Southern Africa,* vol. 5, p. 8, November 2005.

[10] A. Ananthaswamy, "Earth faces sixth mass extinction," in *NewScientist*, 2004, http://www.newscientist.com/article/dn4797-earth-faces-sixth-mass-extinction.html.

[11] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2007, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[12] W. D. Atkinson and A. Gammerman, "An Application of Expert Systems Technology to Biological Identification," *Taxon,* vol. 36, pp. 705-714, November 1987.

[13] M. Attik, L. Bougrain, and F. Alexandre, "Self-organizing Map Initialization," *Lecture notes in computer science,* vol. 3696, pp. 357-362, 2005.

[14] W.-H. Au, K. C. C. Chan, A. K. C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 2, pp. 83-101, April - June 2005.

[15] H.-U. Bauer and T. Villmann, "Growing a Hypercubical Output Space in a Self-Organizing Feature Map," International Computer Science Institute, Berkeley July 1995.

[16] K. S. Bawa, "Cataloguing life in India: the taxonomic imperative," *Current Science,* vol. 98, pp. 151-153, 25 January 2010.

[17] J. Bernatavičienė, G. Dzemyda, o. Kurasova, and V. Maecinkevičius, "Optimal decisions in combining the SOM with nonlinear projection methods," *European Journal of Operational Research,* vol. 173, pp. 729-745, 2006.

[18] M. Bishop, M. Svensen, and C. K. I. Williams, "GTM: Generative topographic mapping," *Neural Computation,* vol. 10, pp. 215-234, 1998.

[19] J. Blackmore and R. Miikkulainen, "Incremental Grid Growing: Encoding High-Demensional Structure into a Two-Dimensional Feature Map," in *IEEE International Conference on Neural Networks, ICNN'93*, Austin, 1993, pp. 450-455.

[20] F. Blayo and P. Demartines, "Data analysis: How to compare Kohonen neural networks to other techniques?," in *Proceedings of IWANN International Workshop on Artificial Neural Networks*, 1991, pp. 469-476.

[21] F. Boero, "Light after dark: the partnership for enhancing expertise in taxonomy," *Trends in Ecology & Evolution,* vol. 16, p. 266, 5 May 2001.

[22] R. Boswell, "Manual for CN2 version 6.1," The Turing Institute Limited, Manual, January 1990.

[23] S. Brosse, J. L. Giraudel, and S. Lek, "Utilisation of non-spervised neural networks and principal component analysis to study fish assemblages," *Ecological Modelling,* vol. 146, pp. 159-66, 1st December 2001.

[24] A. Browne, B. D. Hudson, D. C. Whitley, M. G. Ford, and P. Picton, "Biological data mining with neural networks: implementation and application of a flexible decision tree extraction algorithm to genomic problem domains," *Neurocomputing,* 2003.

[25] R. K. Brummitt, "World Geographical Scheme for Recording Plant Distributions," in *Plant Taxonomic Database Standards*. vol. 2: Working Group on Taxonomic Databases For Plant Sciences (TDWG), 2001, http://www.nhm.ac.uk/hosted_sites/tdwg/TDWG_geo2.pdf.

[26] R. K. Brummitt, "Report of the committee for spermatophyta: 55. Proposal 1584 on *Acacia*," *Taxon* vol. 53, pp. 826-829, 2004.

[27] B. G. Buchanan and E. H. Shortliffe, "Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project," AAAI Press 1984, http://www.aaaipress.org/Classic/Buchanan/buchanan.html.

[28] J. D. Carr, *The South African Acacias*: Conservation Press, 1976.

[29] R. Céréghino, Y.-S. Park, A. Compin, and S. Lek, "Predicting the species richness of aquatic insects in streams using a limited number of environmental variables," *Journal of the North American Benthological Society,* vol. 22, pp. 442-456, 2003.

[30] A. R. Chapman, "Directions for the structure of taxonomic descriptive data," Western Australian Herbarium Department of Conservation and Land Management 2001.

[31] W. W. L. Cheung, T. J. Pitcher, and D. Pauly, "A fuzzy logic expert system to estimate intrinsic extrinction vulnerabilities of marine fishes to fishing," *Biological Conservation,* vol. 124, pp. 97-111, 2005.

[32] T. W. S. Chow and S. Wu, "An Online Cellular Probabilistic Self-Organizing Map for Static and Dynamic Data Sets " *IEEE Transactions on Circuits and Systems,* vol. 51, pp. 732-747, 2004.

[33] S. Chu, J. DeRisi, M. B. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, "The Transcriptional Program of Sporulation in Budding Yeast," *Science,* vol. 282, pp. 699-705, 23 OCTOBER 1998.

[34] W. J. Clancey and R. Letsinger, "NEOMYCIN: Reconfiguring a Rule-Based Expert System for Application to Teaching," *IJCAI,* pp. 829-836, 1981.

[35] W. J. Clancey, "From GUIDON to NEOMYCIN and HERACLES in Twenty Short Lessons: ORN Final Report 1979-1985," August 1986.

[36] J. Y. Clark and K. Warwick, "Artificial Keys for Botanical Identification using a Multilayer Perceptron Neural Network (MLP)," *Artificial Intelligence Review,* vol. 12, pp. 95–115, 1998.

[37] J. Y. Clark, "Artificial neural networks for species identification by taxonomists," *BioSystems,* vol. 72, pp. 131–147, 2003.
[38] P. Clark and T. Niblett, "Induction in Noisy Domains," in *2nd European Machine Learning Conference (EWSL-87)*, Bled, Yugoslavia, 1987, pp. 11-30.
[39] P. Clark and T. Niblett, "The CN2 Induction Algorithm," *Machine Learning,* vol. 3, pp. 261-283, 1989.
[40] P. Clark and R. Boswell, "Rule Induction with CN2: Some Recent Improvements," in *Machine Learning - EWSL-91: European Working Session on Learning*, Porto, Portugal, 1991, pp. 151-163.
[41] P. E. Clark, "CN2 - Rule induction from examples," in *Peter Clark - Software*, http://www.cs.utexas.edu/users/pclark/software/.
[42] K. Coates Palgrave, *Trees of Southern Africa*, 3rd ed.: Struik, 2002.
[43] W. F. Contreras, E. G. Galindo, A. B. Morillas, and P. M. Lorenzo, "An application of expert systems to botanical taxonomy," *Expert Systems with Applications,* vol. 25, pp. 425-430, 2003.
[44] B. Coppin, *Artificial Intelligence Illuminated*, 1 ed.: Jones & Bartlett, 2004.
[45] M. Cottrell and M. Verleysen, "Advances in Self-Organizing Maps," *Neural Networks,* vol. 19, pp. 721–722, 2006.
[46] B. C. Craenen and A. E. Eiben, "Computational Intelligence," in *Encyclopedia of Life Support Sciences, EOLSS*: EOLSS Co. Ltd., http://www.cs.vu.nl/~gusz/papers/Comp-Intell-Craenen-Eiben.ps.
[47] M. J. Dallwitz, "Overview of the DELTA System," http://delta-intkey.com.
[48] M. J. Dallwitz, "A general system for coding taxonomic descriptions," *Taxon,* vol. 29, pp. 41–6, 1980.
[49] M. J. Dallwitz, "A comparison of matrix-based taxonomic identification systems with rule-based systems," in *Proceedings of IFAC Workshop on Expert Systems in Agriculture*, 1992, pp. 215–8.
[50] M. J. Dallwitz, "A comparison of interactive identification programs ", 2000, http://delta-intkey.com.
[51] M. J. Dallwitz, T. A. Paine, and E. J. Zurcher, "Principles of Interactive Keys," 2002, http://biodiversity.uno.edu/delta/.
[52] M. J. Dallwitz, T. A. Paine, and E. J. Zurcher, "User's guide to the DELTA System: a general system for processing taxonomic descriptions," 1993 onwards, http://delta-intkey.com.
[53] L. Davidson and B. Jeppe, *Acacias - A field guide to the identification of the species of Southern Africa*. Johannesburg: Centaur, 1981.
[54] R. Davis, "A DSS for diagnosis and therapy," in *Special Issue: Proceedings of a conference on Decision Support Systems*, Santa Clara, California, 1977, pp. 58-72.
[55] E. J. Dean, "Investigating the Potential Use of Computers and Fuzzy Expert Systems as Tools for the Identification of Biological Organisms," Durban Institute of Technology, Printing Department, Durban 2001.
[56] G. Deboeck, "Software Tools for Self-Organizing Maps," in *Visual Explorations in Finance with Self-Organizing Maps*, G. Deboeck and T. Kohonen, Eds. London: Springer-Verlag, 1998, pp. 179-194.
[57] G. Deboeck and T. Kohonen, *Visual Explorations in Finance with Self-Organizing Maps*. London: Springer-Verlag, 1998.
[58] R. DeSalle, M. G. Egan, and M. Siddall, "The unholy trinity: taxonomy, species delimitation and DNA barcoding," *Phil. Trans. R. Soc. B,* vol. 360, pp. 1905–1916, 2005.

[59] J. Durkin, "Application of Expert Systems in Science," *Ohio Journal of Science,* vol. 90, pp. 171-179, 6 September 1990.
[60] R. A. Dyer, *The Genera of Southern African Flowering Plants* vol. 1. Pretoria: Department of Agricultural Technical Services, 1975.
[61] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA : Genetics,* vol. 95, pp. 14863-14868, December 1998.
[62] M. B. Eisen and P. O. Brown, "DNA arrays for analysis of gene expression," *Methods Enzymol. ,* vol. 303, pp. 179-205, 1999.
[63] M. K. El-Najdawi and A. C. Stylianou, "Expert Support Systems: Integrating AI Technologies," *Communications of the ACM,* vol. 36, December 1993.
[64] A. P. Engelbrecht, S. E. Rouwhorst, and L. Schoeman, "A Building Block Approach to Genetic Programming for Rule Discovery," in *Data Mining: A Heuristic Approach* H. A. Abbass, R. A. Sarker, and C. S. Newton, Eds.: Idea Group Publishing, 2001, pp. 174-189.
[65] A. P. Engelbrecht, *Computational Intelligence: An Introduction*. Chichester: Wiley & Sons, 2002.
[66] W. Fajardo, E. Gibaja, and P. Moral, "G.R.E.E.N. An Expert System to identify Gymnosperms," 2004, http://www.ist.cmu.ac.th/intech/paper/InTech0284.pdf.
[67] L. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms and Applications* Englewood Cliffs, NJ: Prentice-Hall, 1994.
[68] E. A. Feigenbaum, "The art of artificial intelligence: I. Themes and case studies of knowledge engineering," Stanford University STAN-CS-77-621, August 1977.
[69] E. A. Feigenbaum, "Expert Systems: Principles and Practice," Stanford University 1992.
[70] E. A. Feigenbaum and B. G. Buchanan, "DENDRAL and Meta-DENDRAL: roots of knowledge systems and expert system applications," *Artificial Intelligence,* vol. 59, pp. 233-240, 1993.
[71] C. Fernández, E. Soria, J. D. Martín, and A. J. Serrano, "Neural Networks for animal science applications: Two case studies," *Expert Systems with Applications,* vol. 31, pp. 444-450, 2006.
[72] G. B. Fogel and D. W. Corne, "Computational intelligence in bioinformatics," *BioSystems,* vol. 73, pp. 1-4, 2003.
[73] I. France, A. W. G. Duller, G. A. T. Duller, and H. F. Lamb, "A new approach to automated pollen analysis," *Quatern. Sci. Rev.,* vol. 19, pp. 537-546, 2000.
[74] B. Fritzke, "Let it Grow - Self-Organizing Feature Maps with Problem Dependent Cell Structure," in *Proceedings of ICANN-91, International Conference on Artificial Neural Networks*, Espoo, Finland, 1991, pp. 403-408.
[75] B. Fritzke, "Growing Cell Structures - A Self-organizing Network for Unsupervised and Supervised Learning," International Computer Science Institute, Berkeley, California May 1993.
[76] B. Fritzke, "A Growing Neural Gas Network Learns Topologies," in *Advances In Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge MA: MIT Press, 1995.
[77] R. Froese, D. Pauly, and Editors, "FishBase." 2008, www.fishbase.org.
[78] E. L. G. Galindo, "Modelos de Representación del Conocimiento Para la Indentificatión Taxonómica Y Aplicaciones," in *Dpto. Ciencias de la Computación e Inteligencia Artificial*. vol. Ph. D. Granada: Universidad de Granada, 2004.
[79] K. J. Gaston and M. A. O'Neill, "Automated Species Identification: Why Not?," *Phil. Trans.: Biological Sciences,* vol. 359, pp. 655-667, April, 29 2004.

[80] I. D. Gauld, M. A. O'Neill, and K. J. Gaston, "Driving Miss Daisy: the performance of an automated insect identification system," in *Hymenoptera: Evolution, Biodiversity and Biological Control* A. D. Austin and M. Dowton, Eds. Canberra: CSIRO Publishing, 2000, pp. 303-312.

[81] M. Gerstein and R. Jansen, "The current excitement in bioinformatics—analysis of whole-genome expression data: how does it relate to protein structure and function? ," *Current Opinion in Structural Biology,* vol. 10, pp. 574-584, 1 October 2000.

[82] J. L. Giraudel and S. Lek, "A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination," *Ecological Modelling,* vol. 146, pp. 329-339, 2001.

[83] N. Goerke, F. Kintzler, and R. Eckmiller, "Multi-SOMs: A New Approach to Self-Organizing Classification," *Lecure Notes in Computer Science,* vol. 3686/2005, pp. 469-477, 2005.

[84] R. Goodacre, J. Pygall, and D. B. Kell, "Plant seed classification using pyrolysis mass spectrometry with unsupervised learning: The application of auto-associative and Kohonen artificial neural networks," *Chemometrics and Intelligent Laboratory Systems,* vol. 34, pp. 69-83, 1996.

[85] M. Granzow, D. Berrar, W. Dubitzky, A. Schuster, F. J. Azuaje, and R. Eils, "Tumor Classification by Gene Expression Profiling: Comparison and Validation of Five Clustering Methods," *ACM SIGBIO Newsletter,* vol. 21, pp. 16 - 22, April 2001.

[86] I. Guyon, *Neural Networks and Applications*. Amsterdam: Elsevier, 1990.

[87] M. Hajibabaei, M. A. Smith, D. H. Janzen, J. J. Rodriguez, J. B. Whitfield, and P. D. N. Hebert, "A minimalist barcode can identify a specimen whose DNA is degraded," *Molecular Ecology Notes,* 2006.

[88] M. Hajibabaei, G. A. C. Singer, E. L. Clare, and P. D. N. Hebert, "Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring," *BMC Biology,* vol. 5:24 2007.

[89] M. Hajibabaei, G. A. C. Singer, P. D. N. Hebert, and D. A. Hickey, " DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics," in *Trends in Genetics*, 2007.

[90] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning,* vol. 45, pp. 171–186, 2001.

[91] M. H. Hassoun, *Fundamental Artificial Neural Networks*: MIT Press, 1995.

[92] S. Haykin, *Neural Networks : A Comprehensive Foundation*, 2 ed. Upper Saddle River, NJ: Prentice Hall, 1999.

[93] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard, "Biological identifications through DNA barcodes," *Proc. R. Soc. B* vol. 270, pp. 313–321, 2003.

[94] P. D. N. Hebert, S. Ratnasingham, and J. R. deWaard, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," *Proc. R. Soc. Lond. B (Suppl.),* vol. 270, pp. S96-S99, 15 May 2003.

[95] P. D. N. Hebert, M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis, "Identification of Birds through DNA Barcodes," *PLoS Biology,* vol. 2, pp. 1657-1663, October 2004.

[96] G. D. D. Hurst and F. M. Jiggins, "Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts," in *Proc Biol Sci*. vol. 272, 2005.

[97] J. Jalasiewicz, M. Williams, A. Smith, T. L. Barry, A. L. Coe, P. R. Brown, P. Brenchley, D. Cantrill, A. Gale, P. Gibbard, F. J. Gregory, M. W. Hounslow, A. C. Kerr, P. Pearson, R. Know, J. Powell, C. Waters, J. Marshall, M. Oates, P. Rawson,

and P. Stone, "Are we now living in the Anthropocene?," *GSA Today,* vol. 18, pp. 4-8, February 2008.

[98] D. H. Janzen, "Forward," in *Plant Conservation: a natural history approach.*, G. Krupnick and J. Kress, Eds.: University of Chicago Press, 2004.

[99] S. Jockusch, "A Neural Network which adapts its structure to a given set of patterns," in *Parallel Processing in Neural System and Computers*, R. Eckmiller, G. Hartmann, and G. Hauske, Eds. Amsterdam; New York: North-Holland, 1990, pp. 169-172.

[100] S. B. Jones, A. E. Luchsinger, and (eds), *Plant Systematics*. Singapore: McGraw-Hill Book Company Inc., 1987.

[101] J. A. Kangas, T. Kohonen, and J. Laaksonen, "Variants of Self-Organizing Maps," *IEEE Transactions on Neural Networks,* vol. 1, pp. 93-99, 1990.

[102] N. K. Kasabov, *Foundations of neural networks, fuzzy systems, and knowledge engineering*, 2 ed.: Marcel Alencar, 1996.

[103] S. Kaski, "Exploratory Data Analysis by the Self-Organizing Map: Structure of Welfare and Poverty in the World," in *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, Singapore, 1996, pp. 498-507.

[104] S. Kaski, "Data exploration using self-organizing maps." vol. Ph.D.: Helsinki University of Technology, Acta Polytechnica, Scandinavica, 1997, http:\www.cis.hut.fi/-sami/thesis.

[105] S. Kaski, J. Kangas, and T. Kohonen, "Bibliography of Self-Organizing Map (SOM) Papers: 1981--1997," *Neural Computing Surveys,* vol. 1, pp. 102-350, 1998.

[106] M. Y. Kiang, "Extending the Kohonen self-organizing map networks for clustering analysis," *Computational Statistics & Data Analysis,* vol. 38, pp. 161-180, 28 December 2001.

[107] K. Kiviluoto, "Topology preservation in self-organizing maps," in *ICNN'96, IEE International Conference on Neural Networks*, IEEF, Service Center, Piscataway, 1996, pp. 294–299.

[108] K. Kiviluoto, "Comparing 2D and 3D self-organizing maps in financial data visualization," in *Methodologies for the Conception, Design and Application of Soft Computing - Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems (IIZUKA'98)*, Fukuoka, Japan, 1998, pp. 68-71.

[109] T. Kohonen, "Construction of similarity daigrams for phonemes by a self-organizing algorithm," Helsinki University of Technology, Espoo, Finland. Report TKK-F-A463, 1981.

[110] T. Kohonen, "Self-Organized Formation of Topologically Correct Feature Maps," *biological Cybernetics,* vol. 43, pp. 59-69, 1982.

[111] T. Kohonen, *Self-Organization and Associative Memory*, 2 ed. vol. 8. Berlin: Springer-Verlag, 1984.

[112] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, "SOM_PAK: the self-organizing map program package," Report A31, Helsinki University of Technology , Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[113] T. Kohonen, "The self-organizing map," *Neurocomputing,* vol. 21, pp. 1-6, 1998.

[114] T. Kohonen, *Self-organizing maps*, 3rd ed. vol. 30. Berlin: Springer-Verlag, 2001.

[115] T. Kohonen, "Software Tools for SOM," in *Self-organizing maps*, 3rd ed. vol. 30 Berlin: Springer-Verlag, 2001, pp. 311-328.

[116] T. Kohonen, "Self-organizing neural projections," *Neural Networks Special Issue,* vol. 19, pp. 723–733, 2006.

[117] P. Koikkalainen and E. Oja, "Self-organizing hierarchical feature maps," in *Proceedings of IJCNN'90, International Joint Conference on Neural Networks*, San Diego, 1990, pp. 279-284.
[118] P. Koikkalainen, "Progress with the tree-structured self-organizing map," in *Proceedings of ECAI'94, 11th European Conference on Artificial Intelligence*, 1994, pp. 211-215.
[119] P. Koikkalainen, "Fast deterministic self-organizing maps," in *Proceedings of ICANN'95, International Conference on Artificial Neural Networks*, 1995, pp. 63-68.
[120] P. Koikkalainen, "Tree Structured Self-Organizing Maps," in *Kohonen Maps*, M. Oja and S. Kaski, Eds.: Elsevier Science, 1999, pp. 121-130.
[121] W. J. Kress, "DNA 'barcoding' of plants," in *Plant Talk*, 2005, http://www.plant-talk.org/resource/dna.html.
[122] W. J. Kress, K. J. Wurdack, E. A. Zimmer, L. A. Weigt, and D. H. Janzen, "Use of DNA barcodes to identify flowering plants," *PNAS,* vol. 102, pp. 8369–8374, June 7 2005.
[123] N. Laitinen, J. Rantanen, S. Laine, O. Antikainen, E. Räsänen, S. Airaksinen, and J. Yliruusi, "Visualization of particle size and shape distributions using self-organizing maps," *Chemometrics and Intelligent Laboratory Systems,* vol. 62, pp. 47-60, 28 April 2002.
[124] R. Lang and K. Warwick, "The Plastic Self Organizing Map," in *Proceeding of the International Joint Conference on Neural Networks, IJCNN'02*, Honolulu, Hawaii, 2002, pp. 727-732.
[125] P. Langley and H. A. Simon, "Applications of Machine Learning and Rule Induction," *Communications of the ACM,* vol. 38, pp. 55-64, 1995.
[126] N. Lavrač, "Selected techniques for data mining in medicine," *Artificial Intelligence in Medicine,* vol. 16, pp. 3–23, 1999.
[127] N. Lavrač, P. Flach, B. Kavšek, and L. Todorovski, "Rule induction for subgroup discovery with CN2-SD," in *ECML/PKDD/IDDM-2002, 2nd International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning* Helsinki, Finland, 2002.
[128] R. D. Lawrence, "A Scalable Parallel Algorithm for Self-Organizing Maps with Applications to Sparse Data Problems," *Data Mining and Knowledge Discovery,* vol. 3, pp. 171-195, June 1999.
[129] R. Leakey and R. Lewin, "The Sixth Extinction," in *The Sixth Extinction: Patterns of Life and the Future of Humankind*: Anchor, 1995, pp. 232-245.
[130] J. Lederberg, "How DENDRAL was conceived and born," in *ACM Symposium on the History of Medical Informatics*, National Library of Medicine 1987.
[131] O. A. Leistner, "Seed plants of southern Africa: families and genera," in *Strelitzia*. vol. 10 Pretoria: National Botanical Institute, 2000.
[132] Leung and Lam, "Fuzzy concepts in expert systems," *Computer,* vol. 21, pp. 43-56, September 1988.
[133] S.-T. Li, "A web-aware interoperable data mining system," *Expert Systems with Applications,* vol. 22, 27 November 2001 2002.
[134] X. Li and C. F. Eick, "Fast Decision Tree Learning Techniques for Microarray Data Collections," in *The 2003 International Conference on Machine Learning and Applications (ICMLA'03)*, Los Angeles, California, 2003.
[135] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, "DENDRAL: a case study of the first expert system for scientific hypothesis formation," *Artificial Intelligence in Medicine,* vol. 61, pp. 209-261, 1993.

[136] P. J. G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention," *Neural Networks,* vol. 15, pp. 11-39, January 2002.

[137] M. Luckow, C. Hughes, B. Schrire, P. Winter, C. Fagg, R. Fortunato, J. Hurter, L. Rico, F. J. Breteler, A. Bruneau, M. Caccavari, L. Craven, M. Crisp, Alfonso Delgado S., Sebsebe Demissew, Jeffrey J. Doyle, Rosaura Grether, Stephen Harris, Patrick S. Herendeen, Héctor M. Hernández, Ann M. Hirsch, Richard Jobson, Bente B. Klitgaard, Jean-Noël Labat, Mike Lock, Barbara MacKinder, Bernard Pfeil, Beryl B. Simpson, Gideon F. Smith, Mario Sousa S., Jonathan Timberlake, Jos G. van der Maesen, A. E. Van Wyk, Piet Vorster, Christopher K. Willis, J. J. Wieringa, and M. F. Wojciechowski, "*Acacia*: The Case against Moving the Type to Australia," *Taxon,* vol. 54, pp. 513-519, May 2005.

[138] S. C. Madeira and A. L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey," *IEEE Transactions on Computational Biology and Bioinformatics,* vol. 1, pp. 24-45, January-March 2004.

[139] Magill, *Magill's Medical Guide*, Revised ed.: Salem Press, 1998.

[140] P. Mangiameli, S. K. Chen, and D. West, "A comparison of SOM neural network and hierarchical clustering methods," *European Journal of Operational Research,* vol. 93, pp. 402-417, 1996.

[141] T. M. Martinetz and K. J. Schulten, "A "Nerual-Gas" Network Learns Topologies," in *Artificial Neural Networks, Proceedings of ICANN'91, International Conference of Artificial Neural Networks*, Amsterdam, 1991, pp. 397-402.

[142] B. R. Maslin, A. E. Orchard, and J. G. West, "Nomenclatural and classification history of *Acacia* (Leguminosae: Mimosoideae), and the implications of generic subdivision ", 2003.

[143] D. Merkl, M. Dittenbach, and A. Rauber, "Uncovering Hierarchical Structures in Data Using the Growing Hierarchical Self-Organizing Map," *Neurocomputing,* vol. 48, pp. 199-216, 2002.

[144] T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, and A. Waibel, "Machine Learning," *Annual Review of Computer Science,* vol. 4, pp. 417-433, June 1990.

[145] E. Moll, *Trees of Natal*, 2nd ed.: University of Cape Town Eco-Lab Trust Fund, 1992.

[146] E. Moll, "*Acacia* for Africa," in *Veld & Flora*, 2005.

[147] A. Moore, "Re-typing *Acacia*," in *Veld & Flora*, 2006.

[148] G. Moore, "The handling of the proposal to conserve the name *Acacia* at the 17th International Botanical Congress - an attempt at minority rule," *Bothalia,* vol. 37, pp. 109-118, 2007.

[149] C. Moritz and C. Cicero, "DNA Barcoding: Promise and Pitfalls," *PLoS Biology,* vol. 2, pp. 1529-1531, October 2004.

[150] F. Murtagh, "Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering," *Pattern Recognition Letters,* vol. 16, pp. 399-408, April 1995.

[151] A. Narayanan, X. Wu, and Z. R. Yang, "Mining viral protease data to extract cleavage knowledge," *Biodiversity and Conservation,* vol. 18, pp. S5-S13, 2002.

[152] A. Neme and P. Miramontes, "Biological Domain Identification Based in Codon Usage by Means of Rule and Tree Induction," in *Computational Methods in Systems Biology*. vol. 3082/2005, V. Danos and V. Schachter, Eds.: Springer Berlin / Heidelberg, 2005, pp. 221-224

[153] J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong, "Analysis and visualization of gene expression data using Self-Organizing Maps," *Neural Networks,* vol. 15, pp. 953-966, October-November 2002.
[154] N. J. Nilsson, *Artificial Intelligence : A New Synthesis*, 2nd ed. San Mateo, California: Morgan Kaufmann, 1998.
[155] G. A. Norton, D. J. Patterson, and M. Schneider, "LucID: A Multimedia Educational Tool for Identification and Diagnostics," 2000.
[156] C. Ohmann, V. Moustakis, Q. Yang, and K. Lang, "Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain," *Artificial Intelligence in Medicine,* vol. 8, pp. 23-36 February 1996.
[157] M. Oja, S. Kaski, and T. Kohonen, "Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum," *Neural Computing Surveys,* vol. 3, pp. 1-156, 2002.
[158] A. E. Orchard and B. R. Maslin2, "(1584) Proposal to conserve the name *Acacia* (Leguminosae: Mimosoideae) with a conserved type," *Taxon,* vol. 52, pp. 362-363, May 2003.
[159] Oxford Dictionary of English. vol. 2009, http://www.askoxford.com.
[160] R. J. Pankhurst, *Biological Identification: The principles and practice of identification methods in biology*. London: Edward Arnold, 1978.
[161] J. L. Pappas, "Biological taxonomic problem soving using fuzzy decision-making analytical tools," *Fuzzy Sets and Systems,* vol. 157, pp. 1687-1703, 2006.
[162] Y.-S. Park, R. Céréghino, A. Compin, and S. Lek, "Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters," *Ecological Modelling,* vol. 160, pp. 265-280, 15 February 2003.
[163] E. S. Peer, "A Serendipitous Software Framework for Facilitating Collaboration in Computational Intelligence," in *Computer Science, Faculty of Engineering, Built Environment and Information Technology*. vol. *Magister Scientiae* Pretoria: University of Pretoria, 2004, http://cirg.cs.up.ac.za/.
[164] E. P. Phillips, *The Genera of South African Flowering Plants*, 2nd ed. vol. 25. Pretoria: Department of Agriculture, 1951.
[165] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: an overview and their use in medicine," *Journal of Medical Systems,* vol. 26, pp. 445-463, October 2002.
[166] V. Podgorelec, P. Kokol, M. M. Stiglic, M. Heričko, and I. Rozman, "Knowledge discovery with classification rules in a cardiovascular dataset," *Computer Methods and Programs in Biomedicine,* vol. 80 Supplement, pp. S39-S49, 2005.
[167] D. Poole, A. Mackworth, and R. Goebel, "Computational Intelligence and Knowledge," in *Computational Intelligence: A Logical Approach* New York: Oxford University Press, 1998, pp. 1-22.
[168] E. Pooley, *The Complete Field Guide to Trees of Natal, Zululand & Transkei*. Durban: Natal Flora Publications Trust, 1993.
[169] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, pp. 81-106, 1986.
[170] J. R. Quinlan, "Generating production rules from decision trees," in *Tenth International Conference on Articial Inteligence*, Milan, Italy, 1987, pp. 304-307.
[171] J. R. Quinlan, *C4.5: Programs for Machine Learning*: Morgan Kaufmann, 1993.
[172] J. R. Quinlan, "Bagging, boosting, and C4.5," in *Thirteenth National Conference on Artificial Intelligence* Cambridge, MA., 1996, pp. 725-730.
[173] J. R. Quinlan, "Rulequest Research Data Mining Tools: C5.0," 1998, www.rulequest.com/see5.

[174] A. E. Radford, W. C. Dickison, J. R. Massey, and C. R. Bell, "Plant Identification," in *Vascular Plant Systematics* New York: Harper & Row, 1986, pp. 522-536.
[175] A. Rauber, D. Merkl, and M. Dittenbach, "The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data," *IEEE Transactions on Neural Networks,* vol. 13, pp. 1331-1341, November 2002.
[176] J. S. Rodrigues and L. B. Almeida, "Improving the learning speed in topological maps of patterns," in *Proceedings of INNc*, Paris, 1990.
[177] R. Rojas, *Neural Networks: A Systematic Introduction*: Springer-Verlag, 1996.
[178] J. H. Ross, *The Acacia Species of Natal - An Introduction to the Indigenous Species*: The Natal Branch of the Wildlife Protection and Conservation Society of South Africa, 1971.
[179] J. H. Ross, "Fabaceae, subfamily Mimosoideae," in *Flora of Southern Africa*. vol. 16, part 1, J. H. Ross, Ed. Pretoria, South Africa: Botanical Research Institute, Department of Agricultural Technical Services, 1975, p. 159.
[180] J. H. Ross, "A Conspectus of the African *Acacia* Species," in *Memoirs of the Botanical Survey of South Africa*. vol. 44, D. J. B. Killick, Ed. Pretoria: Botanical Research Institute, Department of Agricultural Technical Services, 1979, p. 155.
[181] S. E. Rouwhorst and A. P. Engelbrecht, "Searching the Forest: Using Decision Trees as Building Blocks for Evolutionary Search in Classification Databases," in *2000 Congress on Evolutionary Computation, CEC2000*, La Jolla Marriott, San Diego, USA, 2000, pp. 633 - 638.
[182] S. Russel and P. Norvig, *Artificial Intelligence : A Modern Approach*, 2 ed.: Prentice Hall, 2003.
[183] I. Ruthven and M. Lalmas, "Using Dempster-Shafer's Theory of Evidence to Combine Aspects of Information Use," *Journal of Intelligent Information Systems,* vol. 19, pp. 267–301, 2002.
[184] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *Transactions on Systems, Man, and Cybernetics,* vol. 21, pp. 660-674, May 1991.
[185] T. Samad and S. A. Harp, "Feature map learning with partial training data," in *International Joint Conference on Neural Networks (IJCNN'91)*, Piscataway, NJ, 1991, p. 949.
[186] J. W. Sammon Jr, "A Nonlinear Mapping for Data Structure Analysis," *IEEE Transactions on Computers,* vol. C-18, pp. 401-409, May 1969.
[187] E. V. Samsonova, J. N. Kok, and A. P. IJzerman, "TreeSOM: Cluster analysis in the self-organizing map " *Neural Networks,* vol. 19, pp. 935–949, 2006.
[188] L. M. Santos and H. Du Buf, "Identification by Gabor Features," in *Automatic Diatom Identification*. vol. 51, H. du Buf and M. M. Bayer, Eds., pp. 187-220.
[189] P. H. Schalk and P. Oosterbroek, "Interactive knowledge systems: meeting the demand for disseminating up-to-date biological information," *Biodiversity Letters,* vol. 3, pp. 119-123, Jul.- Sep., 1996.
[190] R. J. Schalkoff, *Artificial Neural Networks*: McGraw-Hill, 1997.
[191] E. Schmidt, M. Lotter, and W. McCleland, *Trees and Shrubs Of Mpumalanga and Kruger National Park*, 1st ed.: Jacana, 2002.
[192] J. F. Schreer, R. J. O'Hara Hines, and K. M. Kovacs, "Classification of Dive Profiles: A Comparison of Statistical Clustering Techniques and Unsupervised Artificial Neural Networks," *Journal of Agriculture , Biological, and Environmental Statistics,* vol. 3, pp. 383-404, 1998.

[193] G. Schwarzer, W. Vach, and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in Medicine,* vol. 19, pp. 541 - 561, February 29 2000.
[194] G. Shafer, *A mathematical theory of exidence*. Princeton: Princeton University Press, 1976.
[195] S. Shanmuganathan, P. Sallis, and J. Buckeridge, "Self-organising map methods in integrated modelling of environmental and economic systems," *Environmental Modelling & Software,* vol. 21, pp. 1247-1256, 2006.
[196] E. H. Shortliffe, S. G. Axline, B. G. Buchanan, T. C. Merigan, and S. N. Cohen, "An Artificial Intelligence Program to Advise Physicians Regarding Antimicrobial Therapy," *Computers and Biomedical Research,* vol. 6, pp. 544-560, 1973.
[197] E. H. Shortliffe, S. G. Axline, B. G. Buchanan, and S. N. Cohen, "Design Consideration for a Program to Provide Consultation in Clinical Therapeutics," in *The 13th San Diego Biomedical Symposium*, 1974.
[198] E. H. Shortliffe and B. G. Buchanan, " A model of inexact reasoning in medicine," *Mathematical Bioscience,* vol. 23, pp. 351–379, 1975.
[199] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System," *Computers and Biomedical Research,* vol. 8, pp. 303-320, 1975.
[200] E. H. Shortliffe, F. S. Rhame, S. G. Axline, S. N. Cohen, B. G. Buchanan, R. Davis, A. C. Scott, R. Chavez-Pardo, and W. J. van Melle, "MYCIN, Computer Program Providing Antimicrobial Therapy Recommendations," in *American Federation for Clinical Research, Western Sectional Meeting*. vol. 33 Carmel, California, 1975.
[201] E. H. Shortliffe, *Computer-Based Medical Consultation, MYCIN*. New York, America: Elsevier, 1976.
[202] E. H. Shortliffe, "MYCIN: A Knowledge-Based Computer Program Applied to Infections Diseases," in *Annual Meeting of the Society for Computer Medicine* Las Vegas, Nevada, 1977.
[203] E. H. Shortliffe, B. G. Buchanan, and E. A. Feigenbaum, "Knowledge Engineering for Infectious Disease Therapy Selection," in *International Conference on Cybernetics and Society*: IEEE, 1979.
[204] E. H. Shortliffe, B. G. Buchanan, and E. A. Feigenbaum, "Knowledge Engineering for Medical Decision Making: A Review of Computer-Based Clinical Decision Aids," in *Proceedings of the IEEE*, 1979, pp. 1207-1224.
[205] T. Shortliffe and R. Davis, "Some Considerations for the Implementation of Knowledge-Based Expert Systems," *SIGART Newsletter,* vol. 55, pp. 9-12, December 1975.
[206] J. Smaldon and A. A. Freitas, "Improving the Interpretability of Classification Rules in Sparce Bioinformatics Datasets," in *Proceeding of AI-2006, the Twenty-sixth SGAI International Conference On Innovative Techniques and Applications of Artificial Intelligence*, 2006, pp. 377-381.
[207] N. Smit, *Guide To The Acacias of South Africa*, Ist ed.: Briza, 1999.
[208] G. F. Smith, M. Buys, M. Walters, D. Herbert, and M. Hamer, "Taxonomic research in South Africa: the state of the discipline," *South African Journal of Science,* vol. 104, pp. 254-258, July/August 2008.
[209] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray

Hybridization," *Molecular Biology of the Cell,* vol. 9, pp. 3273–3297, December 1998.

[210] C. A. Stace, *Ways and Means*, 2nd ed. Cambridge: Cambridge University Press, 1989.

[211] W. T. Stearn, *Botanical Latin*. London: Nelson, 1967.

[212] M. Stoeckle, "Taxonomy, DNA, and the Bar Code of Life " *BioScience,* vol. 53, pp. Viewpoint 2-3, September 2003.

[213] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA : Genetics,* vol. 96, pp. 2907-2912, March 1999.

[214] A. C. Tan and D. Gilbert, "An empirical comparison of supervised machine learning techniques in bioinformatics," in *First Asia Pacific Bioinformatics Conference (APBC 2005)* Adelaide, Australia, 2003.

[215] D. Tautz, P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler, " A plea for DNA taxonomy," *Trends in Ecology & Evolution,* vol. 18, pp. 70-74 2003.

[216] K. Thiele, "The Library of Life," Lisbon 2003.

[217] B. T. Tien and G. van Straten, "A Neuro-Fuzzy Approach to Identify Lettuce Growth and Greenhouse Climate," *Artificial Intelligence Review,* vol. 12, pp. 71-93, 1998.

[218] J. Timberlake, C. Fagg, and R. Barnes, *Field Guide to the Acacias of Zimbabwe*: CBC, 1999.

[219] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of Gene Expression Data Using Self-Organizing Maps," *FEBS Letters,* vol. 451, pp. 142-146, 21 May 1999.

[220] A. Ultsch, "Knowledge Extraction from Self-Organizing Neural Networks," in *Information and classification*, O. Opitz, B. Lausen, and R. Klar, Eds. Berlin: Springer, 1993, pp. 301–306.

[221] A. Ultsch, D. Guimaraes, D. Korus, and H. Li, "Knowledge Extraction from Artificial Neural Networks and Applications," in *Transputer Anwender Treffen/World Transputer Congress*, Aachen, 1993.

[222] A. Ultsch, "The Integration of Neural Networks with Symbolic Knowledge Processing," in *New Approaches in Classification and Data Analysis*, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Eds. New York: Springer Verlag, 1994, pp. 445-454.

[223] A. Ultsch and C. Vetter, "Self-Organizing-Feature-Maps versus Statistical Clustering Methods: A Benchmark," University of Marburg, FG Neuroinformatik & Kuenstliche Intelligenz, Marburg, Research Report Nr 0994, 1994.

[224] A. Ultsch and D. Korus, "Integration of Neural Networks with Knowledge-Based Systems," in *Proc. IEEE Int. Conf. Neural Networks*, Perth, Australia, 1995.

[225] A. Ultsch, "Maps for the Visualization of high-dimensional Data Spaces," in *Proceedings of the Workshop on Self organizing Maps* Kyushu, Japan, 2003, pp. 225-230.

[226] A. Ultsch and L. Hermann, "Architecture of Emergent Self-Organizing Maps to Reduce Projection Errors," in *Proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2005, pp. 1-6.

[227] A. Ultsch, "Using Information Retrieval Methods for a Comparison of Algorithms to find differentially expressed Genes in Microarray Data," DataBionics Research Lab, Department of Computer Science, University of Marburg, Marburg, Technical Report Nr. 12, October 2007.

[228] P. E. Utgoff, "Incremental Induction of Decision Trees," *Machine Learning,* vol. 4, pp. 161-186, 1989.
[229] B. van Wyk and P. van Wyk, *Field Guide To Trees of Southern Africa*, 1st ed.: Struik, 1997.
[230] F. Venter and J. A. Venter, *Making the most of indigenous trees*. Pretoria: Briza, 2002.
[231] J. Vesanto, "SOM-Based Data Visualization Methods," in *Intelligent Data Analysis*. vol. 3, 1999, http://citeseer.ist.psu.edu/vesanto99sombased.html
[232] J. Vesanto, "Using SOM in Data Mining," in *Computer Science and Engineering*. vol. Licentiate of Science in Technology Espoo: Helsinki University of Technology, 2000, http://citeseer.ist.psu.edu/vesanto00using.html.
[233] J. Vesanto and E. Alhoniemi, "Clustering of the Self-Organizing Map," *IEEE Transactions on Neural Networks,* vol. 11, pp. 586-600, May 2000.
[234] J. Vesanto, "Data Exploration Process Based on the Self-Organizing Map," in *Computer Science and Engineering*. vol. Ph. D. Espoo: Helsinki University of Technology, 2002.
[235] J. Vesanto, "SOM algorithm implementation in SOM Toolbox ", Updated 18 March 2005, accessed 23 March 2010, http://www.cis.hut.fi/projects/somtoolbox/documentation/.
[236] C. Von der Malsburg, "Self-Organization of orientation sensitive cells in the striate cortex," *Biological Cybernetics,* vol. 14(2), pp. 85-100, 1973.
[237] E. G. Voss, "The history of keys and phylogenetic trees in systematic biology," *Journal of the Scientific Laboratories,* vol. 43, pp. 1-25, 1952.
[238] W. J. Walley and M. A. O'Connor, "Unsupervised pattern recognition for the interpretation of ecological data," *Ecological Modelling,* vol. 146, pp. 219-230, 1st December 2001.
[239] A. T. Watson, M. A. O'Neill, and I. J. Kitching, "Automated identification of live moths (Macrolepidoptera) using Digital Automated Identification SYstem (DAISY)," *Systematics and Biodiversity,* vol. 1, pp. 287-300, 2003.
[240] P. J. D. Weeks and K. J. Gaston, "Image analysis, neural networks, and the taxonomic impediment to biodiversity studies," *Biodiversity and Conservation,* vol. 6, pp. 263-274, 1997.
[241] A. F. Weller, A. J. Harris, and J. A. Ware, "Artificial neural networks as potential classification tools for dinoflagelate cyst images: A case using the self-organizing map clustering algorithm," *Review of Palaeobotany and Palynology,* vol. 141, pp. 287-382, 2006.
[242] P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," in *Applied Mathematics*. vol. Ph.D. Cambridge: Harvard University, 1974.
[243] Q. D. Wheeler, "Taxonomic triage and the poverty of phylogeny," *Philosophical Transactions of The Royal Society B,* vol. 359, pp. 571-583, 18 March 2004.
[244] R. D. Whitworth, H. Dawson, and E. Baudry, "DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae)," *Proceedings of the Royal Society B,* vol. 274, pp. 1731-1739, 2007.
[245] A. Wilcox and G. Hripcsak, "Knowledge Discovery and Data Mining to Assist Natural Language Understanding," in *Proc AMIA Annu Fall Symp.* , 1998, pp. 835-9.
[246] K. W. Will and D. Rubinoffb, "Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification," *Cladistics,* vol. 20, pp. 47–55, 2004.

[247] K. W. Will, B. D. Mishler, and Q. D. Wheeler, "The Perils of DNA Barcoding and the Need for Integrative Taxonomy," *Syst. Biol.,* vol. 54, pp. 844–851, 2005.
[248] C. Woese and G. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proc Natl Acad Sci U S A,* vol. 74 (11), pp. 5088-90, 1977.
[249] C. Woese, L. Magrum, and G. Fox, "Archaebacteria," *J Mol Evol* vol. 11 (3), pp. 245-51, 1978.
[250] C. Woese, O. Kandler, and M. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.," *Proc Natl Acad Sci U S A,* vol. 87 (12), pp. 4576-9, 1990.
[251] J. B. Woolley and N. D. Stone, "Application of Artificial Intelligence to Systematics: Systex-A Prototype Expert System for Species Identification," *Systematic Zoology,* vol. 36, pp. 248-267, September 1987.
[252] S. M. Wraith, J. S. Alkins, B. G. Buchanan, W. J. Clancey, R. Davis, L. M. Fagan, J. Hannigan, A. C. Scott, E. H. Shortliffe, W. J. van Melle, V. L. Yu, S. G. Axline, and S. N. Cohen, "Computerized Consultation System for Selection of Antimicrobial Therapy," *Am J Hosp Pharm,* vol. 33, pp. 1304-1308, December 1976.
[253] Z. R. Yang and K.-C. Chou, "Mining Biological Data Using Self-Organizing Map," *Journal of Chemical Information and Computer Sciences* vol. 43, pp. 1748-1753, 2003.
[254] V. L. Yu, B. G. Buchanan, E. H. Shortliffe, S. M. Wraith, R. Davis, A. C. Scott, and S. N. Cohen, "Evaluating the Performance of a Computer-Based Consultant," *Computer Programs in Biomedicine,* vol. 9, pp. 95-102, 1979.
[255] V. L. Yu, L. M. Fagan, S. M. Wraith, W. J. Clancey, A. C. Scott, J. Hannigan, R. Blum, L., B. G. Buchanan, and S. N. Cohen, "Antimicrobial Selection by a Computer. A Blinded Evaluation by Infectious Diseases Experts," *Journal of the American Medical Association,* vol. 242, pp. 1279-282, September, 21 1979.
[256] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets and Systems* vol. 69, pp. 125-139, 1995.
[257] L. A. Zadeh, "Outline of a theory of usuality based on fuzzy logic," University of California 1986.
[258] J. M. Zurada, *Introduction to Artificial Neural Systems*: West Publishing Company, 1992.

Appendix A : Acronyms, Abbreviations and Glossary of Terms

ASCII – American Standard Code for Information Interchange is a character-encoding scheme based on the order of the English alphabet. ASCII codes represent text in devices such as computers.

classification – The process of defining and naming classes of organisms [49].

CI – Computational Intelligence.

DELTA – DELTA format Language for TAxonomy is a flexible format for encoding taxonomic data descriptions prior to computer processing.

empirical - Empirical data are data that are produced by experiment or observation. It refers to the use of working hypotheses that are testable using observation or experiment.

FlowerSOM – SOM model developed from a set of KZN *Acacia* flower data (a subset of the whole *Acacia* data set).

glabrous – smooth, hairless.

genus – A genus is a taxonomic unit used in the classification of organisms. In the hierarchy of biological classification genus comes above species. For example *Acacia*.

identification – Identification is the process of assigning a specimen to a (pre-existing) taxon. The name of the taxon can then be used as an index to find known information about the taxon, and therefore about the specimen itself [49].

key – A key is a reference tool or device which aids identification of biological entities such as plants. A key offers a series of choices where for each choice the user must choose between prominent contrasting features (characteristics or attributes) of the entity to be defined. By a process of elimination successively smaller groupings of entities are split off until the specimen can be identified, i.e. the series of choices progressively leads to the definition of the species of the entity. The groupings derived from such a key are artificial rather than natural and this method of identification has the disadvantage that if a mistake is made at any stage identification will be completely wrong.

KZN – KwaZulu-Natal.

LeafSOM – SOM model developed from a set of KZN *Acacia* leaf data (a subset of the whole *Acacia* data set).

PodSOM – SOM model developed from a set of KZN *Acacia* seed and pod data (a subset of the whole *Acacia* data set).

polyclave – multiple-entry access key.

pubescent – hairy, covered with short hairs

Self-organizing map – A SOM is a result of a nonparametric regression process that maps high-dimensional, nonlinearly-related data onto a visual, two-dimensional display in order to perform classification and clustering [114].

SOM - Self-Organizing Map (see Self-organizing map for definition).

Species (sp.) – Species is a taxonomic rank which is used in biological classification. In the hierarchy of biological classification species comes below genus. For example in the name *A. kraussiana* the species is *kraussiana*.

stochastic - Stochastic means being or having random variables. A stochastic model is a tool for estimating probability distributions of potential outcomes by allowing for random variation in one or more inputs over time.

subspecies (subsp.) – A taxonomic rank that is subordinate to species and which usually arises as a consequence of geographical distribution or isolation within a species. For example, *A. brevispica* subsp. *dregeana*.

taxon (pl. taxa) – A taxon is a name designating an organism or group of organisms (no matter the level of the taxonomic hierarchy. The group (one or more) is sufficiently similar to each other to be considered as a single unit. The group is sufficiently dissimilar to any other group to be considered as a candidate for membership of that group.

taxonomist – a biologist who specializes in the classification of organisms into groups on the basis of their structure, origin and behaviour.

taxonomy – The science of biological description, classification and identification [49].

ThornSOM – SOM model developed from a set of KZN *Acacia* habit and thorn data (a subset of the whole *Acacia* data set).

topology – Topology refers to the relationships and characteristics shared in common among data samples.

TreeSOM – SOM model developed from a whole set of KZN *Acacia* data.

TDWG - Taxonomic Data Working Group.

variety (var.) (pl. varieties) – A taxonomic rank below subspecies and as such it has a ternary name, for example *A. luederitzii* var. *retinens*.

whole data set – The data set used in this research project to describe the **KZN** *Acacia* species.

Appendix B : Batch SOM Algorithm

A general algorithm for the SOM method can be summarized as:

```
standardize input data:
  initialise weight vectors;
  for each iteration (t = 0, 1, 2, …)
     for each input vector (i = 0, 1, 2, …)
        for each neuron weight vector (j = 0, 1, 2, …)
           for each vector component (k = 0, 1, 2, …)
```

$$Dist_{i,j} = \sqrt{\sum_{k} (input_{i,k} - weight_{j,k})^2}$$

```
           next k;
        next j;
        find winner;
        adjust winner's neighbourhood;
     next i;
  next t;
  display output;
```