A predictive biogeography of selected alien plant invaders in South Africa

Submitted to the departments of Geography and Zoology & Entomology in full requirement for the degree of Master of Science.

by

Jennifer G. Youthed

Rhodes University
January 1997

Abstract

Five techniques were used to predict the potential biogeography of the four alien plant species, *Acacia longifolia, Acacia mearnsii, Opuntia ficus-indica* and *Solanum sisymbrifolium*. Prediction was based on five environmental factors, median annual rainfall, co-efficient of variation for rainfall, mean monthly maximum temperature for January, mean monthly minimum temperature for July and elevation. A geographical information system was used to manage the data and produce the predictive maps. The models were constructed with presence and absence data and then validated by means of an independent data set and chi-squared tests. Of the five models used, three (the range, principal components analysis and discriminant function analysis) were linear while the other two (artificial neural networks and fuzzy logic) were non-linear. The two non-linear techniques were chosen as a plant's response to its environment is commonly assumed to be non-linear. However, these two techniques did not offer significant advantages over the linear methods. The principal components analysis was particularly useful in ascertaining the variables that were important in determining the distribution of each species. Artifacts on the predictive maps were also proved useful for this purpose.

The techniques that produced the most statistically accurate validation results were the artificial neural networks (77% correct median prediction rate) and the discriminant function analysis (71% correct median prediction rate) while the techniques that performed the worst were the range and the fuzzy classification. The artificial neural network, discriminant function analysis and principal component analysis techniques all show great potential as predictive distribution models.

Acknowledgements

My thanks must go firstly to my supervisors, Dr Trevor Hill and Dr Martin Villet. I wish to thank Martin particularly for his constant advice, guidance, organisation and encouragement.

The generosity of the people who supplied the plant distribution data; Lesley Henderson, Dave Hoare, Martin Hill and Luke Perkins is much appreciated. I am also grateful to Dr Tony Palmer for obtaining the environmental surfaces for me and for advice on IDRISI. The Computing Centre for Water Research (CCWR) in Pietermaritzburg is gratefully acknowledged for allowing me to use data supplied by them.

I owe my thanks as well to the many people who took the time to answer my queries and offer advice: Professor Hulley, Di Donnelly, Dr Neser, Dr Zimmerman, Dr Hoffman, Dr le Maitre, Steve Higgens and Glyn Armstrong. I owe particular thanks to Dr Allon Poole and Dr Mike Burton for taking the time to explain neural networks and fuzzy logic to me.

This study was funded by an FRD core grant.

Table of Contents

SECTION TWO

SECTION THREE

List of figures

List of tables

SECTION I

CHAPTER 1
INTRODUCTION

1.1) INTRODUCTION

For centuries, man has been interested in the distribution of organisms on the surface of the earth and, in particular, the relationship between distribution patterns and climate. For example, many early vegetation maps, like some today, used climate to determine vegetation zones (de Laubenfels, 1975). Today the study of the distribution of patterns of organisms is termed biogeography which Meadows (1985: 1) defines as "a branch or form of general natural science, its immediate area of concern being plants and animals (including man) and their patterns of distribution". These patterns embrace not only past and present distributions, but also future ones. In many cases it is useful to be able to predict where in space (and time) an organism is likely to occur. In being able to do so, one can implement control measures in regions that are predicted to be at risk from organisms with undesirable characteristics, such as disease vectors. Areas suitable for the survival of desirable organisms, such as particular agricultural crops, can also be delimited. Sometimes predictive mapping of an organism needs to be carried out for both of the above reasons; a case in point being that of invasive alien vegetation.

Invasive alien plants are species which have been introduced into countries where they are not native and have the potential to invade indigenous vegetation (Stirton, 1987). As such, they are considered by many to represent a threat to natural ecosystems and should therefore be controlled (Stirton, 1987; Henderson *et al*, 1987; Vermeulen, 1989). However, some of the alien invasive species were originally introduced into a country for some purpose, such as *Acacia mearnsii*, the black wattle, which was brought into South Africa to be cultivated for its tannin-rich bark (Stirton, 1987). Today this species is at the heart of a highly profitable economic industry (De Beer, 1986). Other alien plant species are in a similar position, in the centre of a controversy between those who want them eradiated and those, such as the wattle growers, to whom they provide a livelihood (de Selincourt, 1992). Predictive distribution mapping can offer potential benefits to both sides of this debate.

There is no denying that invasive alien plants do have undesirable characteristics with one of the most undesirable being their ability to out-compete and replace natural flora, rendering once productive or aesthetically beautiful areas an impenetrable mass of alien vegetation (Stirton, 1987). Most invasive plants compete with natural vegetation for light, water, nutrients and space (Wells *et al*, 1986). Some species, for example *Lantana camara* and *Hypericum* spp (Vahrmeijer, 1981) are poisonous to animals and humans, while others, such as *Opuntia* spp, have thorns that may cause injury to livestock. This can have negative economic repercussions for farmers who may lose valuable stock to poisoning or injury.

Invasive alien plants also threaten biodiversity by outcompeting and replacing natural vegetation. This is particularly a problem in the floristically diverse and highly endemic fynbos region of South Africa (Richardson *et al*, 1992). Not only is biodiversity threatened, but also ecotourism (van Wilgen *et al*, 1996) which relies heavily on the uniqueness of the fynbos biome to draw foreign visitors.

Apart from their effect on natural vegetation, alien plants can also have negative impacts on other parts of their host country's ecosystems. For example, by attracting pollinators to themselves, alien plants may reduce the number of visits by pollinators to indigenous plants. Indigenous species may also be shaded or camouflaged by the alien plants, resulting in an increase in the time pollinators take to find the indigenous species (Rebelo, 1987).

Another negative ecosystem impact is the effect that alien vegetation can have on water resources. With reference to South Africa, the Department of Water Affairs and Forestry (DWAF, 1996a: 1) states that "the biggest problem with alien invasive plants is that they use much more water than our local plants and trees". Alien trees in particular, can result in a marked reduction in streamflow (Smith and Bosch, 1989; Van Lill *et al*, 1980). Le Maitre *et al* (1996) estimate that Cape Town could be deprived of 30% of its water supply if alien plant infestation in the catchments of the Western Cape continues unchecked.

Alien plant infestation may also create problems for fire managers by changing the vegetation structure, increasing the fuel load, obstructing access to certain areas and increasing catchment erosion after fires (DWAF 1996b, Richardson *et al*, 1992 and van Wilgen *et al*, 1996).

Predictive models of alien invader spread can serve as both preventative and control measures. By being able to predict areas at high risk from invasion by alien species, steps can be taken to prevent the possible spread of the species into the high risk areas, noticeably by good management methods. For example, overgrazing and the subsequent destruction of natural climax vegetation may open the way for the aggressive alien species which can out-compete the natural vegetation (MacDonald & Jarman, 1985; Tivy, 1993). Control measures can also be intensified at already infested areas that show the greatest risk of spreading. Accurate modelling techniques could allow the prediction of presence/absence for a species in areas that are not easily accessible for fieldwork and in areas, such as the former homelands, that have previously not been sampled. The resulting maps of predicted distribution can give an indication of how much of the country is at risk from invasion by alien plant species.

However, to turn to the other side of the invasive alien plant controversy, predictive mapping can also delimit areas that are optimally suited to the survival of particular alien species. As mentioned above, some people depend on alien plants for a livelihood and the very qualities that make the alien species successful invaders enable them to survive in marginal areas where other vegetation would not (de Selincourt, 1992). For example, the prickly pear, *Opuntia ficus-indica* is able to tolerate very dry conditions due to its succulent nature (Moran & Zimmermann, 1984). This species is an important fruit crop (Cactus Pear Growers Association, 1996), drought fodder plant and vegetable (Zimmerman and Moran, 1991), and is used to make dye, soap and liqueur (Muir, 1986). Many squatters rely on *Acacia saligna*, the port jackson willow, as a source of firewood, building material and even as a fodder crop (de Selincourt, 1992). This species' hardiness and ability to fix nitrogen, thereby enriching the soil, enable it to tolerate harsh conditions where few other species could survive (de Selincourt, 1992).

It is clear that there are advantages in being able to map and predict the biogeography of certain alien plants. The question of course is how one goes about predicting distribution. This study aims to examine several ways in which prediction can be achieved, bearing in mind the statement by Elston & Buckland (1993: 93) that "good models for prediction will be chosen on the basis of simplicity, ease of use and the demonstrable quality of prediction".

1.2) SCOPE AND AIMS

The aims of this thesis are to:

1) Explore some of the computer-based modelling techniques that can be used to predict the distribution of alien plants. Emphasis is placed on little-used techniques or the new use of established ones.

2) Evaluate how the techniques used to predict distribution performed in terms of

   a) accuracy of prediction

   b) simplicity and ease of use

   c) specific advantages

   d) ability to be taken further

3) Delimit areas where the plants are predicted to occur.

It is beyond the scope of this thesis to give a complete analysis of the ecological requirements of the plants. An in-depth discussion of the reasons for their possible success or failure to invade the areas predicted as suitable for invasion is also not intended. The emphasis is on *where* the plants could occur and not *why*.

As the fundamental limits for survival of an organism are set by its abiotic environment (Putman & Wratten, 1985), five abiotic environmental parameters were chosen to predict the distribution of the plant species. While the distribution of an organism within its suitable environment may be further influenced by biotic factors such as the effect of man (Putman & Wratten, 1985), it is beyond the scope of this study to model these interactions as well.

The use of environmental parameters to predict distribution has the added advantage that these factors tend to remain relatively constant over time (Fabricius & Coetzee, 1992). In other words, an area that is predicted as environmentally suitable for invasion at present is still likely to be environmentally suitable in the future. Time was therefore treated as a constant in this study in that it was not modelled explicitly. An area that is predicted as suitable for invasion will remain suitable in environmental terms whether it is already invaded or yet to be invaded.

Finally, a review of available predictive modelling techniques is not intended; instead emphasis is placed on some potentially useful or under-utilized methods, or on the utilisation of established techniques in new ways, in the hope that they will prove to be accurate and cost-effective means of predicting distribution.

CHAPTER 2
THEORETICAL BACKGROUND

2.1) MODELLING AND PREDICTION

Models can be defined as "simplified abstractions of the real world" (von Gadow & van Hensbergen, 1987: 44). They can be used to help us make sense of data and to help us understand parts of complex systems that would be too difficult to model in their entirety. The word 'model' as utilized in this thesis is used in the above sense and 'modelling' is taken to refer to the process of producing the predictive maps.

There are many different types of models, suited to different purposes and selection of the correct model depends on the result desired (Sharov, 1995). There are accounts in the literature of the different families of models and the uses to which they can be put (see for example Higgens & Richarson, 1996; Jeffers, 1982; von Gadow & van Hensbergen, 1987) and the different types of models will thus not be reviewed here.

Much of the difficulty in modelling lies in finding a balance between the number of parameters used and the accuracy of the output desired (Sharov, 1995). The use of only a few parameters often leads to a simple model that is easy to handle but which may not be very accurate; while the use of a large number of parameters may result in greater accuracy but also in a model that is complex and difficult to work with. The success of a model depends on the selection of suitable levels of complexity for the data available and the modelling aims (Sharov, 1995). One must therefore decide what the objectives of the model are and then determine an acceptable level of accuracy to produce the simplest possible model to deliver the level of performance required. Performance of a model is often judged on how well it can predict (Buckland & Elston, 1993). Prediction is useful as it allows one to use available data to determine processes in areas, or times, where there is no available data (Buckland & Elston, 1993).

2.1.1) Bioclimatic factors as a basis for modelling

Chapman & Busby (1994: 183-4) state that "prediction of likely occurrences of a species in unsurveyed areas can be derived from correlation between known species' records and environmental factors, particularly climate". Prediction of distribution using bioclimatic factors has proved to be a popular modelling technique (Accone, 1992; Abrams, 1985; Bauer *et al*, 1994; Boynton, 1989; Knight, 1986; Liebhold *et al*, 1992; Walker, 1990). This popularity can be attributed to a number of reasons. Firstly, climate sets the limits for most organisms, but especially for plants which cannot migrate to escape unfavourable conditions (de Laubenfels, 1975). In many cases there are upper and lower survival thresholds of environmental factors that set the limits on distribution (Putman & Wratten, 1985). Bioclimatic factors thus serve as a sort of 'base level' in models. Climate can provide the first step in a modelling process (Box *et al*, 1993), and other factors that may influence the distribution of the organism within its climatic range, such as fire and competition, can be added later.

Secondly, climatic factors are often readily available; many countries keep meterological records and these records are usually available on a regional scale and stretch back for a number of years.

Bioclimatic variables also have the advantage that they do not change rapidly, i.e. areas that are environmentally suitable now are still likely to be so in the future. This is in contrast to variables such as fire regime or stocking density which may alter quite markedly over a short time period.

In determining the bioclimatic preferences for an organism, one is essentially defining its niche (Putman & Wratten, 1985). Niche is a concept that is often used in ecology, but its actual definition is a source of contention and it has come to mean different things when used in different contexts. According to Schoener (1988), several concepts of a 'niche' can be distinguished.

Grinnell's (1914, in Schoener, 1988) concept of a niche is defined in spatial and dietary terms. It is often regarded as synonymous with the word 'habitat'. A niche is not seen as equivalent to the occupant of the niche. An occupant may occupy a niche, but is not *the* niche.

The niche as defined by Elton (1927, in Schoener, 1988) is often taken to mean the role of the organism in the community, and is thus regarded as different to Grinnell's concept. Schoener (1988) however, points out that the niches of Elton and Grinnell have many similarities and the main difference between the two is that Elton's niche may include more than one species or may be meaningfully empty.

The third concept of a niche, and the one that shall be used for this study, was pioneered by Hutchinson in 1978. Hutchinson saw a niche as being the sum of all the environmental influences acting on the organism. His niche is defined as a regional n-dimensional hyperspace (Schoener, 1988). He further divided niches into fundamental and realized niches. A fundamental niche is "defined by environmental dimensions within which that species can survive and reproduce. A species may be excluded from parts of its fundamental niche because of competition and other biotic interactions. The reduced hypervolume is then termed the realized niche" (Austin *et al*, 1990: 161). This study is essentially defining the fundamental niches of the plant species.

2.1.2) Some issues in bioclimatic modelling

As mentioned above, many organisms have upper and lower bioclimatic limits. This lends itself to the assumption that they approximate a Gaussian distribution (Putman & Wratten, 1985, figure 1). For example, at very low temperatures an organism will not be able to survive; as the temperature increases so does the suitability of the environment for the organism, up to an optimum level. After this optimal level is reached, the environment becomes less suitable with increasing temperature until an upper threshold is reached beyond which survival is unlikely.

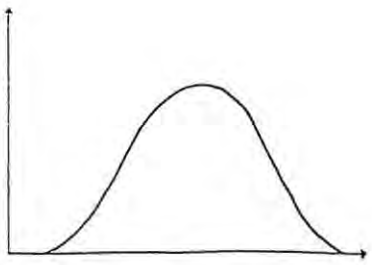


Figure 1. A normal (Gaussian) distribution.

Unfortunately many statistical models are linear in nature and do not take the above relationship between the organism and its environment into account. Non-linear techniques offer a solution to this problem by allowing bell-shaped response curves, instead of only linear ones, to be incorporated into models. Two non-linear methods, fuzzy classification and artificial neural networks are used to model distribution in this study, along with some more conventional linear techniques.

Another issue that needs to be considered when modelling the bioclimatic distribution of an organism, is the time that the organism has had to establish its bioclimatic range. This is particularly important in the case of alien invader vegetation where different species may be introduced to a country at various times. A plant that has only been recently introduced may not have had time to fully establish within its environmental limits and any model based on its present limits may not be modelling its true environmental limits at all. In these cases Box *et al* (1993) suggest that the climatic range of the species in its native country be examined to determine its environmental preferences.

To ensure that the environmental preferences of a plant are as representative as possible, it is desirable to have a large sample size. Unfortunately, as for this study, large sample sizes are not always available. The sample sizes for this project are relatively small, but two extended data sets are used to determine if an increase in sample size significantly affects the prediction quality of the distribution models.

2.2) GEOGRAPHICAL INFORMATION SYSTEMS AND MODELLING

2.2.1) What are geographical information systems?

Geographical information systems (GISs) are defined by Aronoff (1991: 39) as "any manual or computer based set of procedures used to store and manipulate geographically referenced data", and by Korte (1994: 207) as a "system of computer hardware, software, and procedures designed to support the capture, management, manipulation, analysis, modelling, and display of spatially referenced data for solving complex planning and management problems". The

common element between these two definitions is the ability of the system to handle spatially referenced data. This ability to manipulate spatial data, especially to synthesize new data by combining different layers of information, is what tends to distinguish GISs from graphically orientated systems such as CAD (Computer Aided Design) and database management systems (Cowen, 1988; Korte, 1994).

GISs are generally regarded as consisting of four main components (Aronoff, 1991; Young, 1986):

1) An input subsystem

Data input into a GIS is in two basic formats, either raster or vector. Raster format consists of a grid of equal-sized cells called pixels, each of which is given a value. Satellite data is raster in format. Vector data on the other hand, represents objects as a series of points, lines or polygons and attributes may be attached to features. Vector data is often entered into the system by digitizing. There are various advantages and disadvantages to each data type and most modern GIS systems make an effort to support both data formats. Once entered into a GIS, the data sets may be variously termed layers, coverages or surfaces.

2) A data management subsystem

A GIS should allow basic data management ('housekeeping') functions such as copying, deleting, moving and sorting data.

3) A manipulation/analysis subsystem

This subsystem is often regarded as the core of a GIS and perhaps the most important manipulation allowed by geographical information systems is the ability to overlay or merge different coverages to produce a new image (Cowen, 1988; Liebhold *et al*, 1993). Often different overlay options are available, for example layers may be added, subtracted or partially covered. Most GISs also support measurement of spatial objects and the transformation of spatial data.

4) An output subsystem

Output from a GIS is most often in graphical form, although print-outs of tables are further options. Output is usually produced through printers or plotters.

2.2.2) Why GIS?

GISs are particularly powerful modelling tools and have been used to deal with issues ranging from the mapping of natural hazards (Wadge *et al*, 1993) to maintaining a database on the distribution of African Elephants (Michelmore, 1994). They display a number of advantages over conventional mapping methods. While their ability to handle and synthesize large amounts of spatial data is of prime importance, they also offer the ability to update maps and data quickly and easily, something that is not readily accomplished with other mapping methods. This means that a database can be continuously updated and built upon, rather than having to draft a fresh map each time new information is added. Apart from these two main advantages, GISs also offer fast access to the data and rapid processing times as well as advanced graphical capabilities (Elston & Buckland, 1993); they can provide coverage in inaccessible or poorly sampled areas and can produce high-quality output.

By using GIS as a tool to manipulate and analyze the data produced by the predictive techniques, one can make use of the above advantages and hopefully produce modelling techniques superior to the predictive statistical methods used on their own. In this study, emphasis was placed on finding ways to optimally link the GIS to the statistical techniques.

2.2.3) Software

The grid-based geographic information system IDRISI, version 4.1 (Eastman, 1994) and IDRISI for Windows, version 1 (Eastman, 1995) was selected for use in this research. IDRISI was chosen as it is mainly grid-based (although it has vector capabilities as well). Raster format was desirable as the environmental coverages were in raster format, validation of the predictive maps with point data would be easier, as each pixel has a value (McAllister *et al*, 1994) and presence/absence data and the patterns of distribution are more easily represented in grid format (Miller, 1994).

2.3) THE SPECIES AND ENVIRONMENTAL VARIABLES SELECTED FOR MODELLING

2.3.1) Plant species selected

The following alien plant species were selected for modelling:

i) *Acacia longifolia* (Andr.) Willd.

ii) *Acacia mearnsii* De Wild.

iii) *Opuntia ficus-indica* (L.) Mill.

iv) *Solanum sisymbrifolium* Lam.

From the distribution data available, the above four species were selected. Terrestrial species were chosen as water plants are more easily influenced by local conditions such as a sewage spill and require different environmental parameters to terrestrial plants for modelling purposes. Species from different climatic regions of the world were chosen to determine if the predictive techniques could accurately predict distribution over a range of climates. The two *Acacia* species are native to the temperate areas of Australia; *O. ficus-indica* is indigenous to central Mexico (Stirton, 1987) and *S.sisymbrifolium* originates from South America (Henderson, 1995). All of these species also have some economic use, particularly *A. mearnsii* and *O. ficus-indica* which are grown commercially. This makes distribution mapping worthwhile for both control and cultivation purposes.

Acacia longifolia

Also known as the long-leaved or Sydney golden wattle, this *Acacia* is a native of south eastern Australia (Stirton, 1987). As such, it prefers temperate, coastal climates. It was originally introduced into South Africa in 1827 as an ornamental and to bind sand dunes (Stirton, 1987). *Acacia longifolia* is a small, evergreen tree that produces finger-like yellow infloresences (Stirton, 1987). It has phyllodes that are up to 180mm in length with two to five prominent longitudinal veins. Flowering occurs from June to November and afterwards the plant produces narrow, brown seedpods that are constricted between the seeds. The seeds themselves are dark brown with a white aril and seed stalk (Henderson *et al*, 1987).

Wells *et al* (1986) list this wattle as having the following undesirable characteristics: it transforms both landscape and habitat, competes with natural vegetation for space and nutrients, replaces natural vegetation and its seed is a contaminant. There is some evidence that *A. longifolia* may also be poisonous to stock as it has tested positive for hydrocyanic acid (Munday, 1988). Moran *et al* (1986) consider this as one of the most important aggressive invader plants in South Africa due to the copious amount of seed it produces and the high viability of the seeds.

Acacia mearnsii

As a native of south-east Australia, black wattle prefers a temperate climate although it will tolerate some frost. According to De Beer (1986), it will grow in areas with a mean annual rainfall of between 500 and 1500 mm, but prefers high rainfall areas. While it will grow on shallow soils if there is enough water, it prefers well-drained, deep soils. This species was already established in the Cape Town Botanical Gardens as far back as 1858 (De Beer, 1986); however it is generally accepted that this plant was introduced into South Africa from Australia in 1864 by John van der Plank (Stirton, 1987: 48). It is an evergreen tree that grows up to about 15m high; it has dark-green, bipinnate leaves that have many raised nectar glands. The bark is usually grey-brown to black, becoming rough in older trees (De Beer, 1986). The plant produces scented, pale yellow flowers between August and November (Stirton, 1987) followed by black seeds that have a whitish-yellow seed stalk. The seeds of black wattles can remain viable in the soil for over 50 years (Stirton, 1987), making eradication very difficult and *A. mearnsii* is now a declared invader plant in South Africa (Henderson *et al*, 1987).

Acacia mearnsii is grown extensively for its bark which is rich in the tannins used by the tanning industry and it is cultivated, especially in KwaZulu-Natal, for this purpose. It has also been used as a shade and fuelwood tree and was at one stage planted in the Cape Province as a firebreak (De Beer, 1986). The wood is also used for charcoal, to make paper, hardboard and parquet flooring and is often used for wattle-and-daub huts, while the resin may be used for adhesives (Stirton, 1987).

This species is currently the centre of a controversy between the wattle growers, who cultivate it for its bark, and conservationists who want the alien removed. *Acacia mearnsii* is particularly a problem along watercourses where it can alter stream geomorphology

(Rowntree, 1991) and reduce runoff. The Department of Water Affairs and Forestry (1996a) regard *A. mearnsii* as one of the species that have the most impact on the water resources of South Africa. The seeds of this alien are largely waterborne and it can thus spread rapidly down streams, often forming dense, impenetrable thickets that obstruct watercourses. These thickets may also impede access, smother indigenous vegetation, reduce grazing land and render areas aesthetically unpleasing (Macdonald & Jarman, 1985). *Acacia mearnsii* has also been linked to excessive stream bank erosion, the creation of debris dams and changes in channel form and function (Rowntree, 1991). In addition *A. mearnsii* competes with natural vegetation for water, light, nutrients and space (Wells *et al*, 1986).

Opuntia ficus-indica

Most commonly known as the prickly pear, this cactus was thought to have first been introduced into South Africa over 250 years ago (Zimmermann & Moran, 1991). Its native country is Central Mexico (Stirton, 1987). It is a succulent branching shrub, with flattened, oblong cladodes that may be covered with spines. The fruit may also be spiny although spineless varieties have been developed for commercial cultivation. Yellow or orange flowers are produced in November (Stirton, 1987) on or near the margins of the stems. According to the Cactus Pear Growers Association (1996), *O. ficus-indica* prefers moderate to very hot summers and moderate to cold (-10°C) winters; with a rainfall of between 180 and 600mm, preferably during summer. It dislikes hail and foggy weather (coastal and escarpment) and while it will grow in most soils, it needs a fairly good depth of between 800 and 1000mm to grow optimally.

Opuntia ficus-indica has proved to be something of a problem plant, especially in the eastern Cape, where it forms dense, impenetrable thickets and replaces natural grazing. The thorns may cause damage to stock that eat the plant, particularly in times of drought, when other food is unavailable. As the plant will readily regenerate from cladodes that fall to the ground, they are extremely difficult to eradicate. Their other undesirable characteristics (Wells *et al*, 1986) include their ability to replace natural vegetation (especially grazing land), to hinder access to farmlands and to compete for light, water, space and nutrients with other plants. Moran and Zimmermann (1984) attribute the success of this weed to its ability to resist drought conditions due to its waxy cuticle and succulent nature, its ease of reproduction and its success as a competitor.

Solanum sisymbrifolium

Also known as the dense-thorned bitter apple or the wild tomato (Wells *et al*, 1986), this weed is an invader of disturbed and pastoral land (Nel, 1988). It is a natural inhabitant of South America and was probably brought into South Africa along with horse feed during the South African wars in the 1900's (Nel, 1988). As such it is a relatively recent invader in this country, compared to the other three species. *Solanum sisymbrifolium* is a branched, spiny shrub and the leaves are covered with glandular trichomes (Hill, 1994). The plant produces white to bluish flowers and bright red fruit that contain many seeds (Henderson *et al*, 1987).

It tends to be a pioneer plant and has an extensive underground root system which makes it troublesome to eradicate. Apart from being difficult to get rid of, it may be poisonous, its seed is a contaminant and the plant is competitive with a tendency to replace natural vegetation (Wells *et al*, 1986). The plant has little commercial value except that the fruit are a source of the glycoalkaloid, solasodine, which is used, *inter alia*, in oral contraceptives (Hill, 1994).

2.3.2) Environmental variables selected

Bearing in mind the modelling dilemma of complexity versus simplicity, five abiotic environmental variables were chosen to predict the distribution of the four plant species. Due to storage and time processing constraints it was decided to build the model with only a few of the potentially limiting variables of the species' distribution, rather than start with many and spend processing time reducing the number of variables to a manageable size. This was done at the risk of ignoring a possibly important variable. However, the lack of an important variable would probably be obvious and the model could easily be enlarged later. There is also some evidence that in many cases, prediction can be accomplished with only a few key variables, for example, Rogers & Williams (1993) found that they only needed one variable to predict the distribution of the tsetse fly with 82% accuracy. With specific reference to discriminant function analysis they state, "in theory, as more variables are included in the analysis, there should be a greater separation of the presence and absence centroids, and a more accurate prediction of the species concerned. In practise, relatively few variables make a contribution to the predicted distributions, and the other variables can be ignored" (Rogers & Williams, 1993: S82).

Two other factors that need to be considered when selecting variables to model with, are the degree to which they are readily available and the ease with which they can be measured. Predictive models developed with variables that require sophisticated equipment for measurement or well-established meterological networks cannot be easily applied in less developed countries.

Of the parameters easily available, the following five were chosen: median annual rainfall (MAR), co-efficient of variation for rainfall (COV), mean monthly maximum temperature for January (MAXT), mean monthly minimum temperature for July (MINT) and elevation (ELV).

Tivy (1993) states that the two most important limiting factors for vegetation are water and temperature. The use of the mean temperatures for the hottest and coldest months of the year give an idea of the range of temperatures tolerated by the plant species. Elevation is an indirect environmental gradient that is correlated with changes in the direct gradients of temperature and rainfall (Palmer, 1991) and is thus sometimes used as a substitute for these two factors. A combination of median annual rainfall and elevation has been successfully adopted by Palmer (1991) and Palmer & Van Staden (1992) to predict the distribution of certain plant communities in South Africa. Co-efficient of variation for rainfall gives an indication of how variable the rainfall is for a particular area.

CHAPTER 3
DATA AND DATA QUALITY

3.1) PLANT SPECIES DISTRIBUTION DATA

3.1.1) Calibration data

The calibration data set is comprised of the data used to construct the predictive models. The data are of a high resolution (nearest minute) and the data set is drawn from a number of sources. Distribution data in the form of presence and absence for all four species were obtained from southern Natal and the grasslands of the Eastern Cape from Luke Perkins (1996, pers com) and Dave Hoare (1996, pers com) respectively. Additional presence data were gathered from records of the Selmar Schonland Herbarium, Rhodes University, and from GPS (global positioning system) readings made by the author between Grahamstown and Cape St Francis. As the sample sizes for the data sets were small, except for *A. mearnsii* (table 1), it was decided to repeat the predictive techniques with two larger sample sizes (termed the extended data sets) to see what effect sample size would have on the outcome of the predictions.

Two larger samples were created, one for each of the *Acacia* species. Additional data points for *A. longifolia* were obtained from Dennill (1987), while the original sample size for *A. mearnsii* was considered large enough to act as the second extended data set. The small data set for *A. mearnsii* was constructed by randomly cutting out just over half of the original data set. Thus the calibration data consisted of a small sample set for each of the four species (figure 2a, c, e, f) and then an additional extended data set each for *A. longifolia* (figure 2b) and *A. mearnsii* (figure 2d)

An absence data set (figure 3) was obtained along with the presence data and is comprised of sites where none of the four species occur. This data set was used for the discriminant function analysis. Separate absence data sets (figure 4a, b, c) for the artificial neural networks had to be constructed for reasons explained in the chapter on artificial neural networks.

While it would have been ideal to include data from the countries where the plants are native (and have therefore had sufficient time to establish their range), this was not easily obtainable and therefore only the readily available local data was used.

3.1.2) Validation data

These data sets were obtained from Lesley Henderson (1996, pers com) at the National Botanical Institute, Pretoria and cover the whole country. The data comprises quarter degree square (i.e. coarse resolution) records of sightings of the plants.

One of the problems in using a more coarsely resolved data set is that only one part of the quarter degree square can be recorded on the environmental coverages. This is because the coverages have a resolution of one minute by one minute, while the quarter degree squares have a resolution of 15 by 15 minutes. It was decided to record the centre point of the quarter degree square. This could lead to an underestimation of the success of the predictive maps as the plant could occur anywhere within the quarter degree square, but would only be recorded as present if the centre was on an area of presence. But, by the same token, if the quarter degree squares were buffered to allow presence to be recorded anywhere in the square this would also not produce a true reflection of the presence of the plant and could lead to an overestimation of the success of the models. It was decided to take a conservative approach and record only the centre point of each quarter degree square. However, this approach was assumed to give a fairly good estimate of the accuracy of the predictive maps due to spatial autocorrelation; i.e. the fact that neighbouring cells tend to have similar values.

Table 1. The number of points in each of the calibration and validation data sets.

Data set	Calibration			Validation
	Small	Extended	Absence	
Acacia longifolia	14	50	540	50
Acacia mearnsii	25	60	540	257
Opuntia ficus-indica	12	-	540	449
Solanum sisymbrifolium	9	-	540	27

3.2) ENVIRONMENTAL VARIABLES

3.2.1) Median annual rainfall (MAR)

The median annual rainfall surface was constructed by Dent *et al* (1989) at the Computing Centre for Water Research (CCWR) in Pietermaritzburg, South Africa. It is at a 1' by 1' resolution and covers the whole of South Africa. It was constructed from daily and monthly rainfall records from 9 409 rainfall stations. The final version was regressed with physiographic factors and interpolation was performed on the regression residuals.

3.2.2) Co-efficient of variation for rainfall (COV)

The co-efficient of variation for rainfall functions as an indication of the variability of the rainfall for an area and is usually expressed as a percentage. The higher the co-efficient, the more variable (and thus less reliable) the rainfall. This surface was constructed by interpolating from point data obtained from the CCWR database and is also at a 1' by 1' resolution.

3.2.3) Mean monthly maximum temperature for January (MAXT)

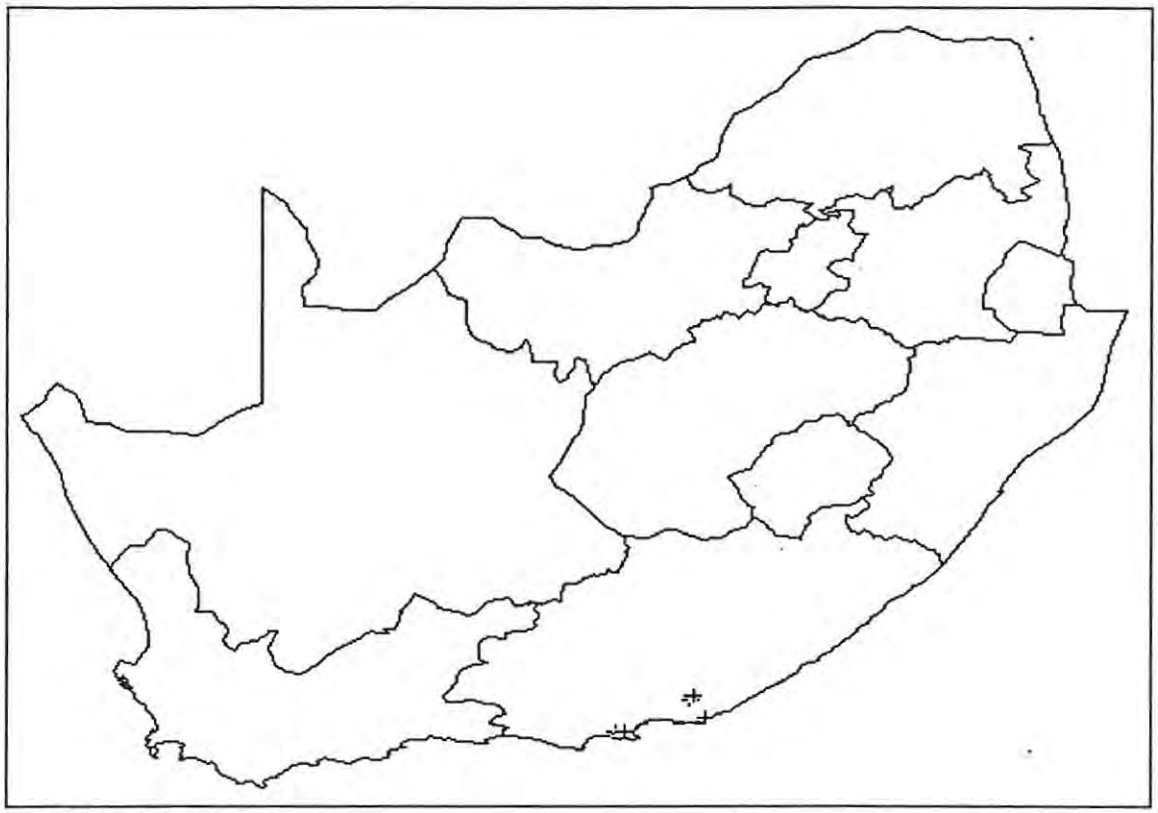
MAXT is a surface of the mean monthly temperature for the hottest month of the year (January) for South Africa. This surface was constructed by interpolation from point data from the CCWR. It also has a resolution of 1' by 1'.

3.2.4) Mean monthly minimum temperature for July (MINT)

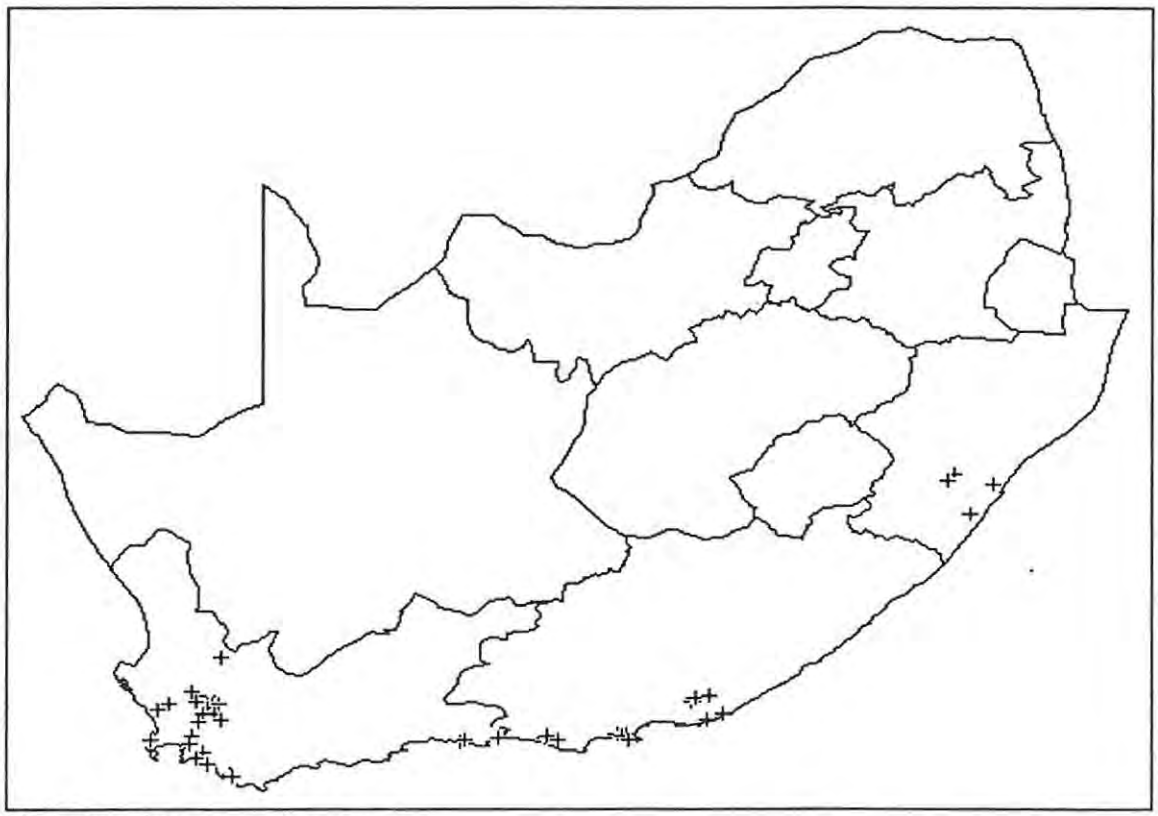
This is an interpolated surface at a 1' by 1' resolution. It is constructed from point data of mean minimum monthly temperature for July (the coldest month) in South Africa.

3.2.5) Elevation (ELV)

ELV is a digital terrain model for the whole of South Africa on a 1' by 1' minute grid. It was produced by Dent *et al* (1989) at the CCWR from topographical maps at the 1:250000 and 1:50000 scale.

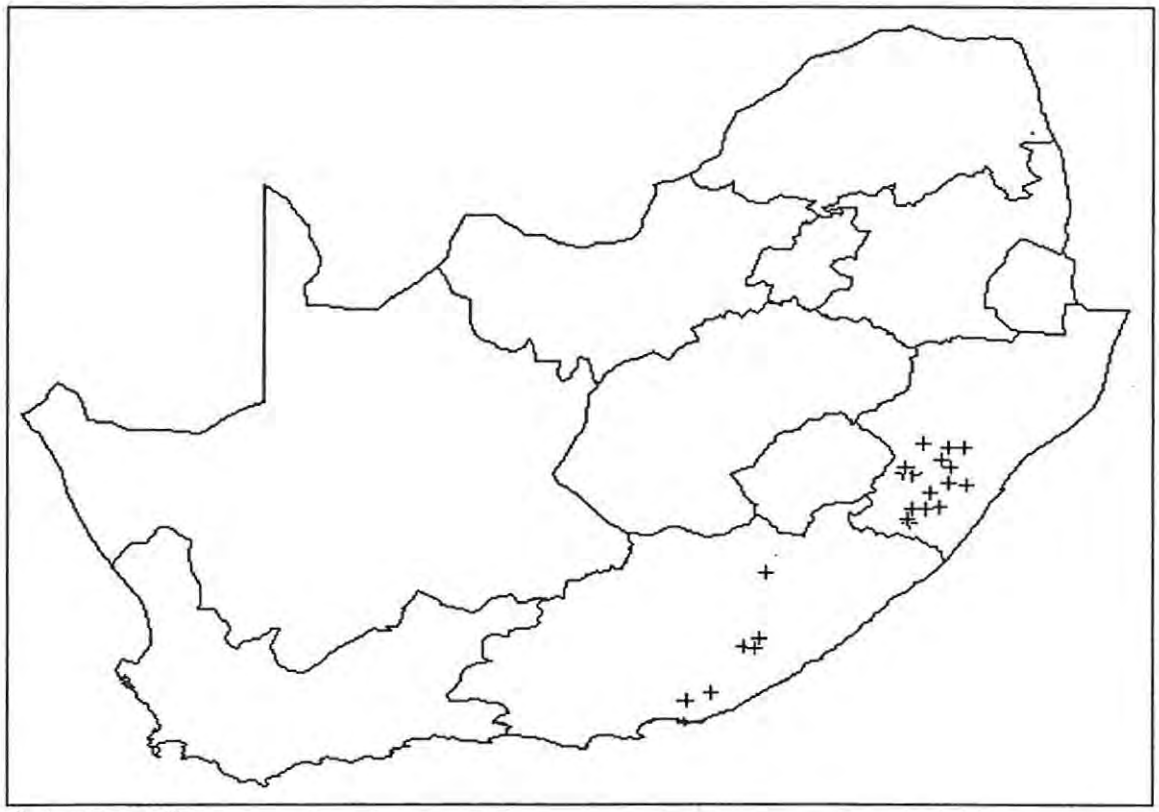


a) *A. longifolia*

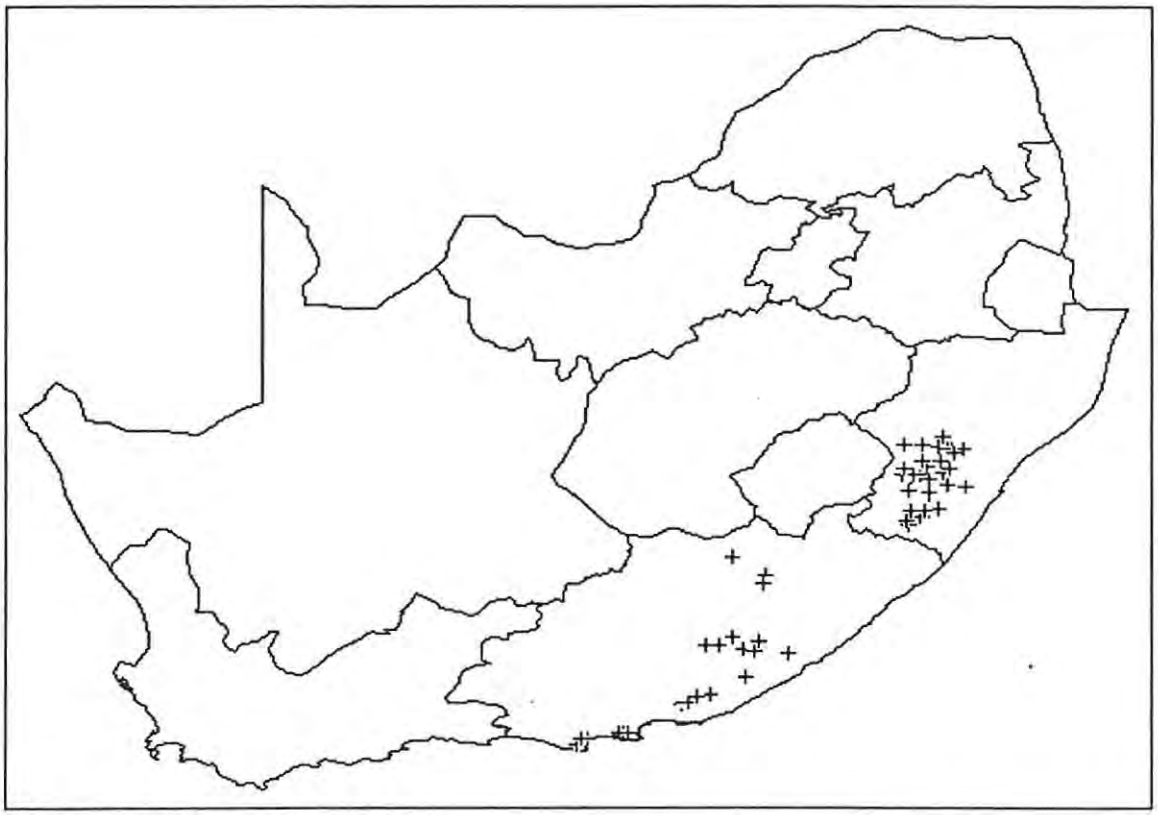


b) *A. longifolia* (extended data set)

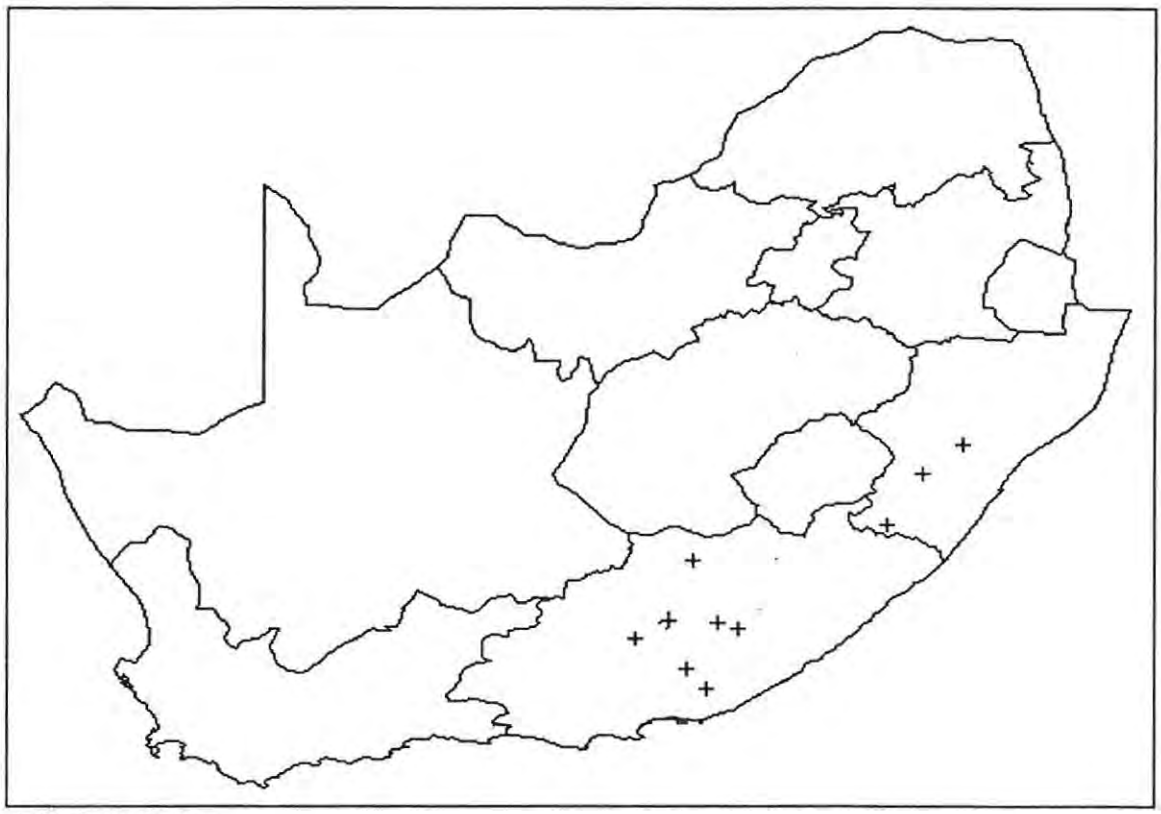
Figure 2. The calibration data sets. Areas of presence are denoted by the '+' sign.



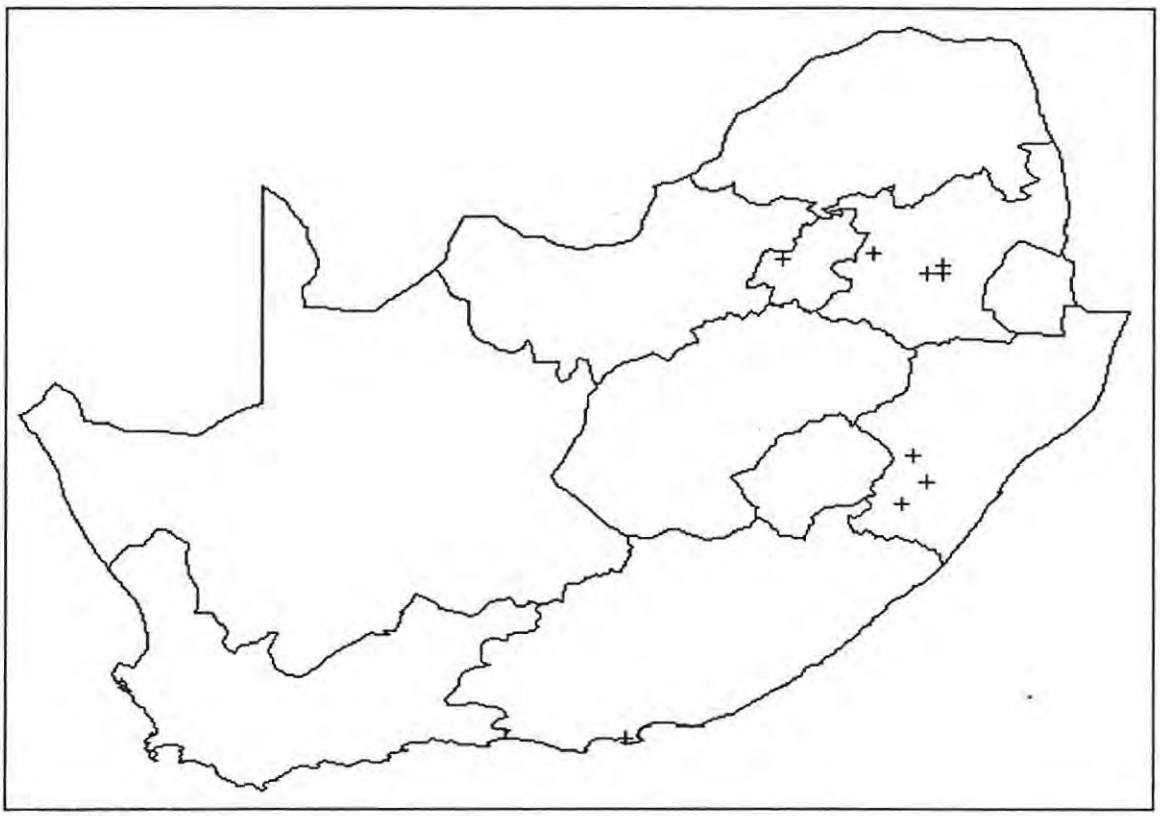
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

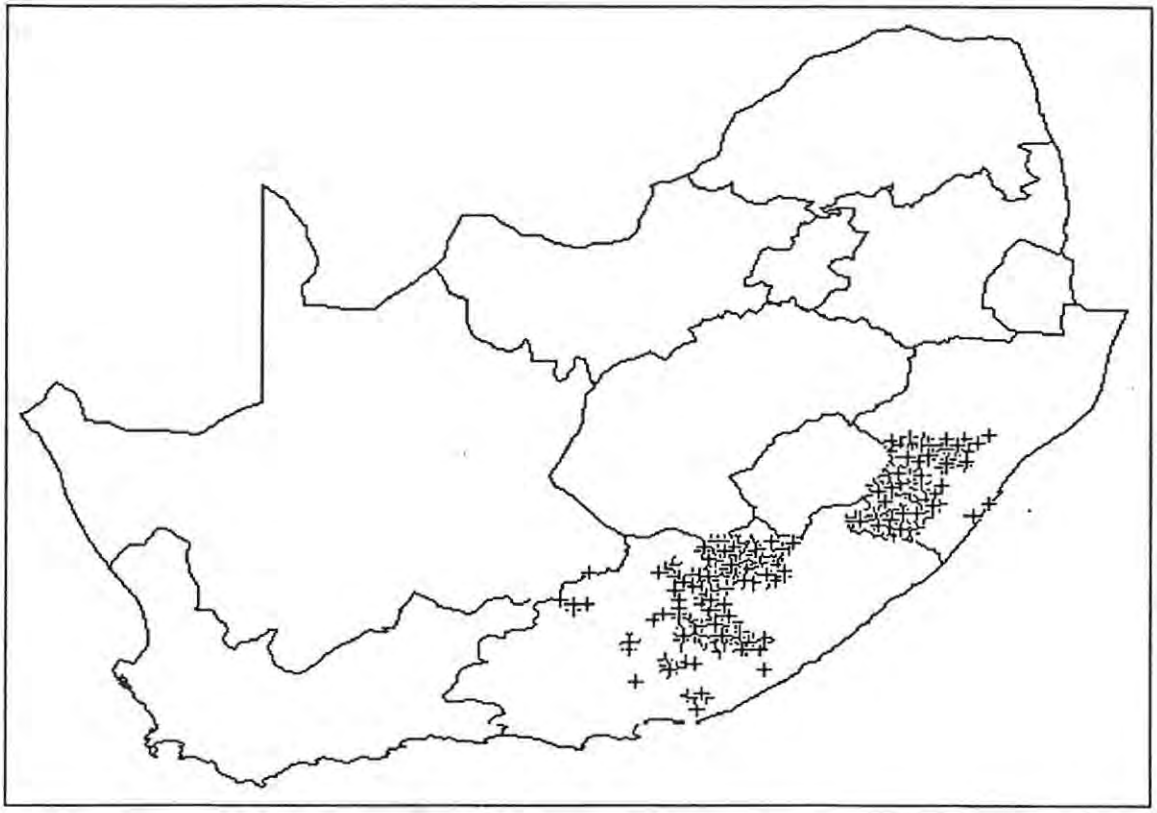
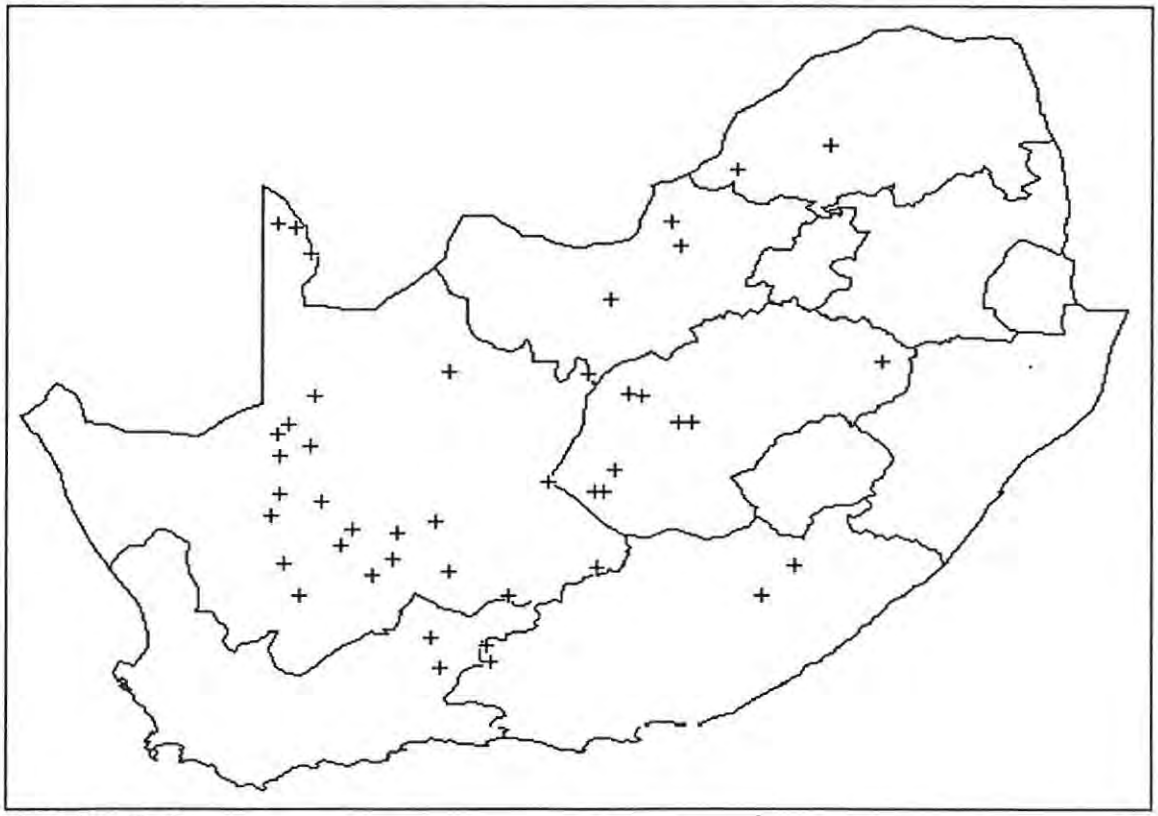


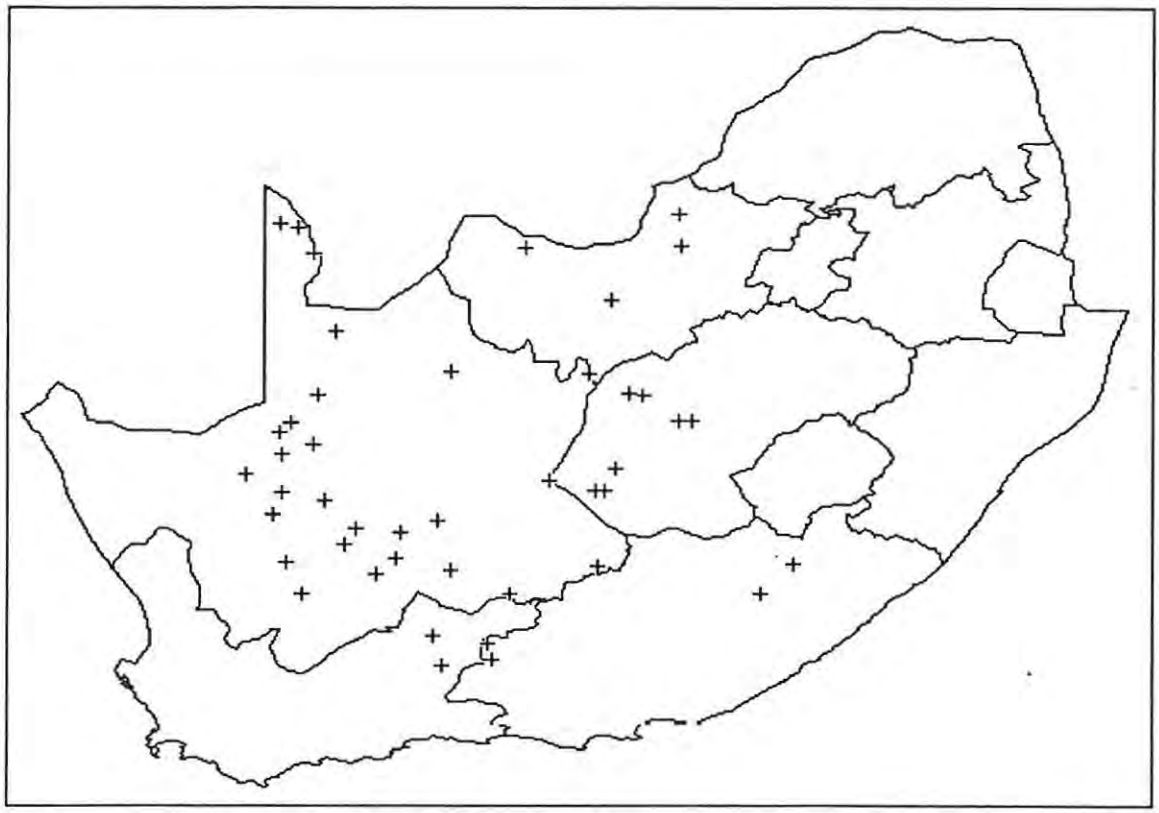
Figure 3. Absence data set for *A. longifolia, A. mearnsii, O. ficus-indca* and *S. sisymbrifolium.*



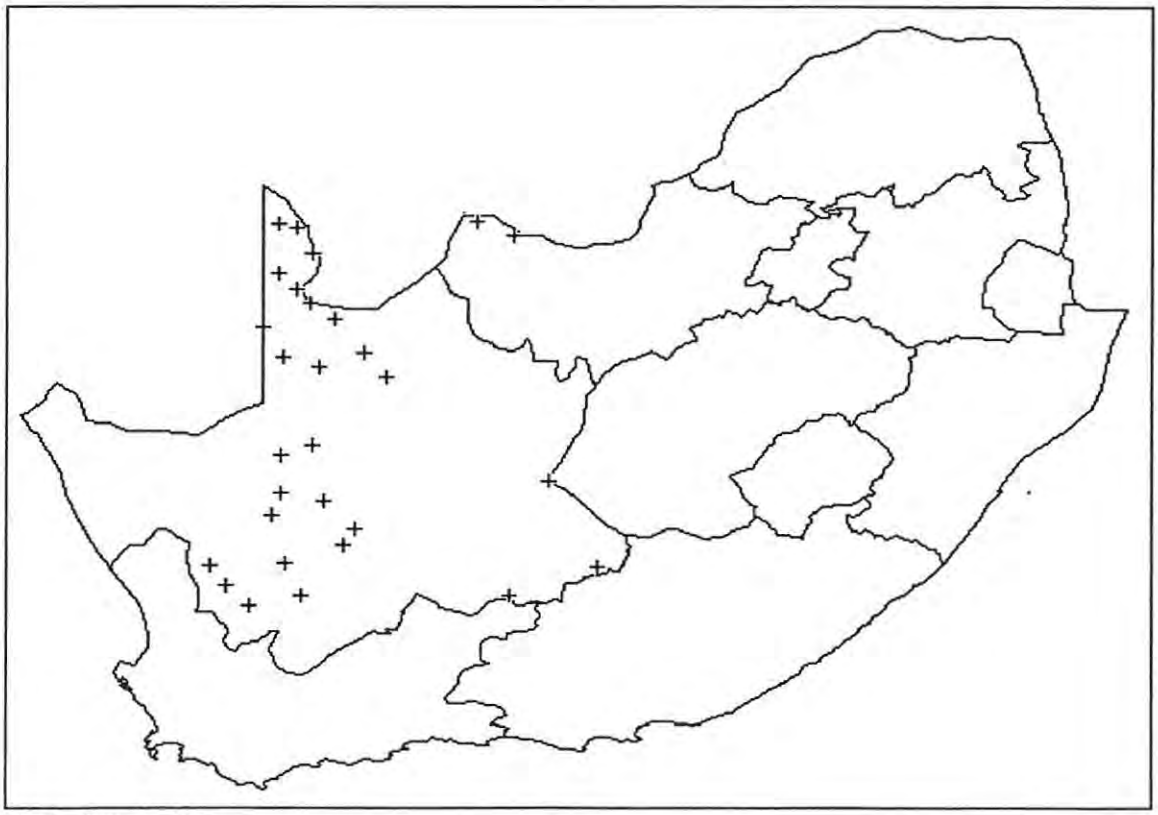
a) *A. longifolia*

Figure 4. The absence data sets for the artificial neural networks.

Areas of absence are denoted by the '+' sign.



b) *A. mearnsii* and *S. sisymbrifolium*



c) *O. ficus-indica*

3.3) DATA QUALITY

While it is generally accepted that data quality is of importance, and that output can only be as good as the original input, one should define what is meant by quality. Chrisman (1991) uses the concept of quality to mean the fitness of the data for use. In other words, levels of quality can vary according to what purpose the data is being used for. He further states that as "error is inescapable, it should be recognised as a fundamental dimension of the data" (Chrisman, 1991: 167). So one must decide what an acceptable level of error is for a particular application and accept that it is not possible to remove all error. This can result in savings in time and money. However, this is not an invitation to sloppy methods and Chrisman (1991) qualifies his previous statement by saying that one should avoid introducing needless error.

Brunsdon and Openshaw (1993), on the other hand, believe that the above approach can lead to error accumulation and has serious consequences such as identifying incorrect locations, failing to find meaningful relationships in the data or discovering false patterns. They maintain that it is not possible to decide on an acceptable level of error as most GISs cannot provide an estimate of the error of the data surfaces. They see this as a major functional failure of GISs. In the absence of error estimation tools in GISs, and the complexity of alternative approaches (see Brunsdon and Openshaw, 1993), it is likely that users will follow the first approach and attempt to decide on an acceptable error level and strive to keep error levels as low as possible.

With these two approaches in mind, there are several data quality issues that should be examined by data managers.

3.3.1) Error

Error is often divided into random error, the accumulated errors of which are assumed to cancel each other out and are therefore quantifiable by statistical confidence limits (Michelmore, 1994); and systematic error or bias, which is methodological and cumulative and should be monitored and noted. The process of keeping track of data, its history and errors is termed lineage (Korte, 1994).

Error in the predictive maps could accumulate from three main sources: errors in the CCWR database, which would affect all of the surfaces, error in the distribution data and user error while manipulating the images. Extensive checking was carried out by Dent *et al* (1989) to ensure that the MAR and ELV surfaces were as free from error as possible. The CCWR flags all suspect data in its database and any suspect values for the temperatures and co-efficient of variation were not used when interpolating these surfaces. The distribution data was checked for duplicate values and suspect co-ordinates. While it is possible that the absence data may contain sites marked as absent but where the plant is actually present (e.g. if the plant was accidentally overlooked, or hidden by another), this was not deemed to occur frequently enough to be a problem. If a site was recorded as the plant being both present and absent, then the absence data was deleted and the presence data used, as the plant was obviously present in some part of the site. All due care was taken to limit user error.

3.3.2) Accuracy

Accuracy can also be divided into two types, positional and attribute accuracy. Both types refer to the "closeness of an observation to its true value" (Chrisman, 1991: 166), i.e. either to the true position of the observation on the earth's surface (positional accuracy), or to the attribute it represents (attribute accuracy).

Positional accuracy is often a problem when geocoding specimens, especially herbarium specimens which may have no more than a name describing their location and no precise co-ordinates (Lindenmayer *et al*, 1991). However, the positional accuracy for the calibration data set is high as the presence and absence sites were recorded to the nearest minute using a GPS. Accuracy is affected by the original sources of data and error may be compounded by overlaying, misclassification, changes between raster and vector data and other data manipulation processes in the GIS (Buckland and Elston, 1993; Goodchild, 1991). It is important to keep a check on the errors in the data and the resulting estimate of accuracy of the end result. The accuracy of a map often depends on the original compiler, who must decide what should or should not be represented. He in turn may be hampered by the resolution of the image which will determine how much can be fitted onto it (the ability of GIS to store information as different layers has relieved this problem somewhat). In general, the coarser the resolution of the map, the fewer the objects that can be represented on it. Goodchild (1991: 196) points out that this is often acceptable as "the map is intended to give

a visual impression of spatial variation, not an exact inventory". Again the concept of 'fitness for use' can be applied.

3.3.3) Precision

Precision refers to the number of decimal places that can be handled. Precision is often diminished in databases due to storage constraints or rounding-off operations. Most GISs support a high level of precision, but then require large amounts of storage space and longer processing times. The data manager must often decided whether to sacrifice precision or space and time.

All of the original coverages, except for the temperature surfaces, were integer in format i.e. they had no decimal places; but the discriminant function analysis, neural network and fuzzy classification produced coverages that were precise to six decimal places. This level of precision was maintained as far as possible in subsequent manipulations of the surfaces. The results of the validation of the maps were rounded off to the nearest percent. The temperature surfaces had one decimal place, but were multiplied by ten so that they became integer images as this made them easier to work with.

3.3.4) Resolution

Resolution refers to the spatial scale of a map, and as discussed above, can affect map accuracy. In general, the coarser the resolution of an image, the fewer the objects that can be represented on it. One of the great advantages of GIS is that it allows maps to be displayed at various scales, regardless of the original input resolution (Michelmore, 1994). However, caution should be exercised when making decisions based on these maps and decisions should not be made on maps that have a finer resolution than their original input resolution. The original scale of input data should always be recorded in the database. Modelling is generally carried out using fine resolution data and then validated with low resolution data, which is the strategy adopted here. This can be problematic if only low resolution data are available and Buckland and Elston (1993) have investigated constructing models using a mixture of spatial scales.

Dent *et al* (1989) consider the one by one minute resolution of their surfaces to be optimum, as a coarser resolution would result in loss of data and a finer scale would make data handling

difficult and time-consuming. There is a slight discrepancy of a few pixels in boundary matching between coverages, particularly along the coast; however, this was not deemed to be a major problem for a study at this scale.

3.3.5) Interpolation

Interpolation is a method used to create a continuous surface from a set of scattered data points (O'Conaill *et al*, 1994). There are many different methods used to interpolate surfaces and the one chosen often depends on one's objectives as each method has its own advantages and disadvantages (Isaaks & Srivastava, 1989). IDRISI uses inverse distance weighting to create a continuous surface (Eastman, 1994); i.e. the closer a cell is to a data point, the closer it is in value to it. The interpolation searches for the closest six points to the cell whose value is to be determined and calculates the cell value based on its distance from those points. If more than six points are found within the search radius, then the search radius is temporarily decreased, and if fewer than six points are found, the radius is temporarily increased (Eastman, 1994). An inverse distance exponent of 2 was used. Both of the temperature surfaces and the COV surface were interpolated using this method by running the INTERPOL module in IDRISI. Isaaks and Srivastava (1989) consider inverse distance interpolation to be the best method to use when the objective is to minimize the largest errors.

While it is preferable for the interpolation method to give some estimate of the error (O'Conaill *et al*, 1994), this is not always possible and the surfaces cannot be assumed to be error-free, particularly if the data set is not evenly spaced. In some cases where data points are widely spaced, accumulated error may leave artifacts (in the form of circular patches and swirls on the surfaces). Although the interpolated surfaces used in this study may have a few areas where they are not very accurate, this is unlikely to adversely affect the predictive maps, which were based on country-wide environmental patterns and not on microclimates.

SECTION II

CHAPTER 4
THE RANGE AND INTERQUARTILE RANGE

4.1) INTRODUCTION

Using the range to model distribution is not a new concept. Hutchinson's fundamental niche is essentially defined by the upper and lower abiotic limits within which a species can survive, i.e. its bioclimatic range (Putman & Wratten, 1984). Box *et al* (1993) termed the areas between two climatic extremes a 'climatic envelope' and used this envelope with a median success rate of 85% to 88% to predict plant distribution. Lindenmayer *et al* (1991) used climatic range to determine the distribution of a rare possum and then used the range between the 10th and 90th percentiles to narrow the distribution down to a few 'core' areas. The range also plays an important role in the Bioclimtic Prediction System, BIOCLIM (Chapman & Busby, 1994). BIOCLIM uses climatic parameters to predict the potential distribution of plant species. Amongst its climatic parameters are ranges of temperature and precipitation (Richardson & McMahon, 1992). This system has been used to predict the potential distribution of the pest plant, *Chondrilla juncea*, in western Australia (Panetta & Dodd, 1987) and to identify potential planting regions for *Eucalyptus nitens* in South Africa (Richardson & McMahon, 1992). The range can thus serve as an important modelling and predictive technique, especially since its values are usually easily derived and widely understood.

The range can be defined as the difference between two extreme values (Mowforth, 1979). By extracting all the values for a particular environmental variable from sites where the plant is present, one can determine the minimum and maximum values of that variable. Assuming that these two values represent the extremes of that variable tolerated by the plant, then all the areas with environmental values falling between the two extremes could be regarded as potentially suitable for that particular species. Similarly, areas with environmental values falling outside the range could be regarded as unsuitable habitats for the establishment of the species.

One of the problems in using the range to predict potential distribution is that the range represents extreme values (Gregory, 1973; Mowforth, 1979) and one extreme value can skew the results. The range is also affected by sampling bias. One of the techniques used to dampen this effect is to make use of the interquartile range. The interquartile range is the middle half of the data; in other words, the top 25% and bottom 25% of the values are discarded. This removes the extreme values at either end of the scale.

4.2) METHODS

The maximum and minimum values for the environmental variables (i.e. what was assumed to be the range) for each plant species were calculated by extracting the environmental variables from IDRISI for each site where the plant was recorded as present; putting these variables into a database and then ranking them to obtain the maximum and minimum values.

Each coverage was reclassified using IDRISI to code all the areas between the maximum and minimum values as having a score of one (i.e. areas falling within the range for a particular environmental variable were given a score of one), while all other areas were given a score of zero. These reclassified coverages were then added together using an overlay function. The resulting predictive map has values ranging from zero to five; with zero representing areas where none of the environmental parameters are suitable, and five, areas where all five environmental parameters are within the range tolerated by the plant. Thus the image gives a prediction of areas most suitable for invasion in terms of the number of environmental preferences being met. Predictive maps for the interquartile range were obtained using the same technique of reclassifying and overlaying as above, but using the 25th and 75th percentiles as the minimum and maximum values.

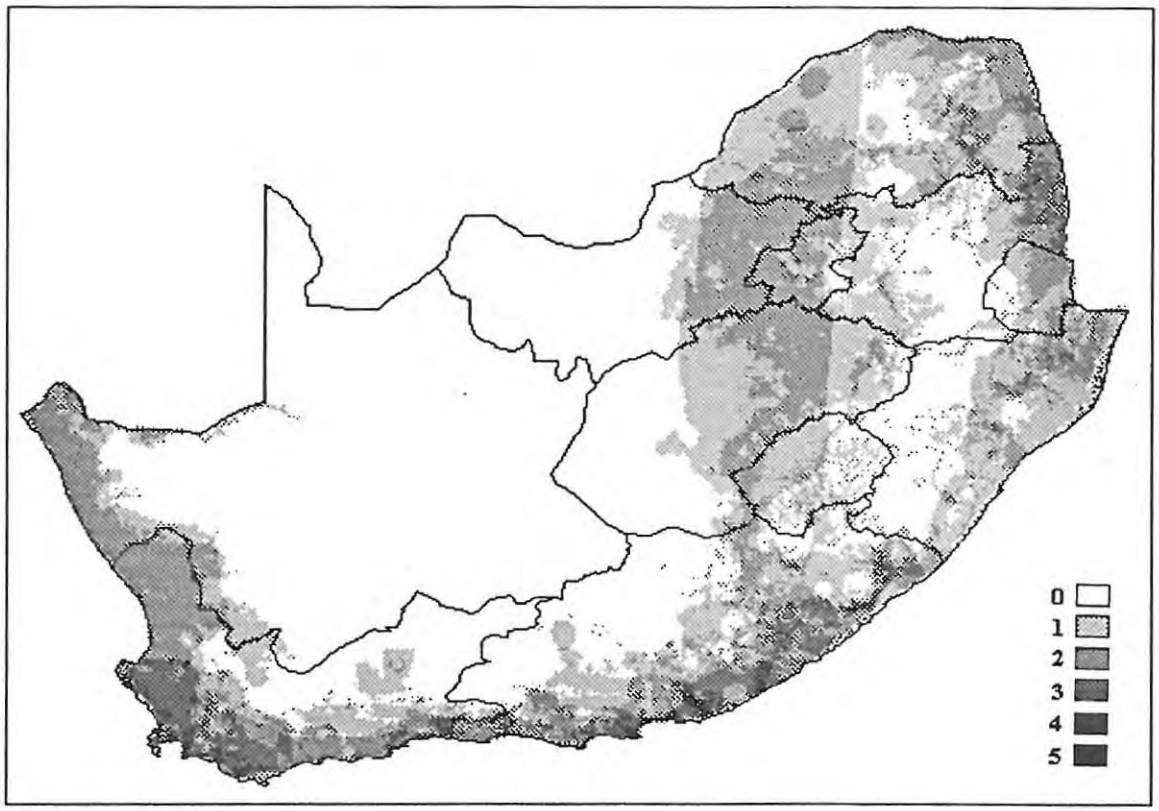
Validation of the accuracy of the predictive maps was done by overlaying the quarter degree records onto the predictive maps. The value on the predictive map was extracted at each quarter degree site to determine if the areas of predicted maximum suitability corresponded to the areas where the plant is known to occur.

Chi-squared tests were performed on the validation results to determine if the predictive maps were significant departures from randomness. The tests took the form of 1 by 2 tables, with a category for correct prediction of actual presence (i.e. where the plant is actually present and is predicted as being present) and a category for incorrect prediction of actual presence (i.e. where the plant is predicted as absent, but is actually present). Thus a statistically significant result either indicates that the map is a good predictor of distribution (many actual presences) or that the map is predicting areas where the plants are not present (i.e. the second category for the chi-squared tests, where the plants are predicted as absent but are not).

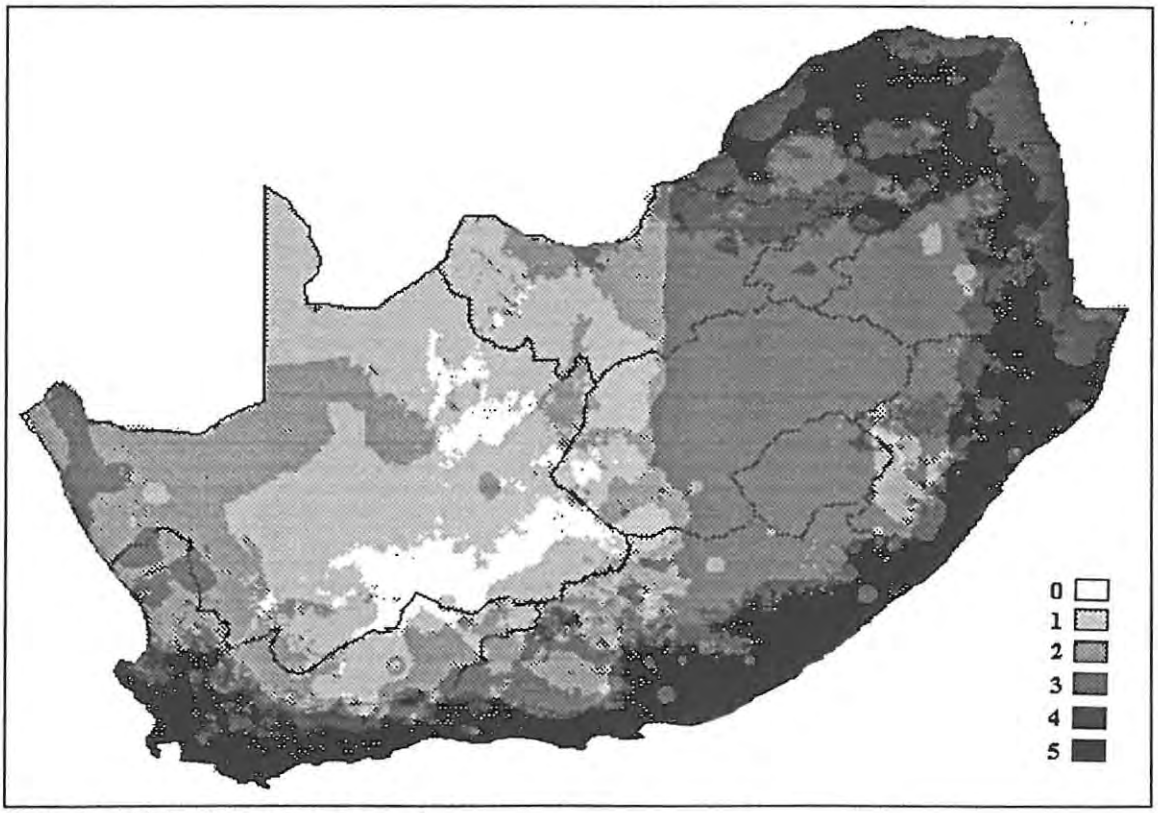
4.3) RESULTS

The maps produced using the range (figure 5a - f) and interquartile range (figure 6a - f) show the environmental suitability of areas for invasion on a graded scale of zero to five, with five representing areas optimally suited to invasion. This scale is represented in shades of grey on the figures, with white representing areas of total unsuitability and black areas of maximum suitability. In general, the darker the shade, the greater the suitability of the area for the species.

The validation results (tables 2 and 3) show the percentage of quarter degree square records falling within each area of the predictive maps derived from the range and interquartile ranges respectively. A large chi-squared result (tables 4 and 5) and a small significance level (below 0.05) indicate that the maps are significant departures from randomness.

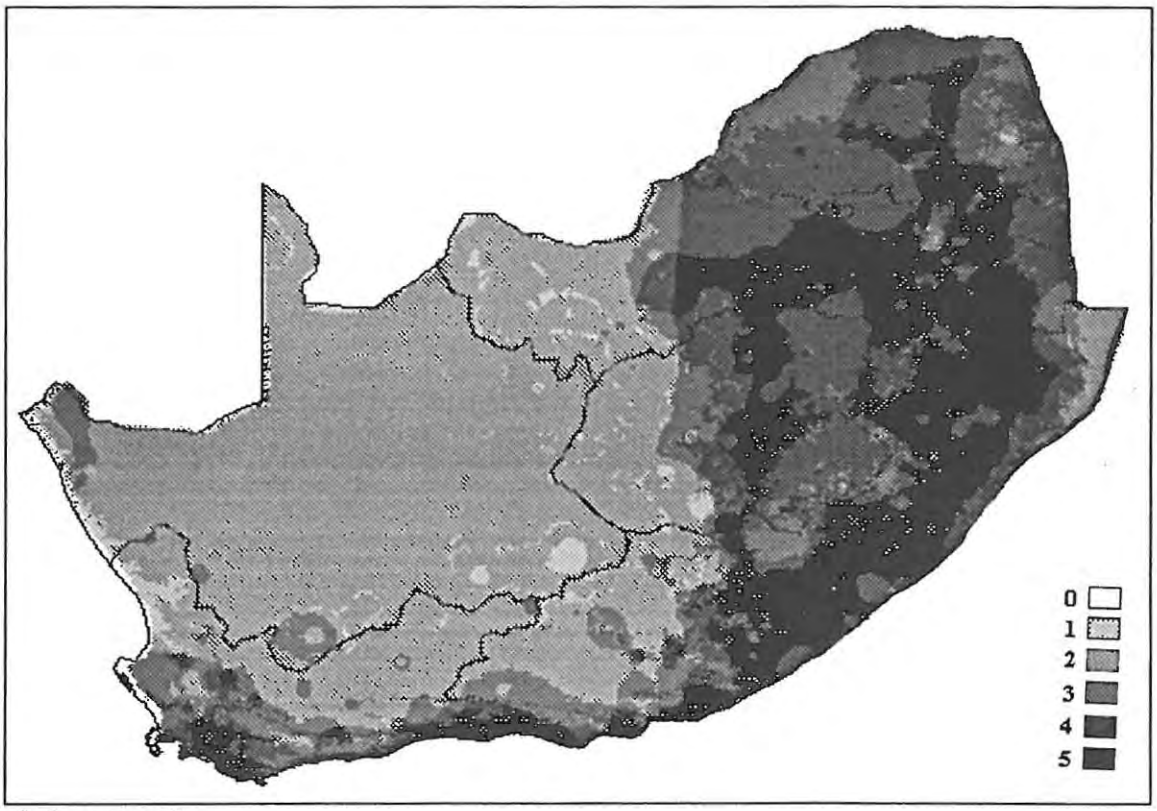


a) *A. longifolia*

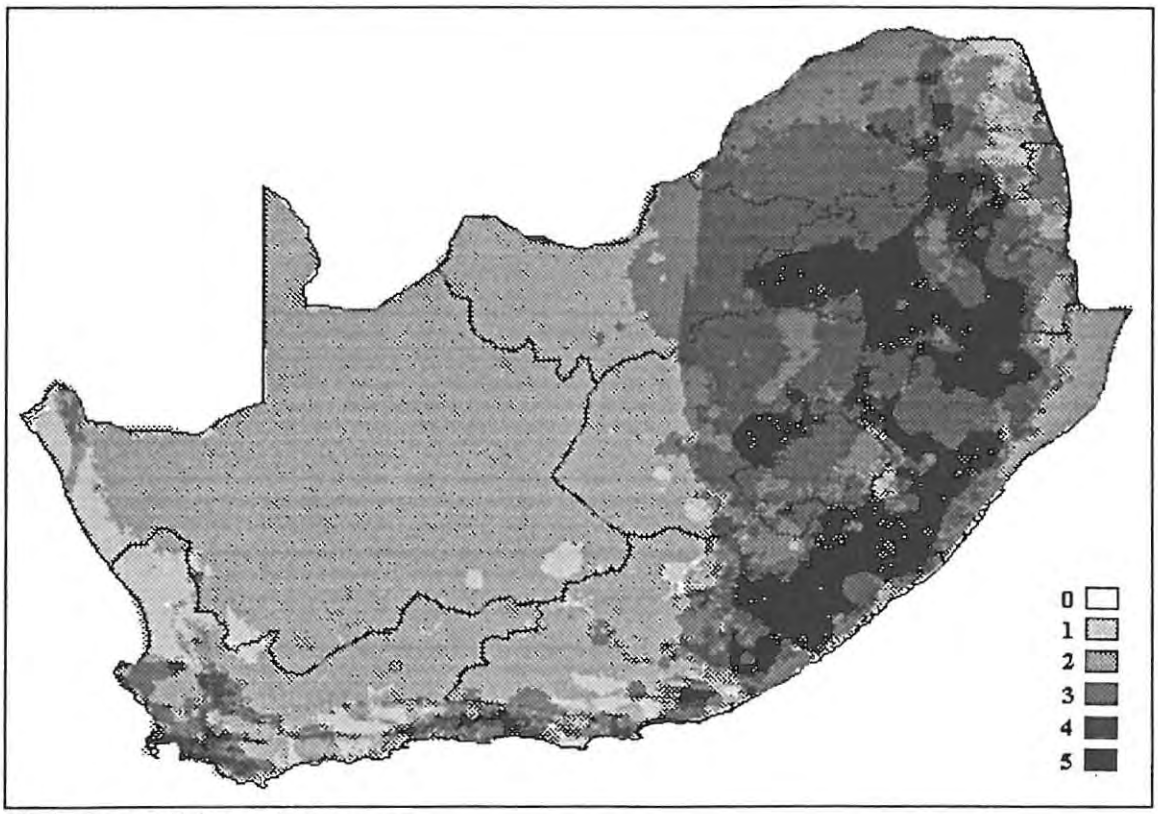


b) *A. longifolia* (extended data set)

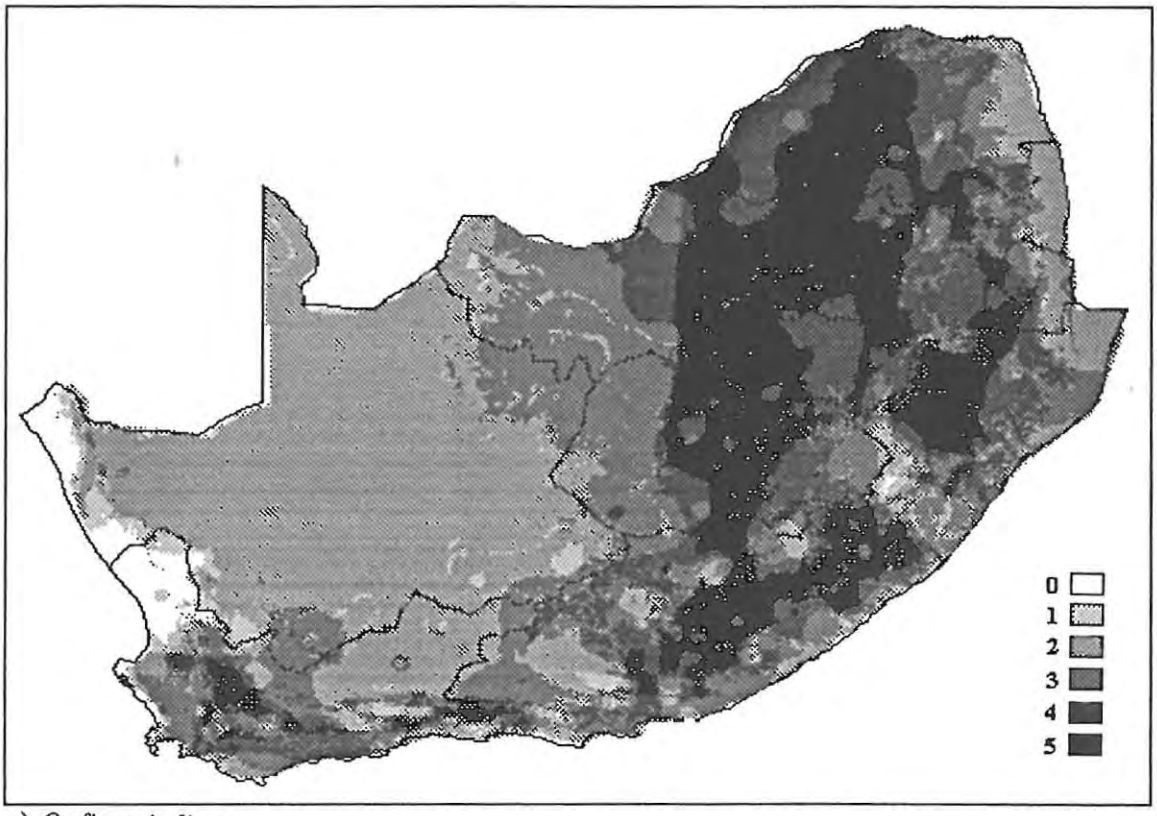
Figure 5. Potential distribution maps derived from the range. The figures represent the number of environmental parameters that are within the range tolerated by the plant. 0 represents areas where none of the parameters are suitable, 2 represents areas where two parameters are suitable, and so on up to 5 which represents areas where all five parameters are suitable.



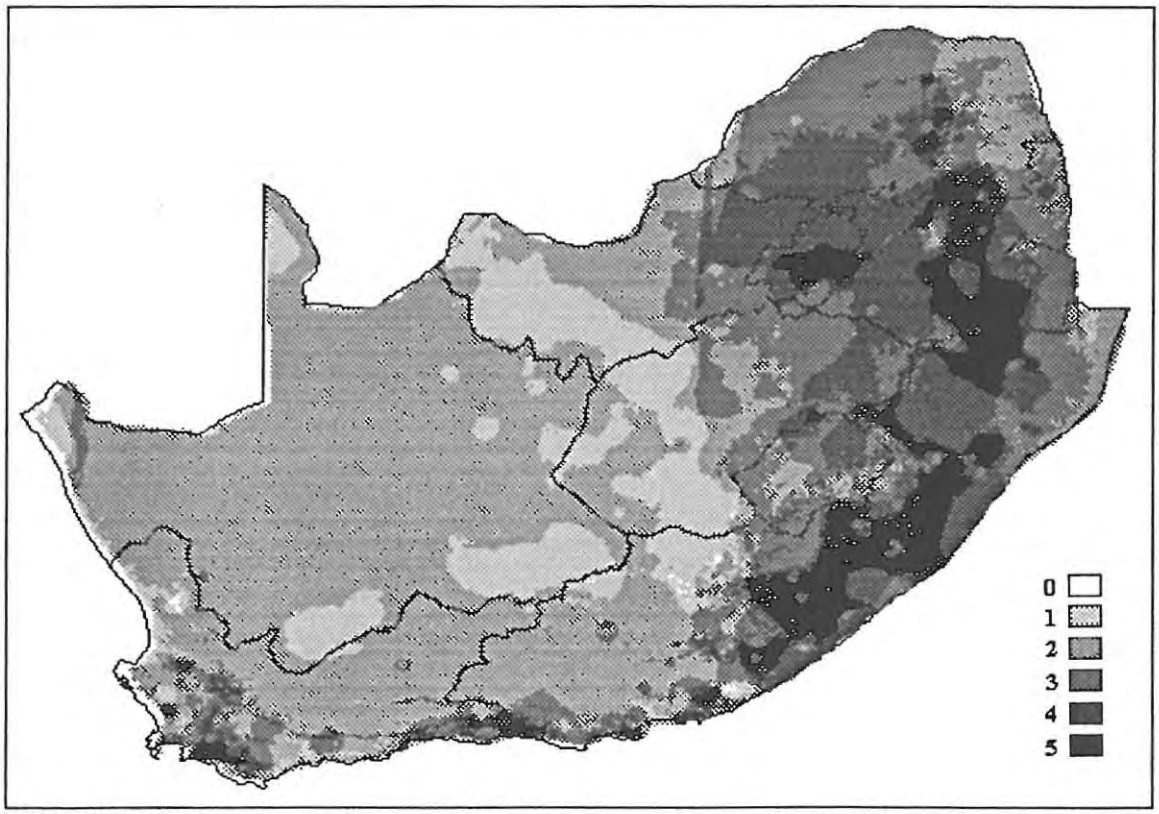
c) *A. mearnsii*



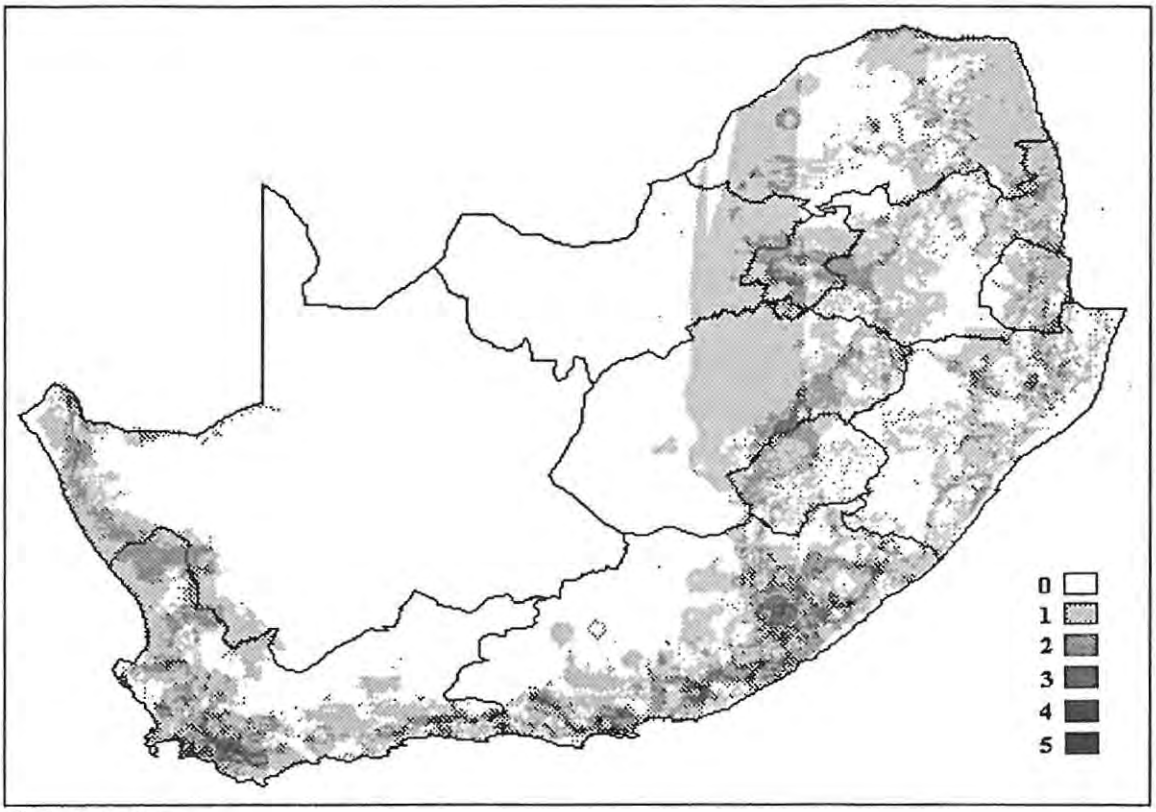
d) *A. mearnsii* (extended data set)



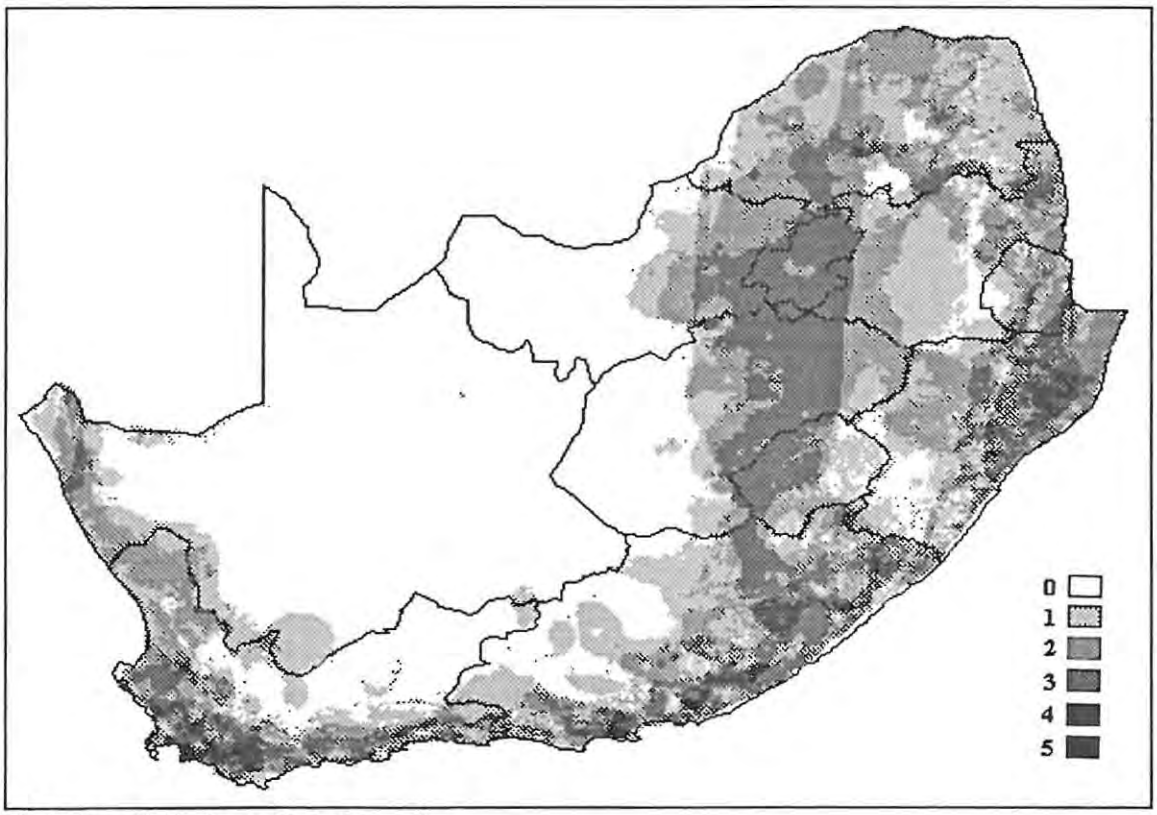
e) *O. ficus-indica*



f) *S. sisymbrifolium*

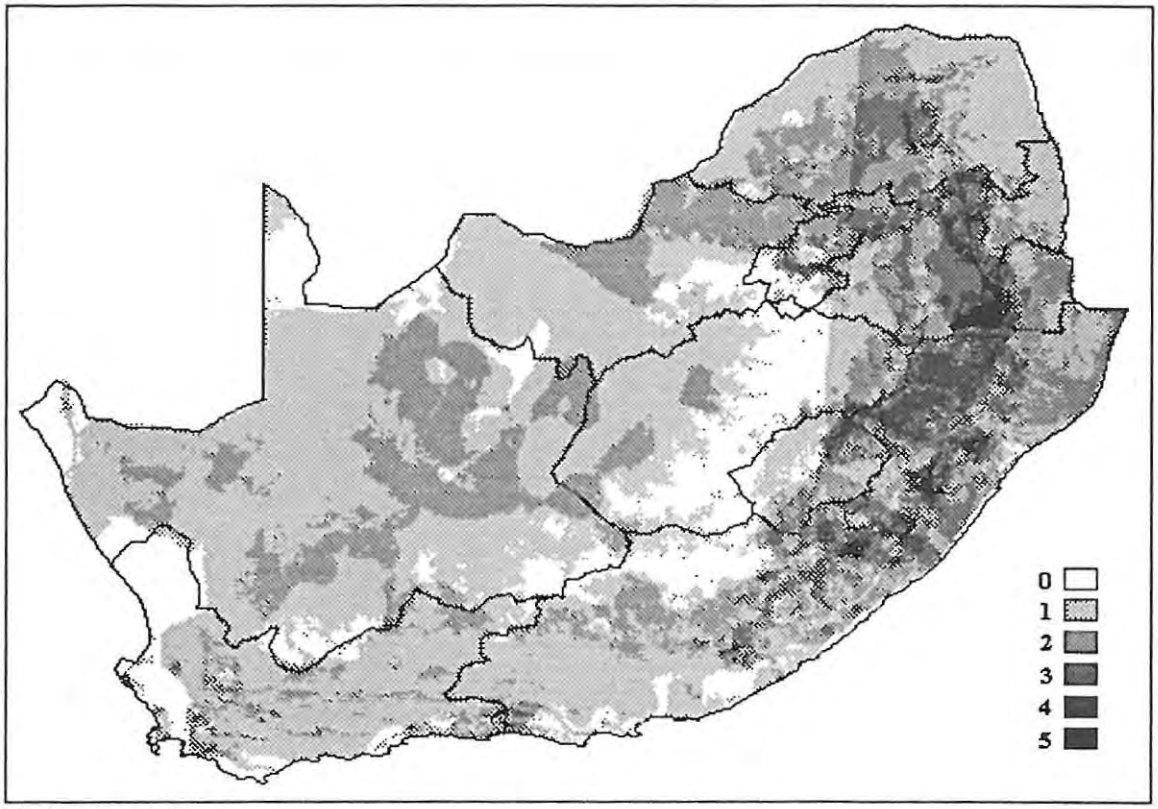


a) *A. longifolia*

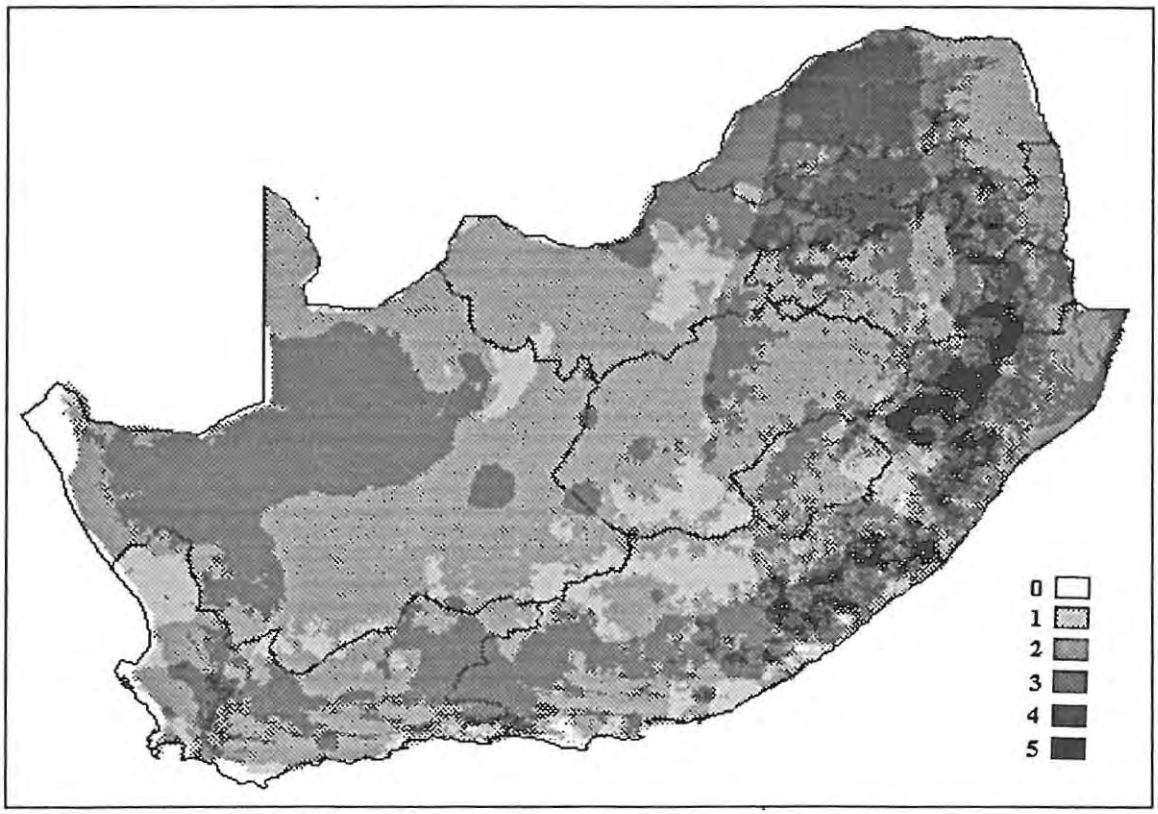


b) *A. longifolia* (extended data set)

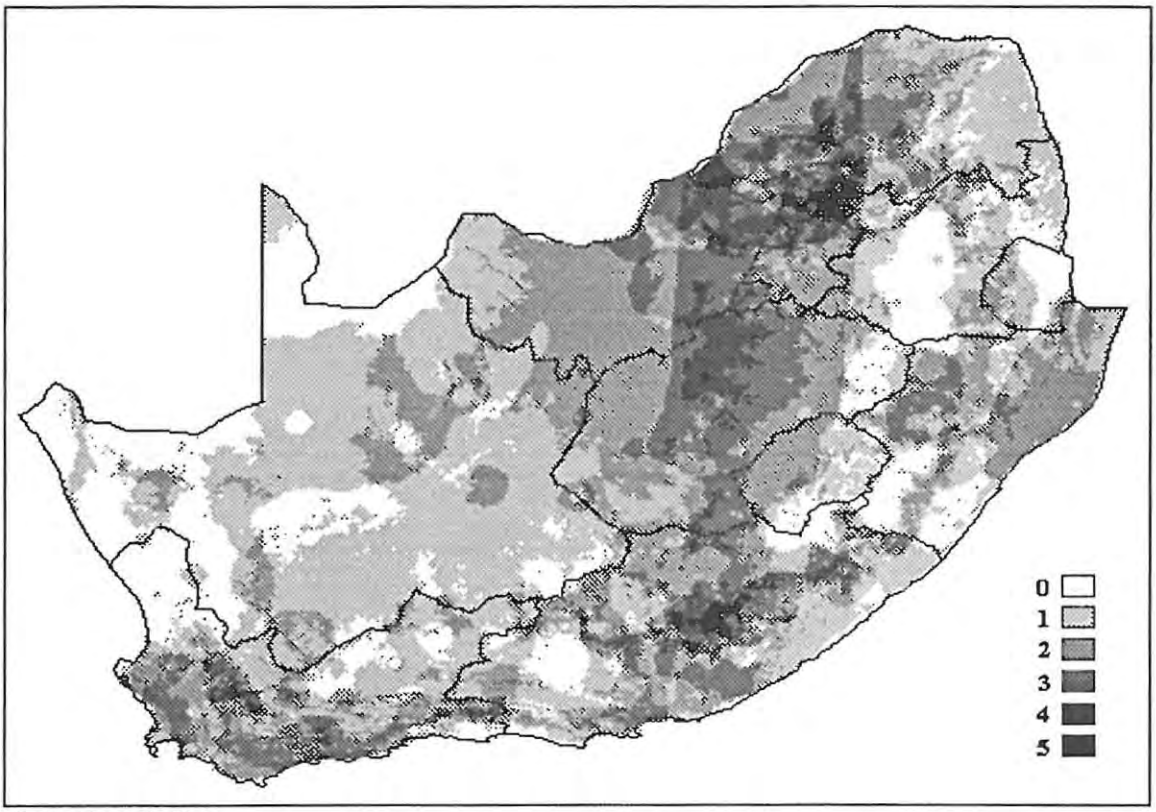
Figure 6. Potential distribution maps derived from the interquartile range. The figures represent the number of environmental parameters that are within the range tolerated by the plant. 0 represents areas where none of the parameters are suitable, 2 represents areas where two parameters are suitable, and so on up to 5 which represents areas where all five parameters are suitable.



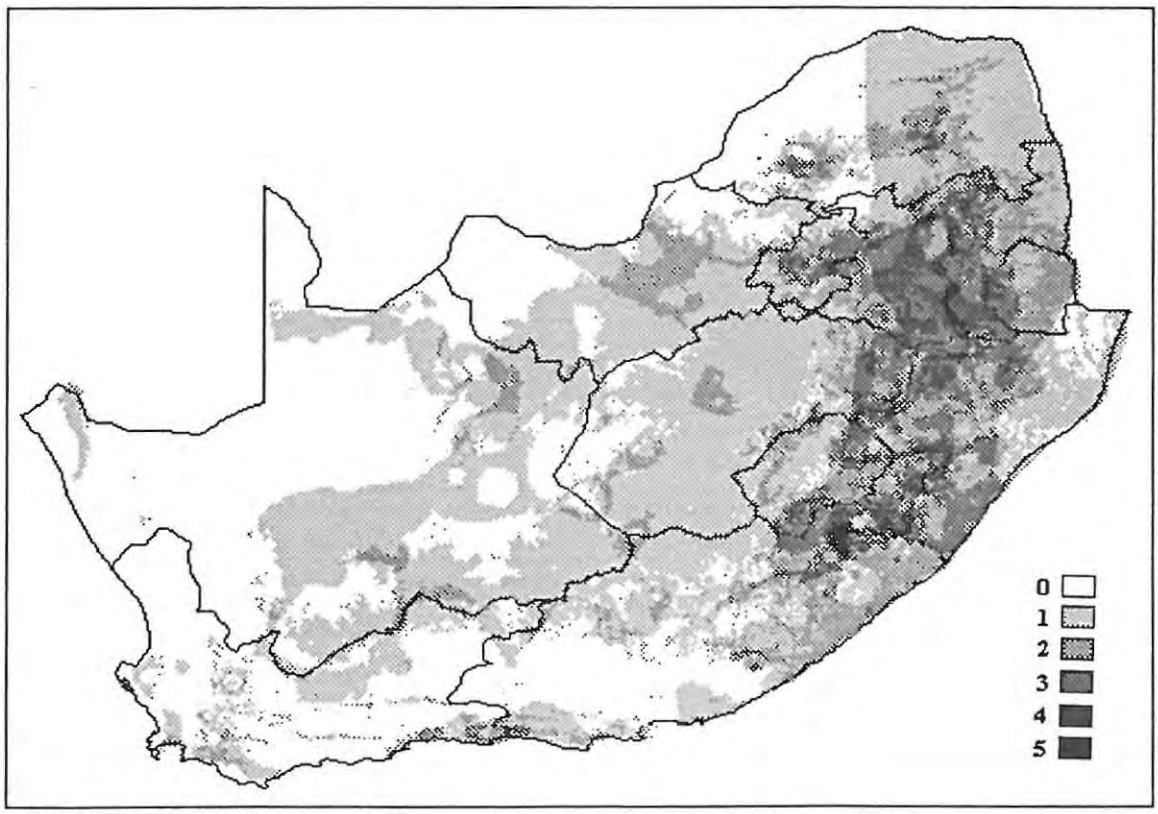
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

Table 2. The percentage of quarter degree square records falling within each area on the range map. a) refers to the small data sets and b) to the extended data sets. The numbers 0 to 5 refer to the coding on the predictive map, with 0 representing areas where none of the environmental values fell into the range (or interquartile range) tolerated by the species, 1 representing areas where one environmental variable was within range, and so forth, with 5 representing areas where all five of the parameters were within range (i.e. areas of maximum suitability according to environmental preferences).

Species		0	1	2	3	4	5	Total
A. longifolia	a)	20	28	24	14	12	2	100
	b)	14	2	0	6	24	54	100
A. mearnsii	a)	6	3	12	14	30	35	100
	b)	3	4	4	7	22	60	100
O. ficus-indica		0	4	13	24	24	35	100
S. sisymbrifolium		0	3	11	30	26	30	100

Table 3. The percentage of quarter degree square records falling within each area on the interquartile range map. a) refers to the small data sets and b) to the extended data sets. The numbers 0 to 5 refer to the coding on the predictive map, with 0 representing areas where none of the environmental values fell into the range (or interquartile range) tolerated by the species, 1 representing areas where one environmental variable was within range, and so forth, with 5 representing areas where all five of the parameters were within range (i.e. areas of maximum suitability according to environmental preferences).

Species		0	1	2	3	4	5	Total
A. longifolia	a)	30	36	24	6	2	2	100
	b)	28	18	20	18	10	6	100
A. mearnsii	a)	16	4	21	43	11	5	100
	b)	7	7	28	33	19	6	100
O. ficus-indica		11	23	32	24	10	0	100
S. sisymbrifolium		11	37	22	19	7	4	100

Table 4. Chi-squared results and significance levels for the range predictive maps. a) refers to the small data sets and b) to the extended ones.

		Chi-squared result	Significance level
A. longifolia	a)	48.08	0.0000
	b)	0.32	0.5716
A. mearnsii	a)	23.07	0.0000
	b)	10.93	0.0009
O. ficus-indica		41.80	0.0000
S. sisymbrifolium		4.48	0.0343

Table 5. Chi-squared results and significance levels for the interquartile range predictive maps. a) refers to the small data sets and b) to the extended ones.

		Chi-squared result	Significance level
A. longifolia	a)	46.08	0.0000
	b)	38.72	0.0000
A. mearnsii	a)	207.63	0.0000
	b)	196.98	0.0000
O. ficus-indica		404.60	0.0000
S. sisymbrifolium		23.15	0.0000

4.4) DISCUSSION

The range maps indicate that all of the species show a preference for the mid-eastern parts of the country and the fynbos region. KwaZulu-Natal and Mpumalanga appear to be environmentally suited to the growth of *A. mearnsii* (figure 5c, d). *Acacia longifolia* shows a fairly limited distribution and does not appear to enjoy very high altitude areas such as the Drakensberg, or the more arid areas, which is not surprising as this species occupies low-lying, temperate areas in its native country (Stirton, 1987).

Although the maps for *A. mearnsii*, *O. ficus-indica* and *S. sisymbrifolium* depict substantial areas where two or three environmental parameters are suitable for the plant, it is likely that only the areas where all five environmental parameters are within the range tolerated by the

species that will prove to be suitable for invasion. Box *et al* (1993: 629) based his climatic envelope model on the assumption that "a species will occur at a site as long as none of the species' climatic limits is exceeded by the local climatic data." Thus, if even one of the bioclimatic variables is not within the range tolerated by the species, then the area is unlikely to be suitable for invasion.

However, areas where four of the five parameters are within the range tolerated by the plant may prove marginally suitable for invasion or may function as invasion corridors along which invasion can spread to other areas. Alternatively, the species may occur there, but in very low densities, or may be constantly invading the area, but is unable to gain a foothold (Rogers & Williams, 1993).

While one would expect the interquartile range to predict only the areas of maximum suitability for invasion by removing the extremes for each environmental parameter, as Lindenmayer *et al* (1993) did in their study, this did not prove to be the case here. The interquartile ranges for all four species (figure 6a - f) show very restricted distributions to the maps produced from the range, with very few areas of maximum suitability being predicted. Certain areas on the maps where the plants were known to occur showed that the likelihood of the plant occurring there was non-existent. For example, *S. sisymbrifolium* is know to occur at Grahamstown, yet the map indicates that this area is not at all climatically suitable for this species. This may be due to the fact that most of the data sets for presence were small (only 9 recorded presences in one instance). The cutting out of 50% of the data for the interquartile range reduced the data set to one not large enough to predict distribution accurately.

With regards to the validation results, only the two extended data sets for the predictive range maps demonstrated a greater than 50% predictive success rate (table 2). The chi-squared results however, indicate that the predictive map for one of the extended data sets (for *A. longifolia*) is not significant. The remaining chi-squared results (table 5) indicate that the maps are significant departures from randomness, implying that these maps are statistically poor predictors of true presence, with the possible exception of the map produced for the extended data set for *A. mearnsii*. The chi-squared results for the interquartile range maps are all significant (table 5) implying that these maps are all statistically poor predictors of presence.

For both the range and the interquartile range, an increase in sample size resulted in an increase in the percentage of quarter degree square sites falling into the areas of predicted maximum suitability. This improvement in the predictive success of the coverages with the addition of more data points may be due to the larger sample sizes being more representative of the total population. This appears to be especially true for the interquartile range. For example, despite the relatively small sample size (n = 12) for *O. ficus-indica*, the range map achieved prediction success equal to that for the small data set (n = 25) for *A. mearnsii* and better than that of the small data set (n = 14) for *A. longifolia* (table 1). This could be attributed to the fact that *O. ficus-indica* has had enough time to establish within its full range in South Africa. However, the predictive interquartile map for this species was not at all successful, perhaps because the sample size was too small to be representative of the interquartile range of the species.

4.5) CONCLUSIONS

Using the minimum and maximum values of the environmental parameters (the 'climatic envelope') to predict the potential distribution of the species does not appear to be very successful. This may be due to the relatively small sample sizes used for this study, the interquartile range in particular appearing to require larger sample sizes than those used here to produce accurate predictive maps. The length of time that a species has had to establish itself in its host country also appears to be of importance; generally, the longer the plant has been in its host country, the more likely it is to have established within its full bioclimatic range. This implies that there may be a better chance that the range obtained from sample sites for the species will be representative of its true environmental range.

CHAPTER 5
PRINCIPAL COMPONENTS ANALYSIS

5.1) INTRODUCTION

Principal components analysis (PCA) is a multivariate statistical technique that determines uncorrelated linear combinations of variables that explain the variability in a data set. The data are not transformed, but are simply restated (Jackson, 1983) with the first principal component being the axis that summarizes most of the variability. If many original variables in the data set can be represented by only two or three principal components (PCs) then there may be redundancy in the data, i.e. most of the variables measure similar things (Manly, 1986) and the principal components adding little to the data set may be discarded. Thus principal components analysis may be used to reduce the dimensionality of a data set. Similarly, PCA can be used to determine the most important unrelated variables for distinguishing between groups (Jeffers, 1967).

Principal component analysis has many applications. One of its most common uses is to reduce a data set by determining linear combinations of the variables that explain most of the variation in the data set. This can help to determine the most important variables (or groups of variables) for a particular study. For example, Jeffers (1967) used PCA to determine the factors important in evaluating the strength of home-grown timber for pitprops and in distinguishing aphids into four groups.

PCA may also be used for modelling purposes. Buckland & Elston (1993) performed a linear multiple regression on principal component scores to model the distribution of red deer and two bird species. Wong (1968) used PCA to help construct a multiple regression model to predict mean annual flood in New England.

Sometimes PCA is used for both modelling and data reduction as in the cases of Osborne and Tigar (1992) and Menozzi *et al* (1978). Both of these studies used PCA to reduce their data sets and then used the PCA scores to model distribution.

Whilst in the context of this study PCA was not used with the express purpose of reducing the data set, it did indicate which variables were possibly redundant. Its main role was to indicate the environmental variables important for each of the four species for prediction purposes. It should be borne in mind that the predictions are made using the environmental factors only and not the plants' distributions. One of the ways in which distribution may be incorporated in the analysis is by using the co-ordinates of the sites as variables in the PCA (Buckland & Elston, 1993; Osborne & Tigar, 1992). This method was not used in this study for technical software reasons.

5.2) METHODS

IDRISI has a principal components analysis module. This was used to extract the first three standardized principal components from the five coverages. The median annual rainfall (MAR), elevation (ELV), maximum temperature (MAXT) and minimum temperature (MINT) coverages had to be rescaled to between 0 and 255 before a PCA could be run on them as the PCA would only accept byte binary data (i.e. negative values and values over 255 have to be rescaled). The inter-relationships of the environmental variables are given by a correlation matrix, which is automatically generated by the PCA routine in IDRISI (Eastman, 1994).

The PCA scores for each presence site (high resolution data) were extracted from each of the three principle component coverages through use of the QUERY option in IDRISI. These values were ranked and the range (i.e. maximum and minimum value) for each component for each species was obtained. These ranges were used to reclassify the principal component coverages into areas of presence and absence according to the presence data.

Validation was carried out by overlaying the validation data set onto the predictive maps. The value on the predictive maps for each pixel where the plant was present on the validation data set was then extracted and the percentage of sites correctly predicted as present was calculated.

Two by two chi-squared tests were calculated to determine which of the predictive maps were significantly different from ones that would be randomly produced. The chi-squared tests also give an indication of whether the maps are significantly predicting actual presence or whether their significance is due to the prediction of areas of false presences (i.e. areas where the plant is predicted to be present but is actually absent).

5.3) RESULTS

The percentage variance explained by each principal component (table 6) gives an indication of how much redundancy there is in the data. If the first principal component explains most of the variability, then many of the variables measure the same thing (Manly, 1986).

Table 6. The percentage variance and cumulative total expressed by each principal component.

	Percentage variance	Cumulative total
PC 1	63.55	63.55%
PC 2	18.88	82.43%
PC 3	14.71	97.14%

IDRISI calculates the weights (co-efficients) that are attributed to each input variable for each of the PCA axes (table 7). These weights give an idea of which variable is important in explaining most of the variability for each principal component. Inter-relationships between the variables can be indicated by means of correlation; a high correlation co-efficient (either positive or negative) indicates that two variables are highly inter-related (table 8).

Table 7. The weightings of the input variables for the principal components.

	MAR	COV	MAXT	MINT	ELV
PC 1	0.71	0.90	0.94	0.54	0.84
PC 2	-0.24	0.20	0.09	-0.82	0.41
PC 3	-0.64	0.36	0.30	0.17	-0.28

Table 8. Correlation matrix of the environmental variables median annual rainfall (MAR), coefficient of variation for rainfall (COV), mean maximum temperature (MAXT), mean minimum temperature (MINT) and elevation (ELV).

	MAR	COV	MAXT	MINT	ELV
MAR	1	0.38	0.46	0.45	0.64
COV		1	0.95	0.37	0.71
MAXT			1	0.46	0.72
MINT				1	0.09
ELV					1

A predictive map using each of the three principal components was produced for the four species (figures 7, 8, 9, a - f). The three principal components themselves are depicted in figures 10, 11 and 12.

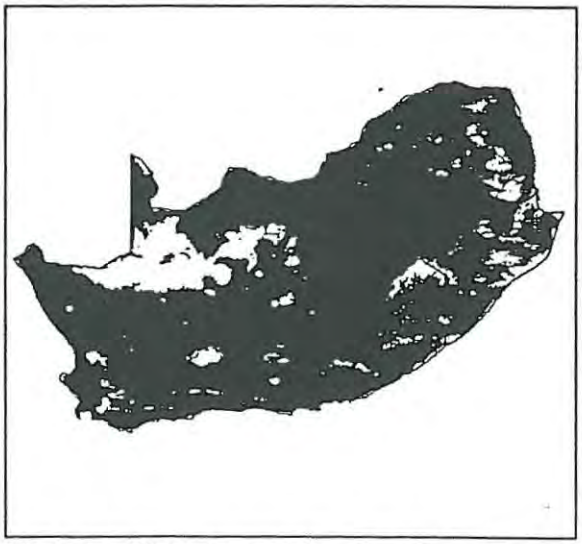
The results of the validation of the predictive maps for each component are expressed as percentages (table 9). The percentages were calculated as the number of quarter degree squares correctly classified as present on the predictive maps.

Chi-squared results with a significance level of less than 0.05 were taken to indicate predictive maps that showed a significant departure from randomness (tables 10, 11 and 12).

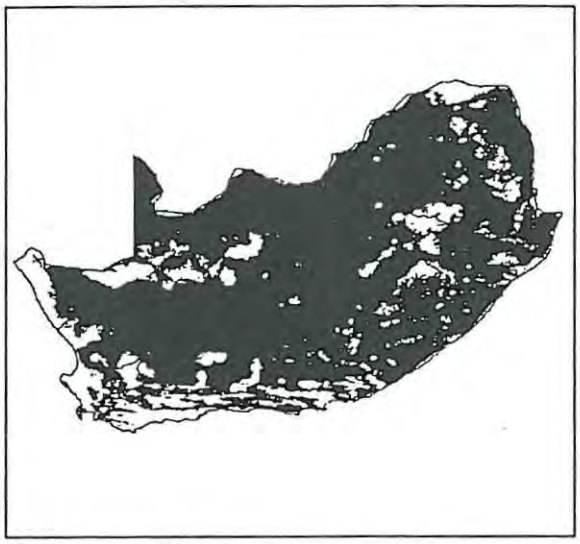
Table 9. The percentage of sites correctly classified as present. a) refers to the small data sets and b) to the extended data sets.

		PC 1	PC 2	PC 3
A. longifolia	a)	72	48	34
	b)	78	86	86
A. mearnsii	a)	68	72	72
	b)	77	93	80
O. ficus-indica		69	81	80
S. sisymbrifolium		96	44	33

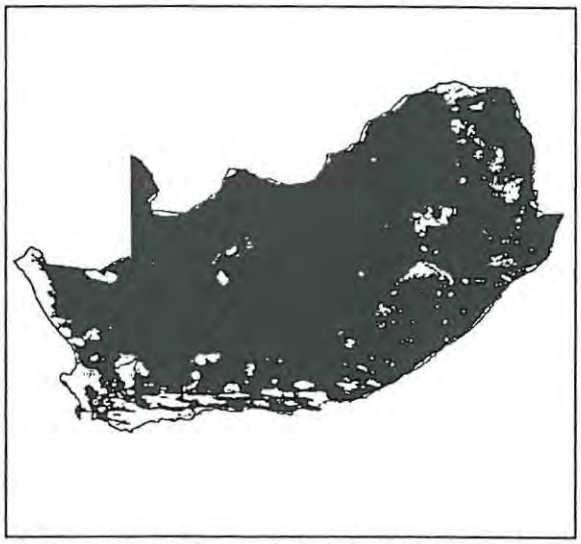
a) *A. longifolia*



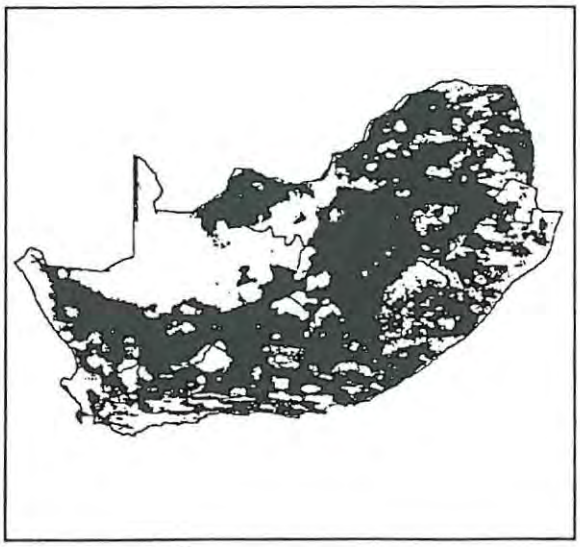
b) *A. longifolia* (extended data set)



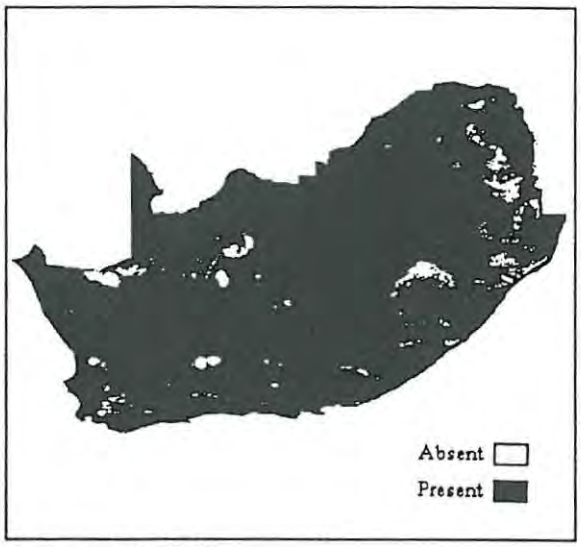
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)

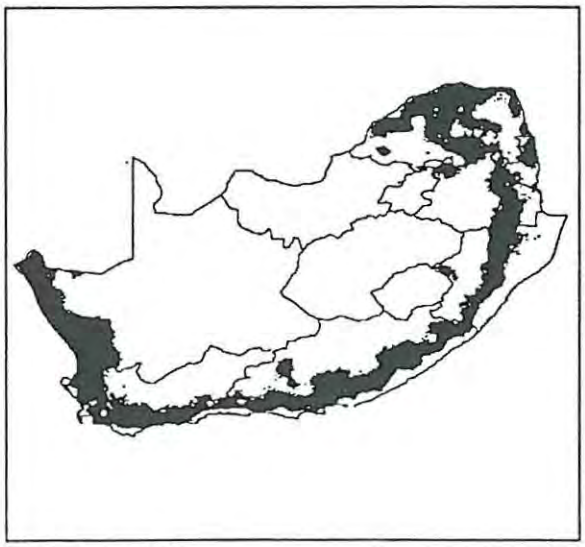


e) *O. ficus-indica*

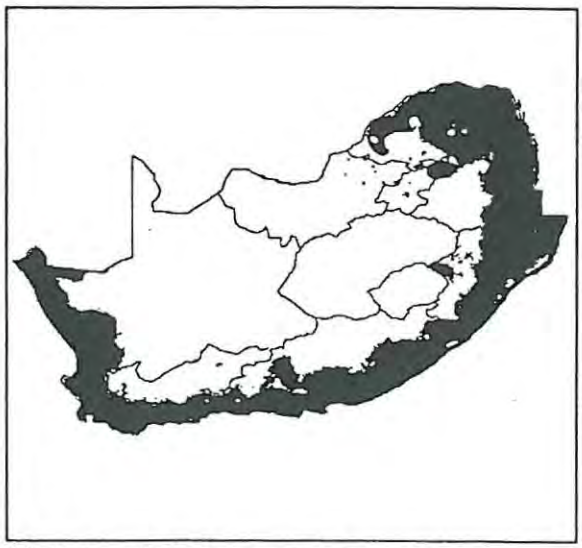


f) *S. sisymbrifolium*

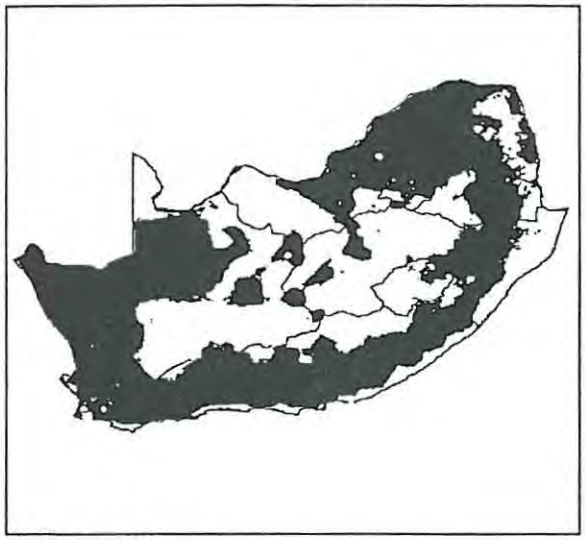
Figure 7. Potential distribution maps derived from the first principal component.



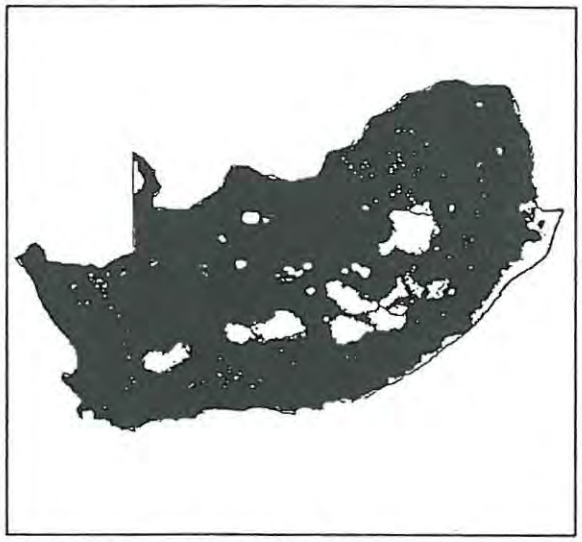
a) *A. longifolia*



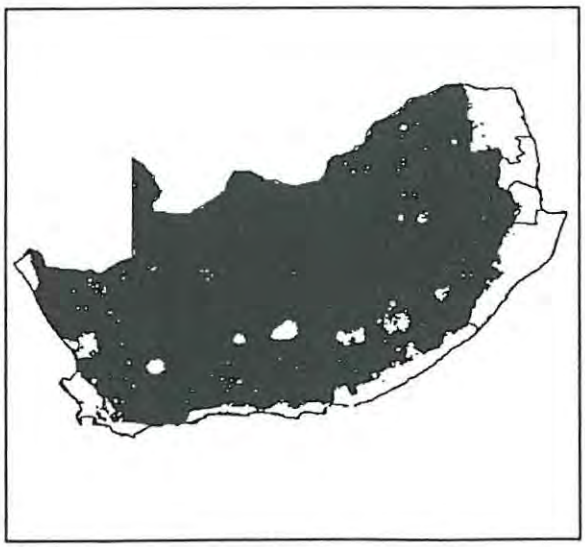
b) *A. longifolia* (extended data set)



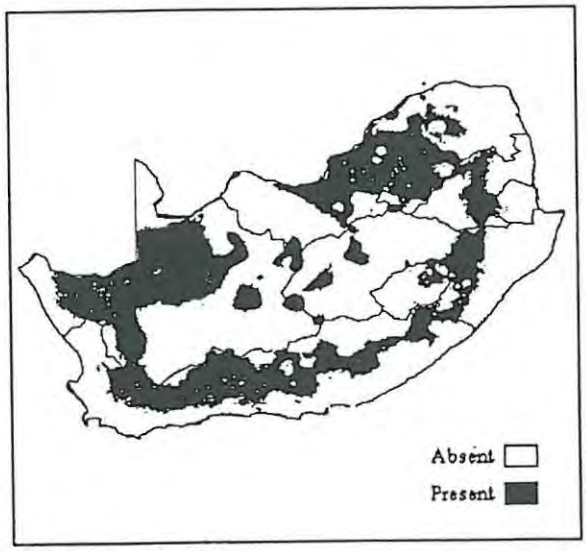
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)

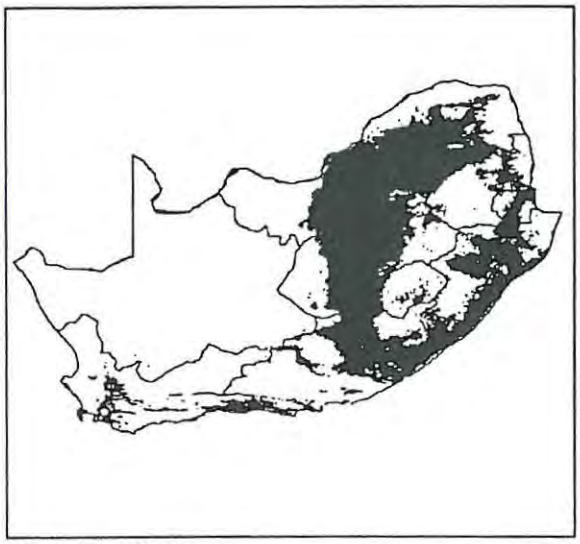


e) *O. ficus-indica*

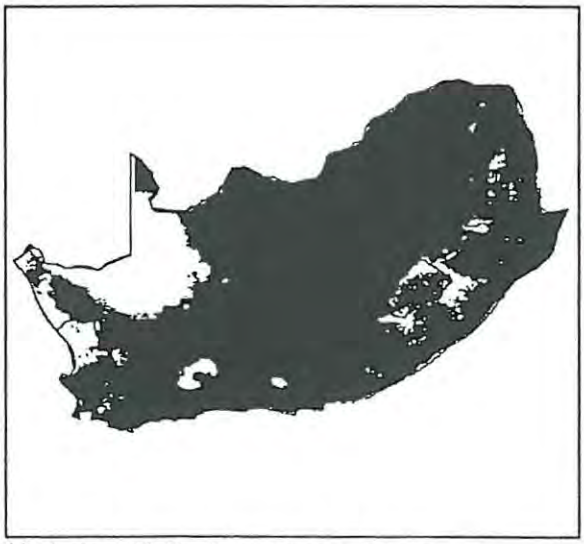


f) *S. sisymbrifolium*

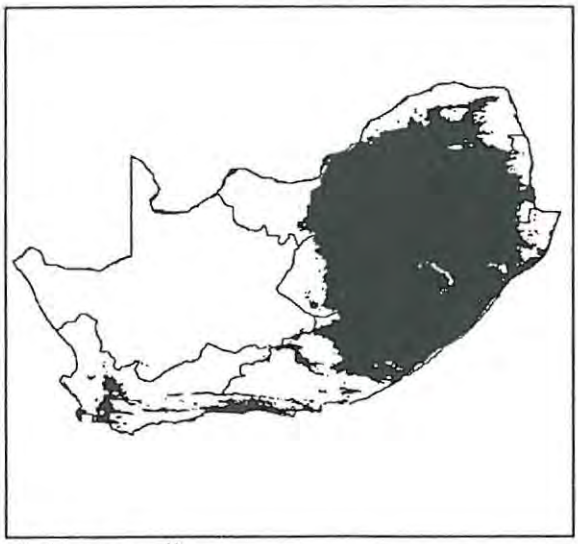
Figure 8. Potential distribution maps derived from the second principal component.



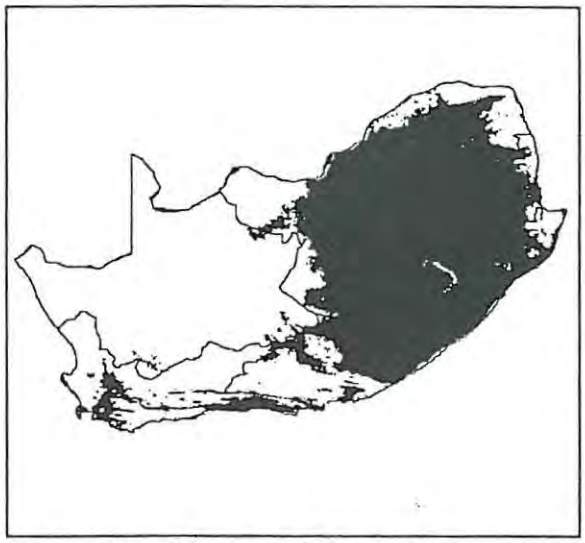
a) *A. longifolia*



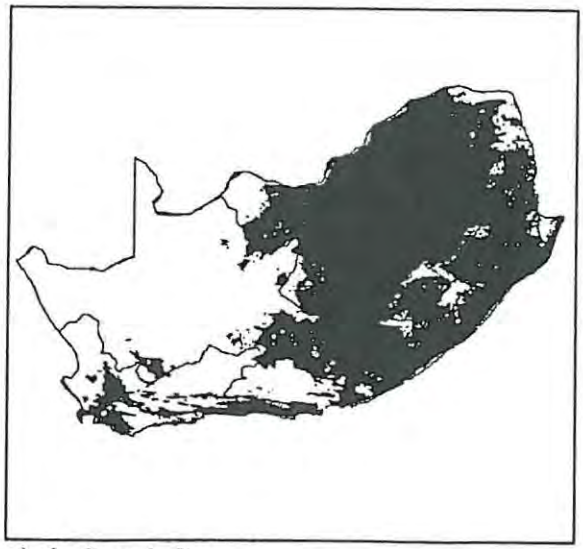
b) *A. longifolia* (extended data set)



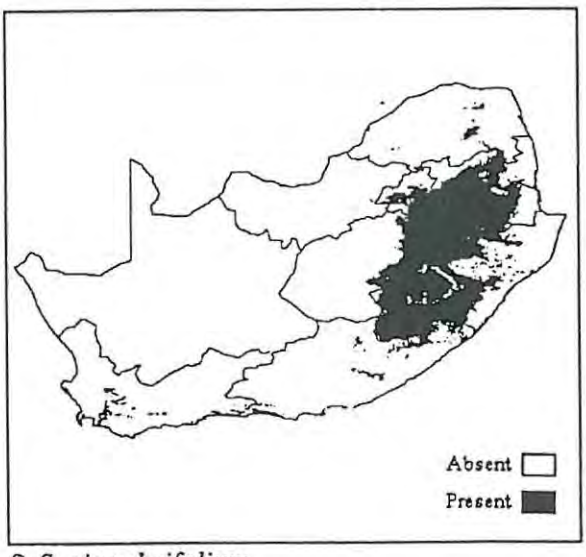
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

Figure 9. Potential distribution maps derived from the third principal component.

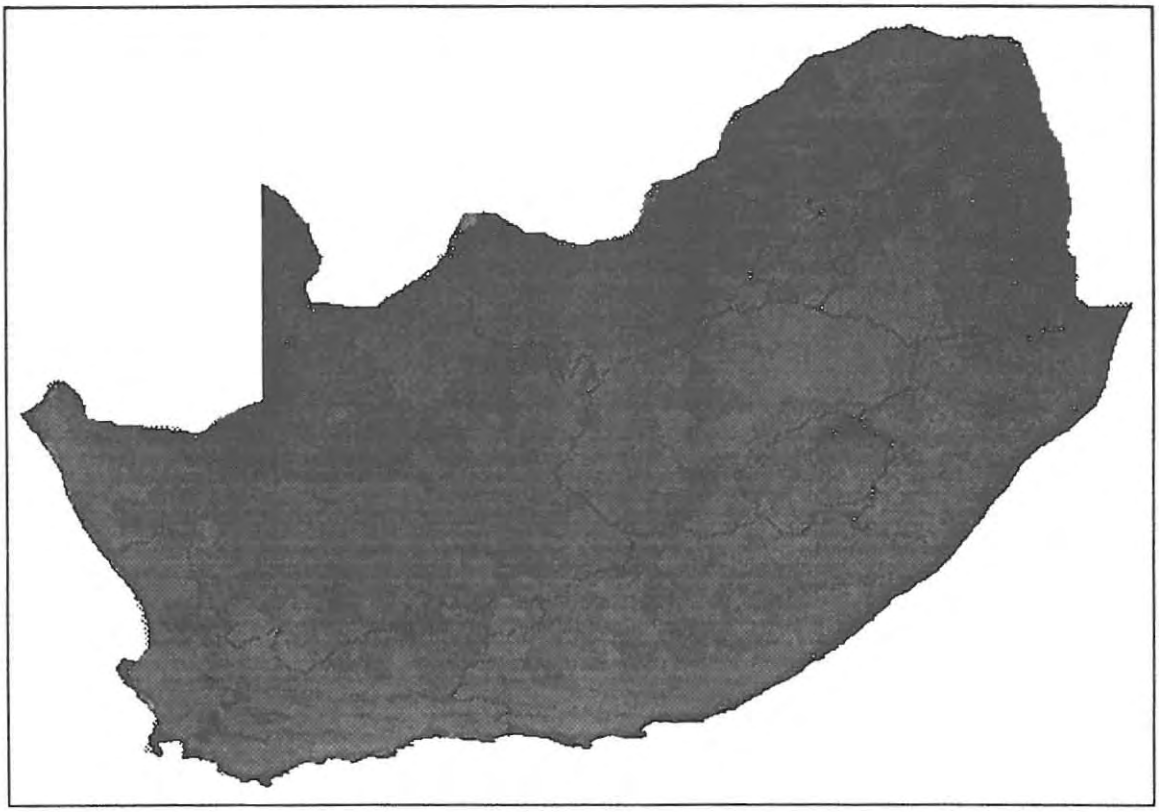


Figure 10. Image of the first principal component.

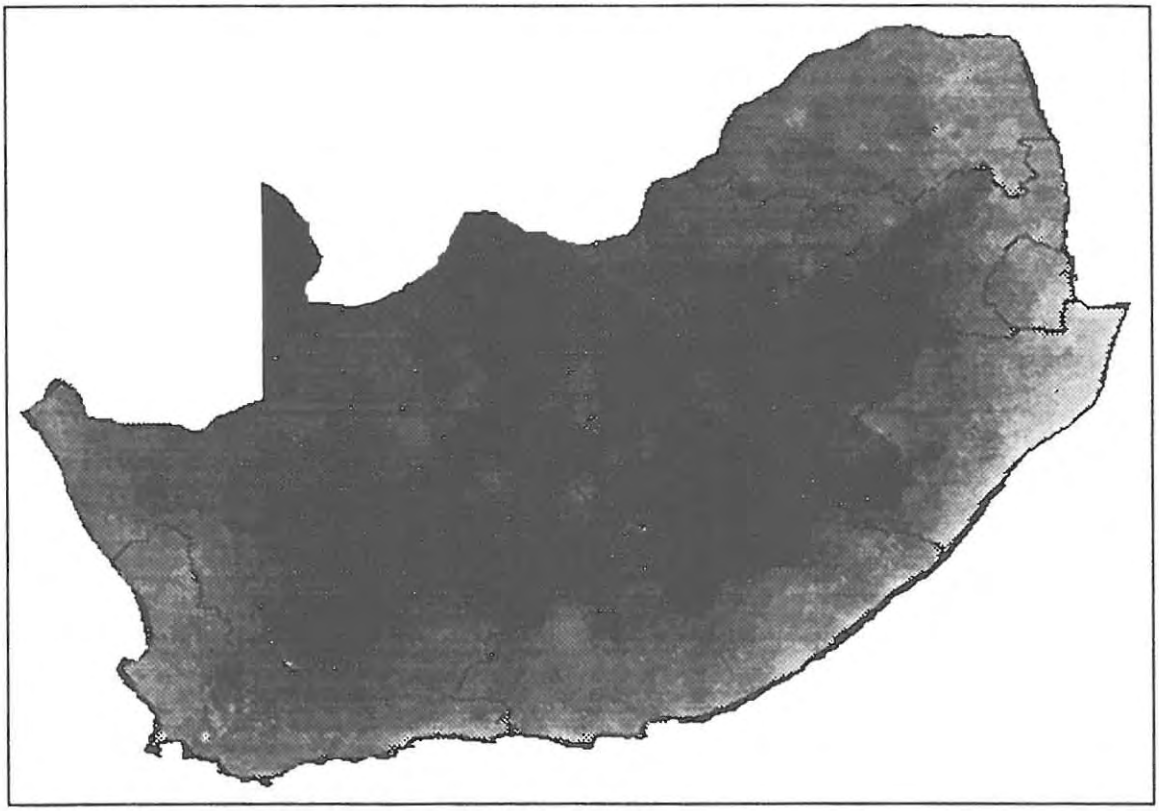


Figure 11. Image of the second principal component.

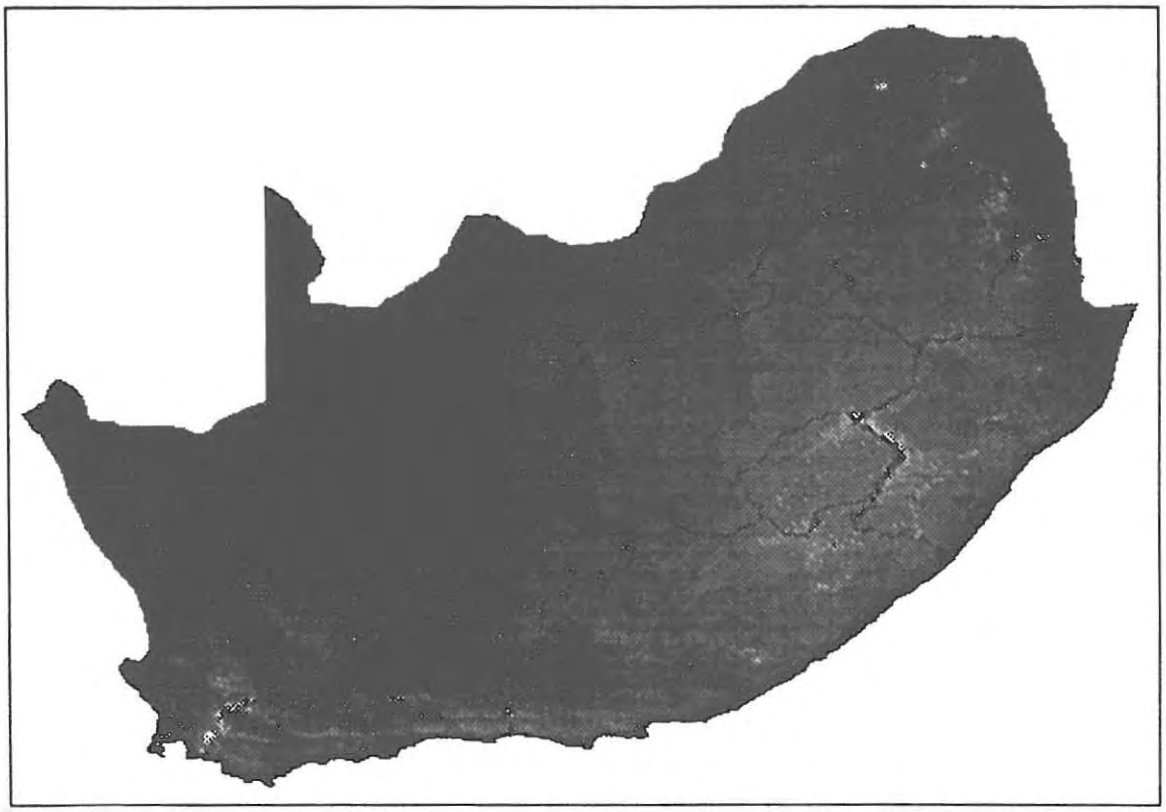


Figure 12. Image of the third principal component.

Table 10. Chi-squared results and significance levels for the predicted maps produced from the first principal component. a) refers to the small data sets and b) to the extended data sets.

		Chi-squared result	Significance level
A. longifolia	a)	8.18	0.0427
	b)	2.14	0.3429
A. mearnsii	a)	40.44	0.0000
	b)	54.80	0.0000
O. ficus-indica		6.22	0.1013
S. sisymbrifolium		0.01	0.9950

Table 11. Chi-squared results and significance levels for the predicted maps produced from the second principal component. a) refers to the small data sets and b) to the extended data sets.

		Chi-squared result	Significance level
A. longifolia	a)	86.14	0.0000
	b)	78.09	0.0000
A. mearnsii	a)	10.87	0.0124
	b)	30.31	0.0000
O. ficus-indica		10.79	0.0129
S. sisymbrifolium		0.07	0.9956

Table 12. Chi-squared results and significance levels for the predicted maps produced from the third principal component. a) refers to the small data sets and b) to the extended data sets.

		Chi-squared result	Significance level
A. longifolia	a)	2.13	0.5456
	b)	4.56	0.2068
A. mearnsii	a)	67.43	0.0000
	b)	48.37	0.0000
O. ficus-indica		2.43	0.4880
S. sisymbrifolium		4.05	0.2557

5.4) DISCUSSION

Almost 65% of the information is summarized in the first component, but components 2 and 3 still retain enough of the environmental variability to make useful predictions.

The number of variables used for prediction are sufficient. The first component does not summarize all of the variability, so one can assume that not all the variables measure the same thing; but at the same time, the measure of variability given by the first principal component is not so small as to suggest that not enough variables were chosen (Manly, 1986).

Each environmental input is weighted by the PCA, giving an indication of which variable accounts for the most variation for a particular principal component. According to the weightings, the first PC relates to a combination of MAXT, COV and ELV; the second to MINT and the third to MAR. The correlation co-efficients also show a strong relationship between MAXT, COV and ELV, indicating that there may be some redundancy amongst these three variables that form the first PC. The second and third PC axes do not show as much redundancy (table 7), although the correlation co-efficients (table 8) indicate that there is a moderate degree of correlation between MAR and ELV.

The principal component coverages (figures 10, 11, 12) demonstrate visually how the variables that are explained by each principal component are distributed in South Africa. By examining these coverages and their patterns, one can gain an idea of how the distribution of the species under each principal component will be affected. For example, if a species shows sensitivity to minimum temperature, which is explained by PC 2, through examination of the second principal component coverage one would expect the species to be present along the coastal areas and in the north-eastern parts of the country (light areas on the coverage). This is the case for *A. longifolia* (figure 8b), but not for *O. ficus-indica* (figure 8e), which shows a preference for the colder areas of the country (figure 11).

It is important to note that the weightings are only related to the climatic variables; and not to the plants directly, as their calculation is not involved in any way with the distribution data sets. However, since the distribution of the plants often depends on climate (de Laubenfels, 1975), it is useful to determine which climatic variables are important to each species. One

can deduce this by examining which principal component results in the best prediction of distribution for a species (as each principal component is related to a variable, or combination of variables).

A high degree of correlation between the variables usually indicates that only a few basic dimensions have been measured (Jeffers, 1967). This is confirmed by high percentages of variance being explained by the first few principal components (Manly, 1986).

From the validation results for *A. longifolia*, the best predictive maps appeared to be those produced for the small data set for PC 1 and for the extended data sets for all three PCs (Table 9). According to the chi-squared results however, all of the predictive maps were significant departures from randomness except for those produced for PC 3 and the extended data set for PC 1. Closer analysis of the chi-squared results and the coverages revealed that the maps that were significant had fairly few predictions of false presence (an incorrect prediction of presence) or false absence (an incorrect prediction of absence). The best predictive maps for this species appear to be for the small data set for PC 1 (making a combination of MAXT, COV and ELV important in determining distribution) and the extended data set for PC 2 (where MINT is important).

Chi-squared test results for *A. mearnsii* indicated that all of the predictive maps produced were significant. The validation results also showed good prediction on all three PCs for this species. However only the small data set for PC 2 did not predict large numbers of false presences and false absences, suggesting that MINT is important in determining the distribution of this species.

Only PC 2 for *O. ficus-indica* produced a predictive map that was significantly different from one that would have been produced randomly (table 11), but the map contained significant amounts of false presences and false absences, which meant that its validation result was good (table 9) but that it was not a good predictor of true presence and absence. For *S. sisymbrifolium*, none of the predictive maps produced were significant. Validation results for PC 1 suggested otherwise, but analysis of the chi-squared result indicated that the map (figure 7f) was predicting many areas of false presence (i.e. predicting presence where the plant should be absent). It may be that none of the environmental variables used here are

determining factors for the distribution of this species, and that some factor not included in this study is more important.

Principal component 2 appeared to be the best predictor for the extended data set for *A. longifolia* and for the small data set for *A. mearnsii* in terms of predictive maps that were significant departures from randomness and were good predictors of true presence. It may be that MINT is an important determinant of distribution for these two species.

An examination of the coverages (figures 7 - 9) demonstrates a similarity in appearance between the predictive maps for each species for the separate principal components. This is because each PC is related to a particular environmental variable (or a combination thereof as in the case of PC 1). For example, in figure 9 it is clear that MAR is important (areas predicted as present are generally in the wetter eastern half of the country). An accurate prediction map (i.e. one with few false presences or false absences) will only result for a species if the particular environmental factor that the PC is expressing is important to the distribution of that species. Many species, however, depend on a combination of variables to determine their distribution. If a PC expresses that particular combination, then it is likely to be a very good predictor of that species' distribution.

Sample size does not appear to be of great importance to this technique. While it appears to affect the validation results by improving the rate of prediction with an increase in sample size (table 9), the chi-squared tests reveal that an increase in sample size does not result in more accurate prediction maps.

5.5) CONCLUSIONS

The percentage of variance extracted from the principal components suggests that a sufficient number of variables were chosen to model with. The first principal component relates to a combination of MAXT, COV and ELV, the second to MINT and the third to MAR. The second principal component gave the best accurate prediction results and appears to be important in determining the distribution of the two *Acacia* species. The one significant map

produced for *O. ficus-indica* was not a good predictor of distribution for this technique as it contained many false presences and false absences. No significant maps were produced for *S. sisymbrifolium*, indicating that perhaps some other variable not used in this study may be important for the distribution of this species. The important variables determining the distribution of *A. longifolia* were MINT and a combination of MAXT, ELV and COV. For *A. mearnsii*, MINT was the important determining variable. For *O. ficus-indica* and *S. sisymbrifolium*, it appeared that the first three PCs did not express the variable, or combination of variables that were the most important in determining the distribution of these two species. This is not a limitation of this technique, but rather a pointer that some other variables may be important or that more PCs may need to be extracted.

PCA thus provided a good prediction technique as long as the PCs expressed the variables (or combinations thereof) that were most important to the distribution of a particular species.

CHAPTER 6
DISCRIMINANT FUNCTION ANALYSIS

6.1) INTRODUCTION

Discriminant function analysis (DFA) is a multivariate statistical technique that enables one to determine independent linear combinations of variables which will function to separate two or more groups. It can be applied to classify observations into groups and to predict into which group a new observation would fall (Jackson, 1983). Huberty (1992) makes a distinction between descriptive and predictive discriminant analysis. This study utilizes the latter.

Descriptive discriminant analysis is the type of discriminant analysis most commonly used by ecologists (Williams, 1983). It is frequently used to determine the best combination of variables for use in classifying observations into predetermined groups. The function derived from descriptive discriminant analysis is termed a canonical variable (Williams, 1983).

Predictive discriminant analysis, on the other hand, is used to predict into which group a new observation is likely to fall, based on a combination of variables. The functions used to predict are termed discriminant or classification functions (Williams, 1983).

Predictive discriminant analysis requires a data set containing the discriminant variables for each observation as well as knowledge of which observation belongs to which group (Huberty, 1992). Membership to a group is indicated by an identifying variable. To predict into which group an observation is likely to fall, the results from the discriminant function and the discriminant variables are used. The discriminant variables used in this study were the five environmental coverages, viz. median annual rainfall (MAR), co-efficient of variation for rainfall (COV), mean monthly maximum temperature (MAXT), mean monthly minimum temperature (MINT) and elevation (ELV). The two groups of absence and presence were given the identifying variables of 0 (plant absent) and 1 (plant present).

Although a normal distribution and equal variances for each group are assumed for discriminant function analysis, "it turns out in practise that the discriminant analysis model is surprisingly robust...the procedure is found to work well even when its assumptions are not met" (Jackson, 1983: 106). Several other authors concur (Knoke, 1982; Williams, 1983). Williams (1983) points out that while assumptions of normal distribution and equal variances are seldom met for ecological data, this does not mean that statistical techniques like discriminant analysis cannot be utilised. Instead, these methods may be used as exploratory techniques and, as such, can provide useful tools in contributing new insights into the data.

The applications of discriminant analysis are diverse. For example, it has been used to classify pollen assemblages into groups (Lui and Lam, 1985), to predict macro-invertebrate distribution in rivers (Moss *et al*, 1987), to predict the distribution of *Portulacaria afra* in the Eastern Cape (Gibson, 1995) and to map the rural-urban fringe (Fresenmaier *et al*, 1979). Rogers and Williams (1993) used linear discriminant analysis to predict the distribution of *Glossina morsitans*, the tsetse fly, in Tanzania, Zimbabwe and Kenya and to reduce the dimensionality of their data. Caughley *et al* (1987) used discriminant function analysis to group the distributions of three kangaroo species according to climate and two studies have used DFA to predict the presence or absence of zebra mussel in the Great Lakes of North America (Koutnic & Padilla, 1994; Ramcharan *et al*, 1992).

6.2) METHODS

The discriminant analysis was performed by the graphics and statistical computer package, Statgraphics version 7 (Manugistics, 1992). Presence for a species was coded as 1 and absence as 0. A separate analysis was carried out for each species. The unstandardized discriminant function co-efficients obtained from the calculations were used to create a coverage of discriminant function scores in the geographical information system. This was achieved by multiplying each layer in the GIS by its respective discriminant function co-efficient, adding the layers together and then adding the constant. This follows the form of the discriminant function i.e.:

$$D = \text{constant} + (k_1.MAR) + (k_2.COV) + (k_3.MAXT) + (k_4.MINT) + (k_5.ELV)$$

Where:

D = the discriminant function

k_i = unstandardized co-efficients

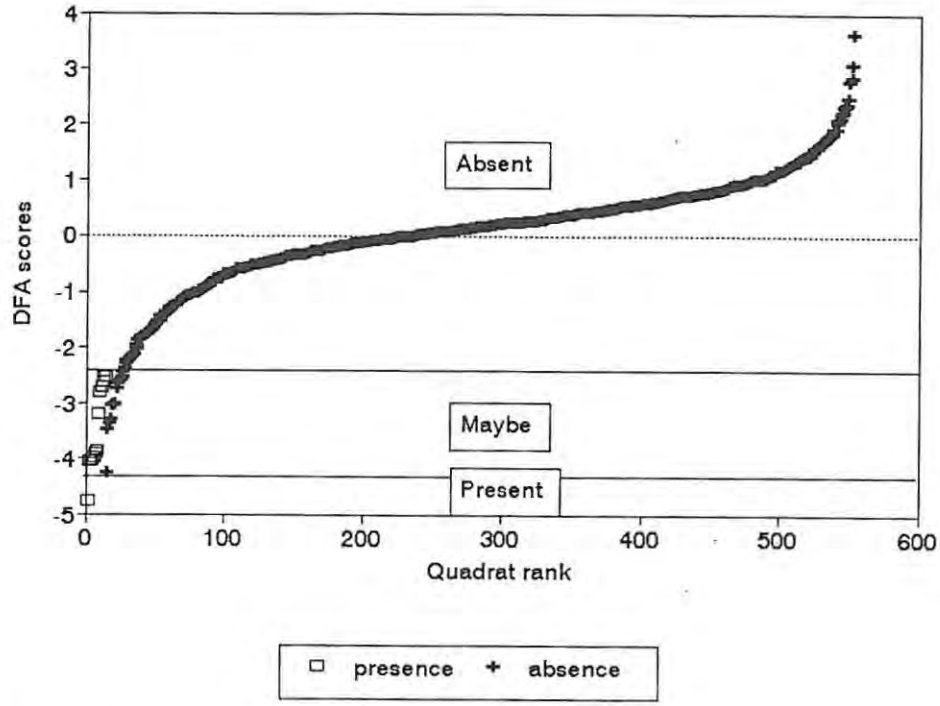
MAR, COV, MAXT, MINT and ELV are the environmental coverages in the GIS.

A separate discriminant coverage was created for each species. Each coverage was then reclassified into areas of predicted presence and absence based on the discriminant scores. The division between presence and absence for each species was determined by two methods:

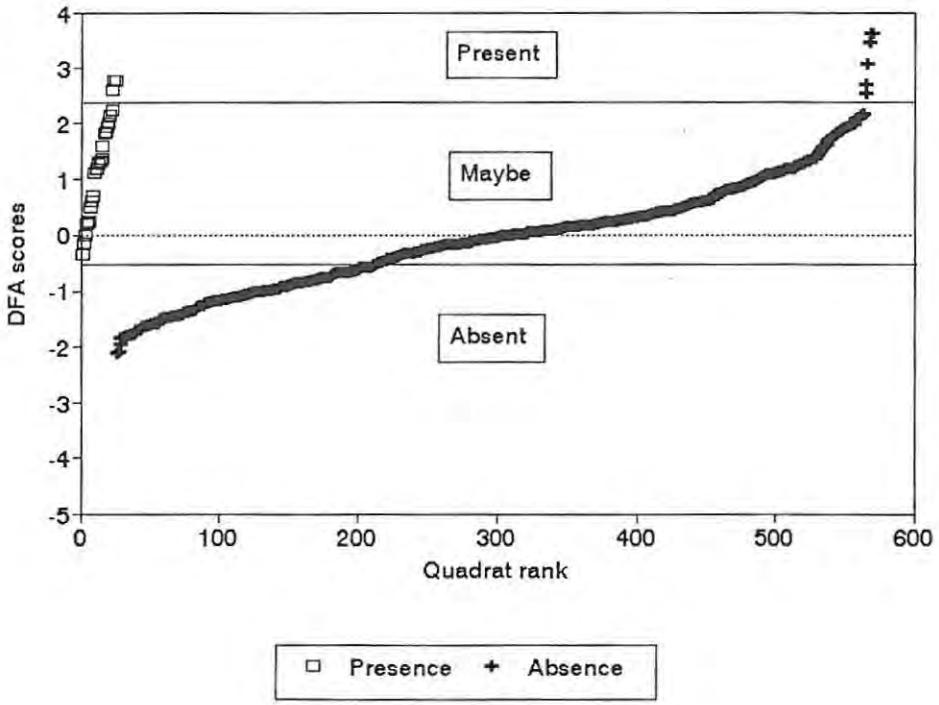
1) By finding the midpoint between the centroids of the two groups. The resulting maps were coded as 1 for areas of predicted presence and 0 for areas of predicted absence. This is a relatively rough method, but is quick and easy to perform. This is also the method used by the software itself to produce the classification tables.

2) For the second method, the discriminant function scores for each species were taken into an electronic spreadsheet package and ranked in ascending order (presence and absence separately). A graph was then plotted with presence as the first series and absence as the second series. From these graphs, thresholds between areas of definite presence, uncertain presence or absence and definite absence could be determined (figure 13a - f). These values were then used to reclassify the discriminant function coverage and the resulting maps were coded 1 for areas of predicted presence, 2 for areas of uncertainty and 3 as areas of predicted absence.

A check on the accuracy of the predictive maps was made through the use of quarter degree square records. The value at each recorded quarter degree site on the predictive map was extracted and the percentage of quarter degree squares correctly predicted as present was calculated. Posterior probabilities were not calculated as base maps with prior probabilities were not available.

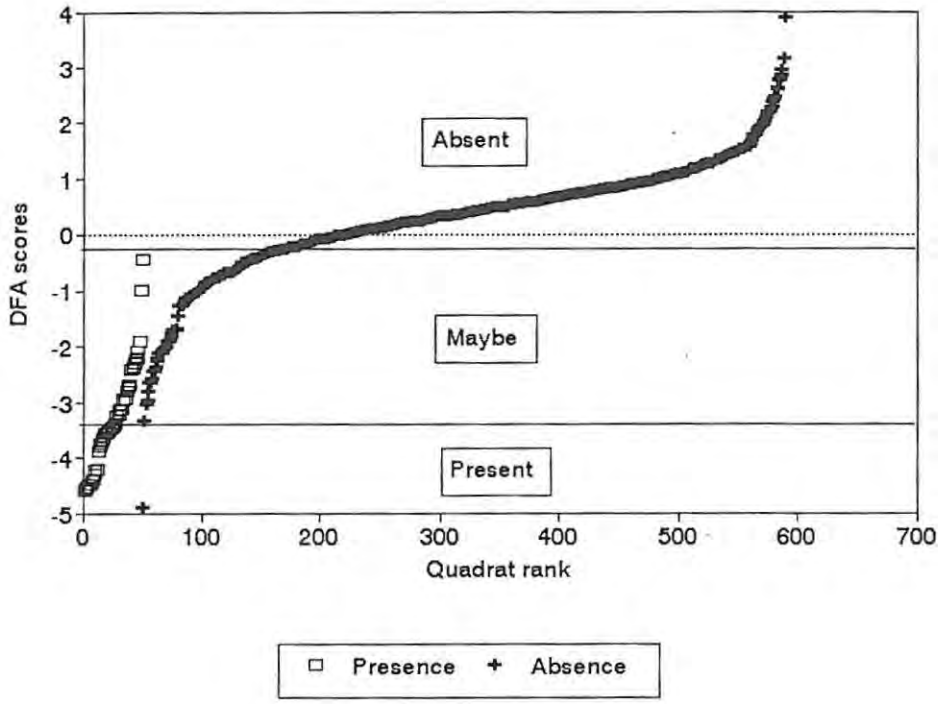


a) *A. longifolia*

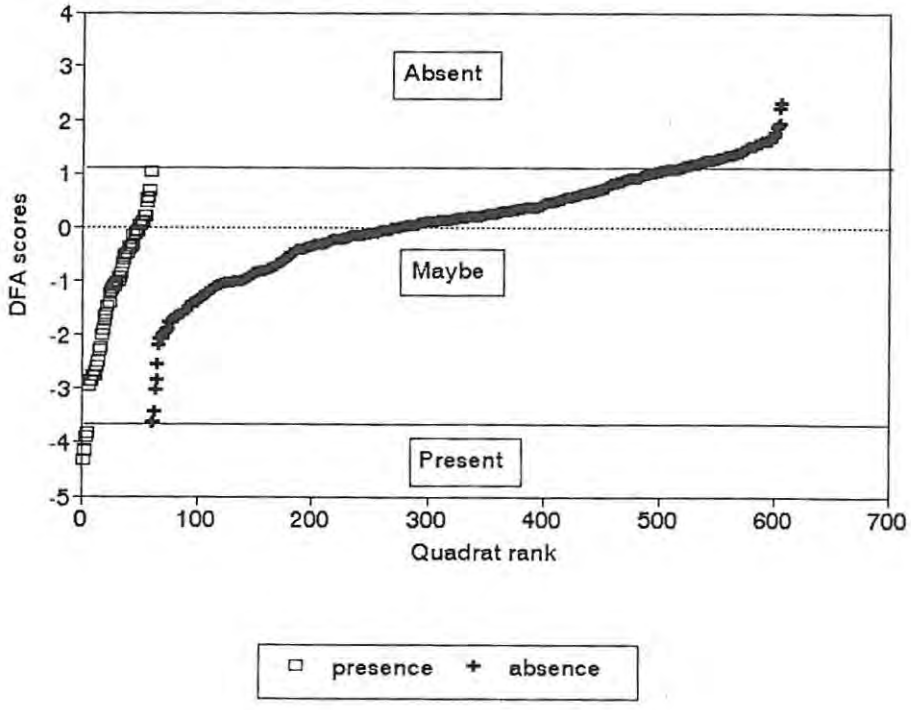


b) *A. longifolia* (extended data set)

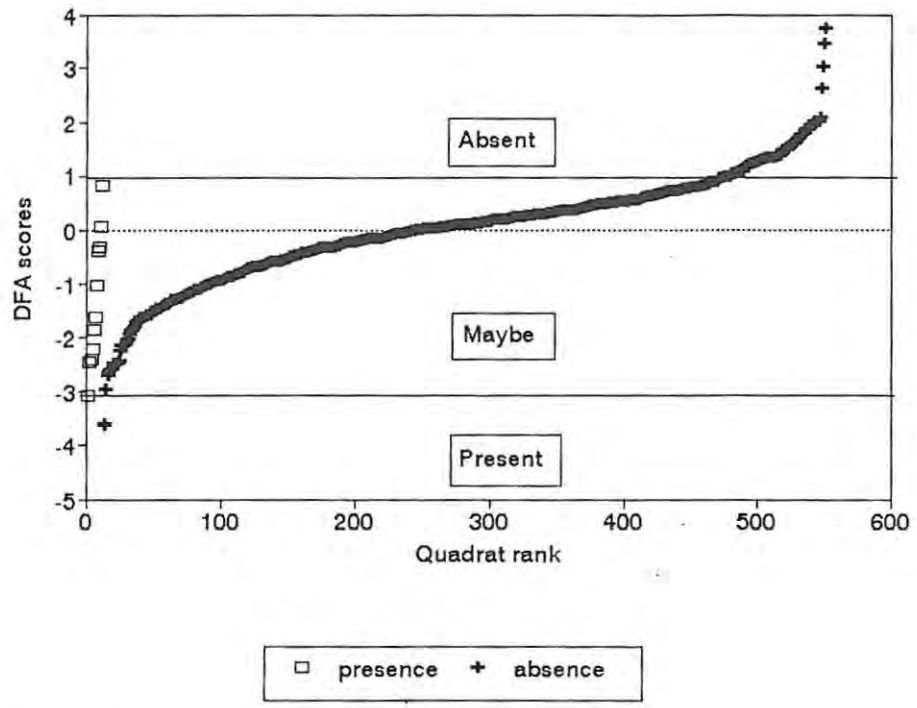
Figure 13. The graphs used to determine the thresholds between presence, possible presence and absence for DFA Method 2.



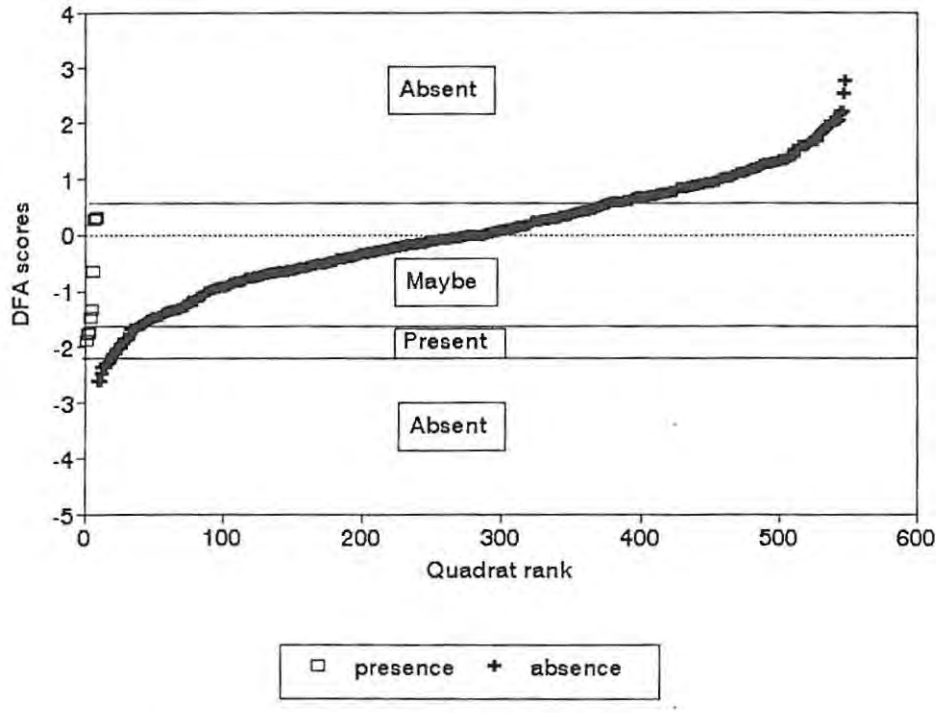
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

6.3) RESULTS

One discriminant function was produced for each of the four species. The small data set for *A. longifolia* showed the best classification results (table 13a) with 100% of the plants being correctly predicted as present and 94% correctly classified as absent, while the extended data set for *A. mearnsii* showed the least accurate classification of the four species (table 14b) with 60% being correctly predicted as present and approximately 79% as absent. The classification results for *O. ficus-indica* and *S. sisymbrifolium* (tables 15 and 16) were intermediate.

Table 13. Method 1 classification matrix for *A. longifolia* (numbers in brackets are percentages). 1 = present, 0 = absent. a) refers to the small data set and b) to the extended data set.

a)

	Predicted	group	Total
Actual group	0	1	
0	509 (94)	31 (6)	540 (100)
1	0 (0)	14 (100)	14 (100)

b)

	Predicted	group	Total
Actual group	0	1	
0	511 (95)	29 (5)	540 (100)
1	2 (4)	48 (96)	50 (100)

Table 14. Method 1 classification matrix for *A. mearnsii* (numbers in brackets are percentages). 1 = present, 0 = absent. a) refers to the small data set and b) to the extended data set.

a)

	Predicted	group	Total
Actual group	0	1	
0	422 (77)	123 (23)	545 (100)
1	7 (28)	18 (72)	25 (100)

b)

	Predicted	group	Total
Actual group	0	1	
0	428 (79)	117 (21)	545 (100)
1	24 (40)	36 (60)	60 (100)

Table 15. Method 1 classification matrix for *O. ficus-indica* (numbers in brackets are percentages). 1 = present, 0 = absent.

	Predicted	group	Total
Actual group	0	1	
0	430 (80)	109 (20)	539 (100)
1	4 (33)	8 (67)	12 (100)

Table 16. Method 1 classification matrix for *S. sisymbrifolium* (numbers in brackets are percentages). 1 = present, 0 = absent.

	Predicted	group	Total
Actual group	0	1	
0	391 (73)	148 (27)	539 (100)
1	3 (33)	6 (67)	9 (100)

The standardized co-efficients (table 17) can be used to determine which variable has the most weight in differentiating between the groups (Jackson, 1983). For *A. mearnsii*, the co-efficient most important for discriminating between the presence and absence groups was minimum temperature, for *S. sisymbrifolium*, median annual rainfall and for *A. longifolia* and *O. ficus-indica*, elevation.

Table 17. Standardized co-efficients from the discriminant function analyses. a) refers to the small data sets and b) refers to the extended data sets.

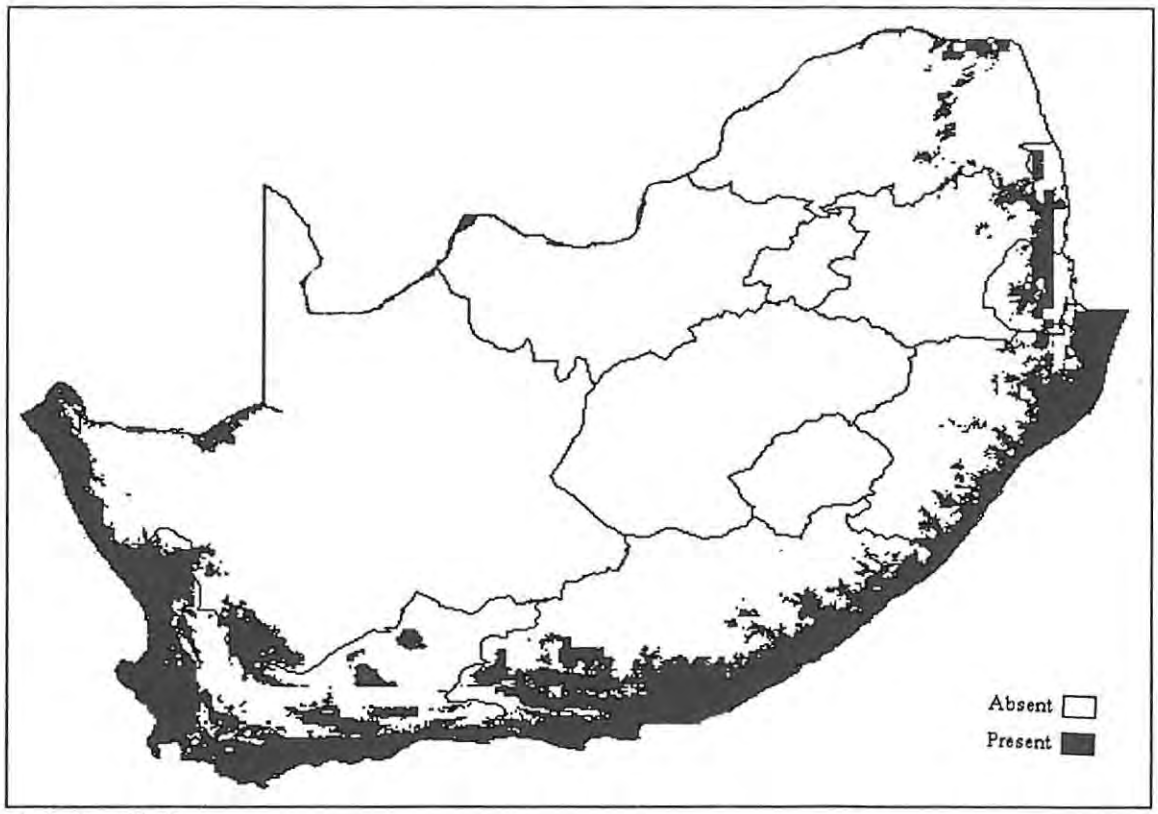
Species		MAR	COV	MAXT	MINT	ELV
A. longifolia	a)	-0.22	-0.23	0.40	-0.03	0.92
	b)	-0.01	0.14	-0.07	0.27	1.16
A. mearnsii	a)	0.02	-0.08	-0.61	0.57	-0.41
	b)	-0.22	-0.02	0.50	-0.59	0.40
O. ficus-indica		0.48	0.45	-0.45	0.20	0.72
S. sisymbrifolium		0.63	0.00	0.58	-0.21	0.31

The predictive maps produced by Method 1 show the predicted likelihood of the plant species being either present or absent (figure 14a - f). The coverages produced by Method 2 divided the potential distribution of the plant species into three categories, present, absent and maybe present (figure 15a - f). The latter category depicts areas of uncertainty where the plant may or may not be present. This category is a result of the DFA classification which did not discriminate clearly between presence and absence.

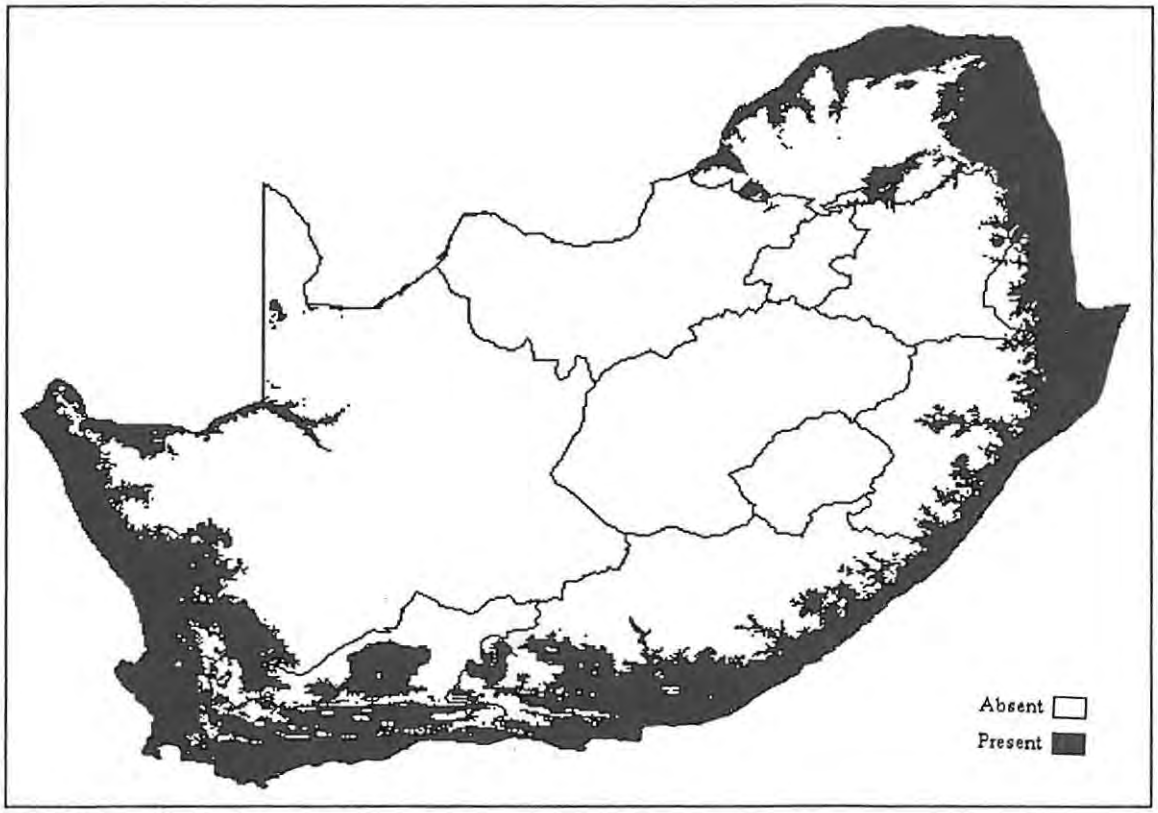
The results of the validation of the predictive maps are expressed as percentages of sites correctly predicted as present for each species for Method 1 and as the sites correctly predicted as present and as maybe present by Method 2 (table 18). This was undertaken to allow for comparison between the two methods, as in Method 1 the maybe present category (where the DFA made some misclassifications) was not separated out as in Method 2, but formed part of the presence category.

Table 18. Percentage of sites correctly predicted as present and maybe present. a) refers to the small data sets and b) to the extended ones.

Species		Method 1	Method 2
A. longifolia	a)	74	57
	b)	95	74
A. mearnsii	a)	64	84
	b)	59	74
O. ficus-indica		76	68
S. sisymbrifolium		28	37

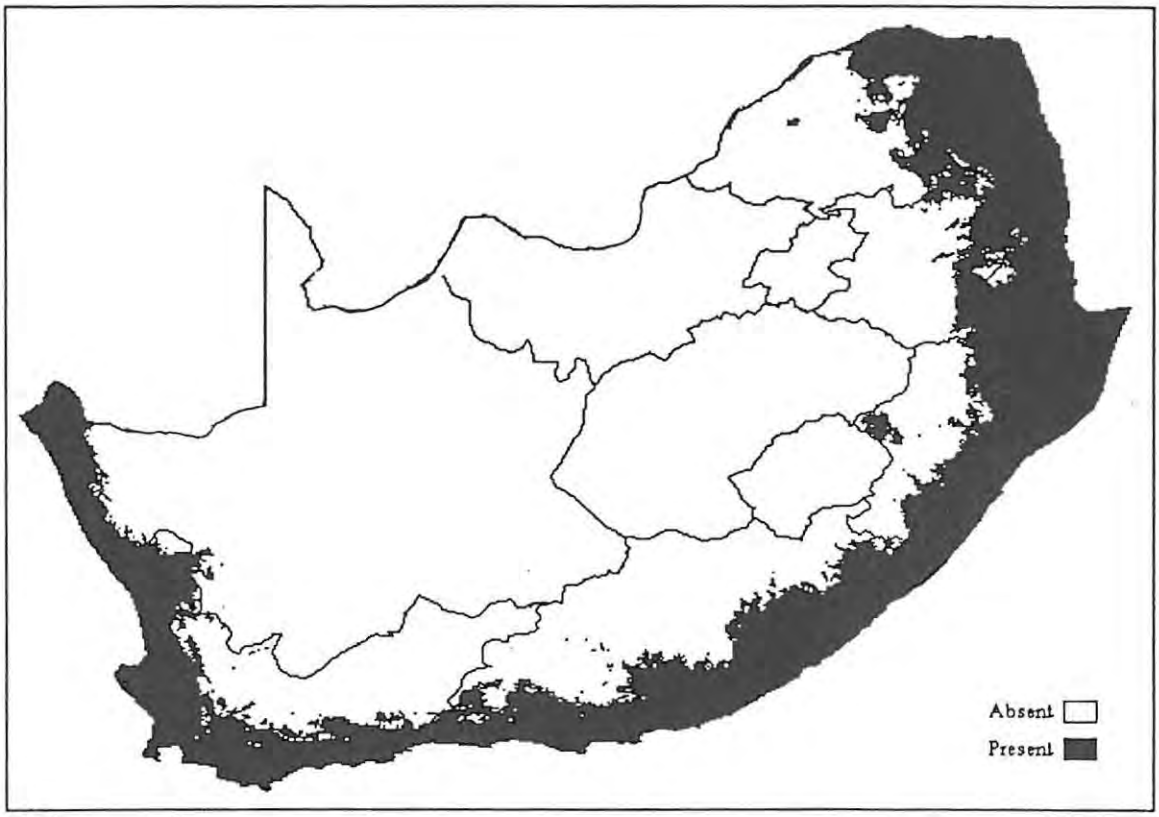


a) *A. longifolia*

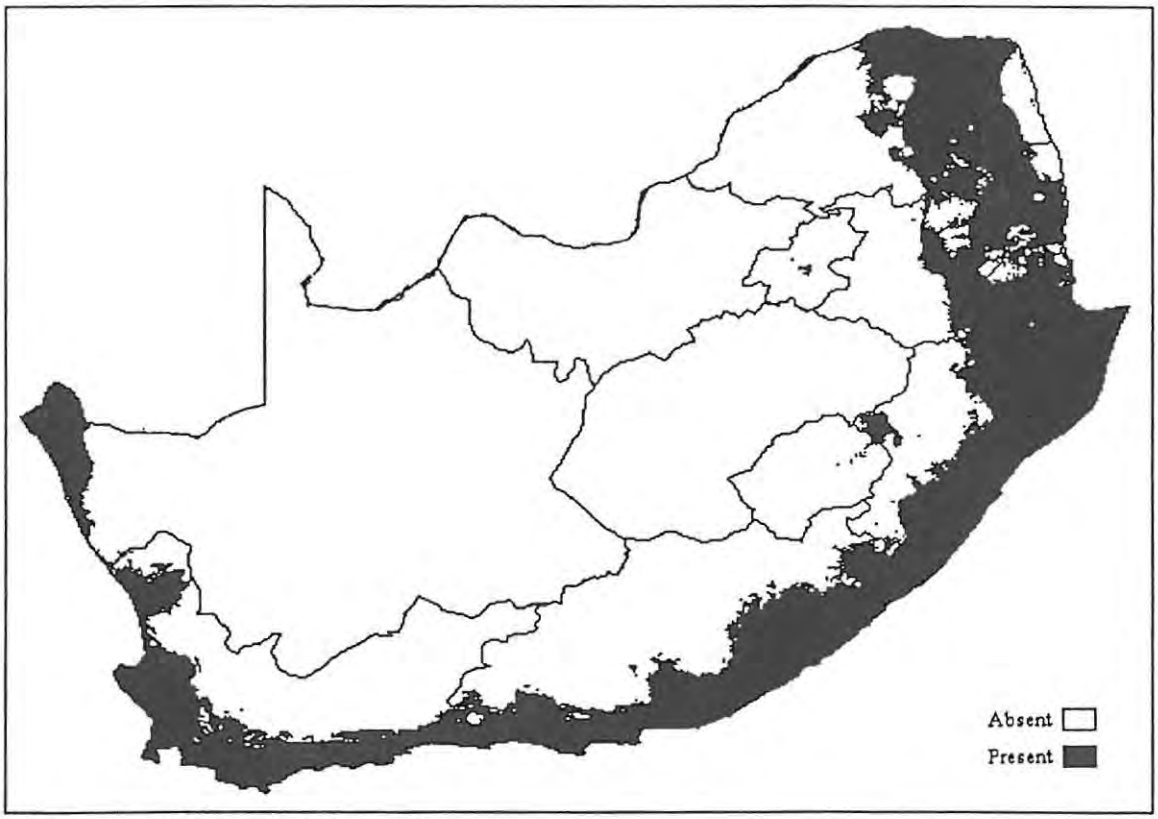


b) *A. longifolia* (extended data set)

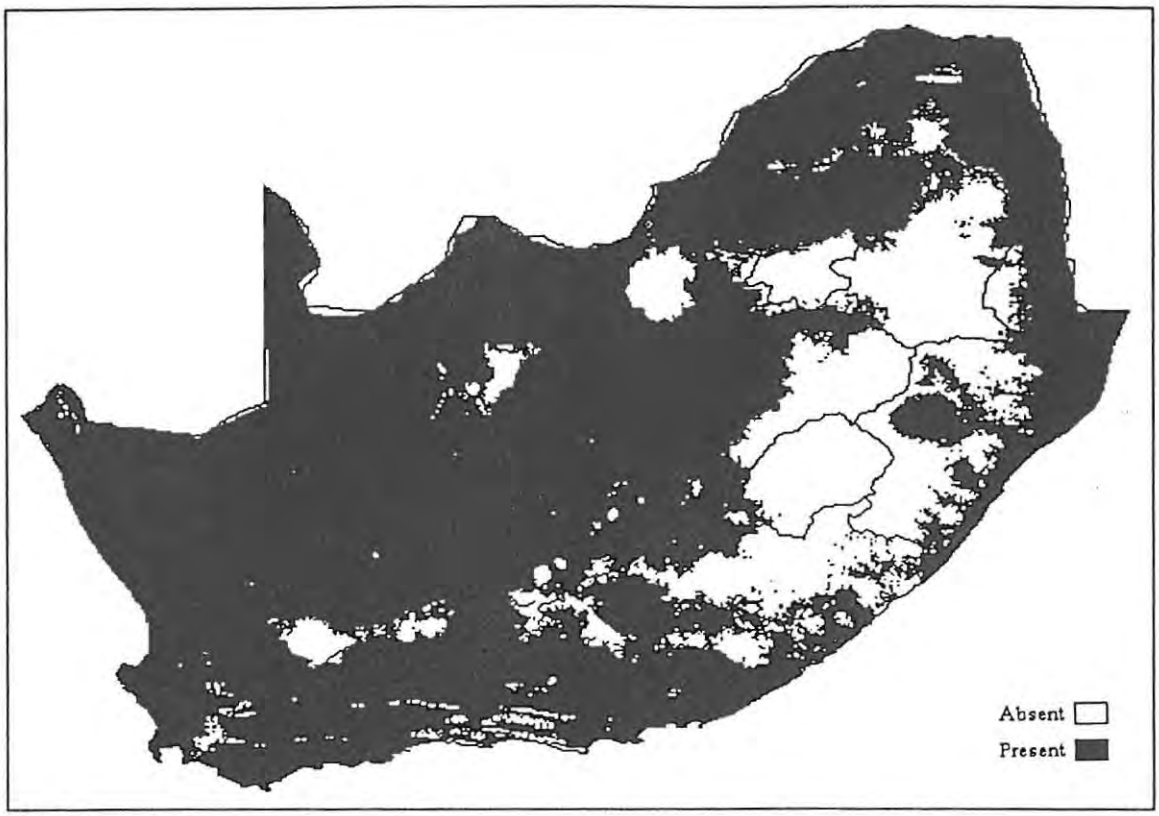
Figure 14. Potential distribution maps derived from the discriminant function analysis, Method 1.



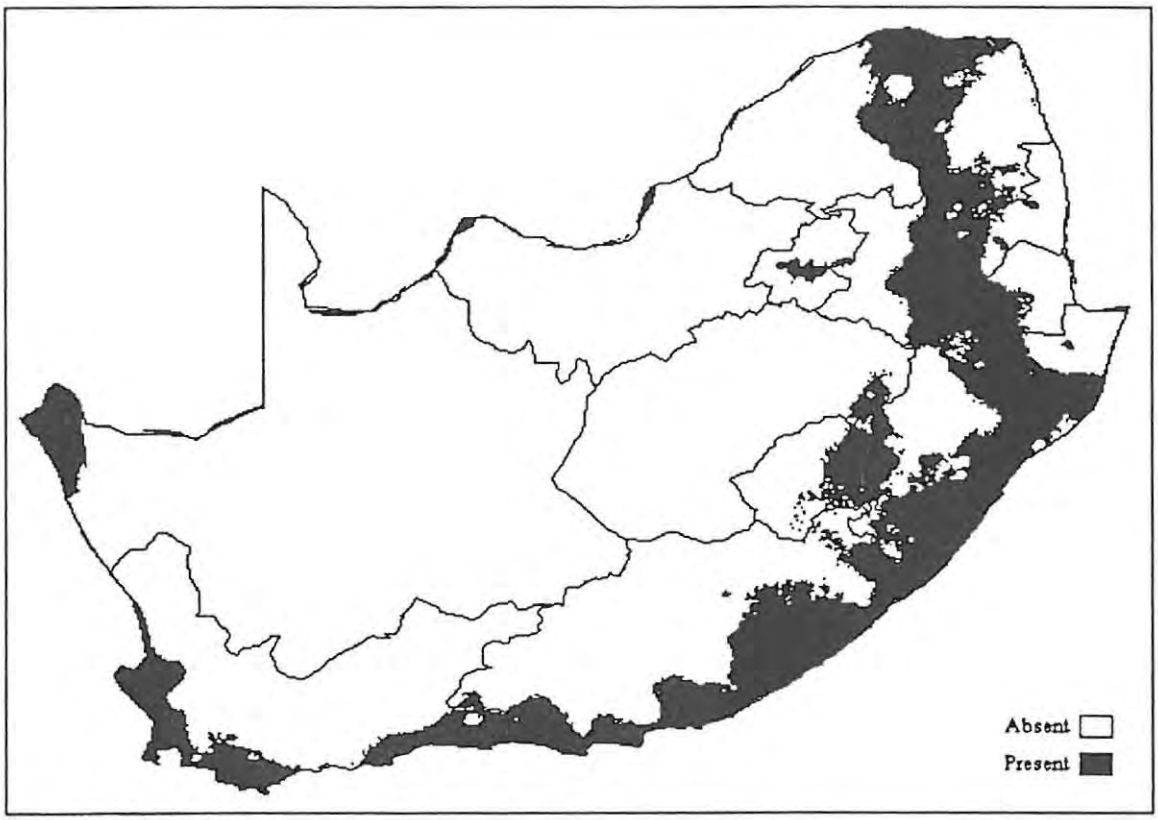
c) *A. mearnsii*



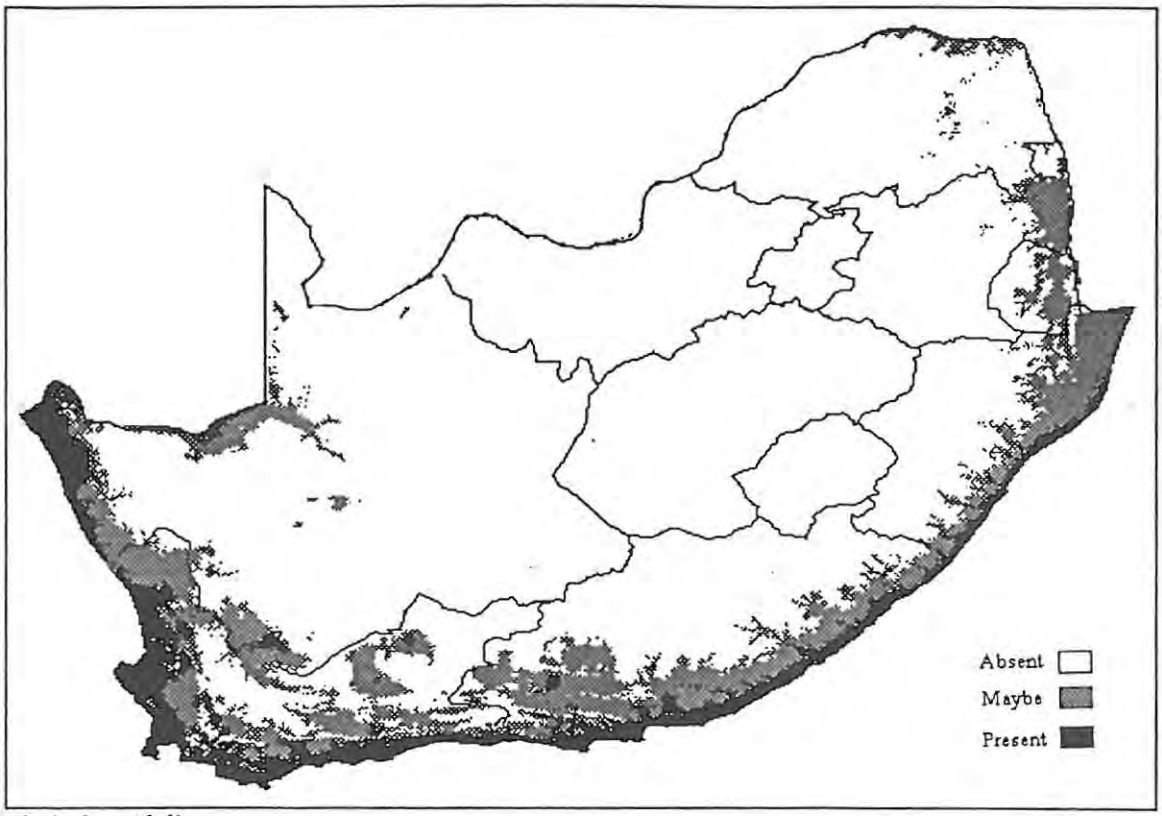
d) *A. mearnsii* (extended data set)



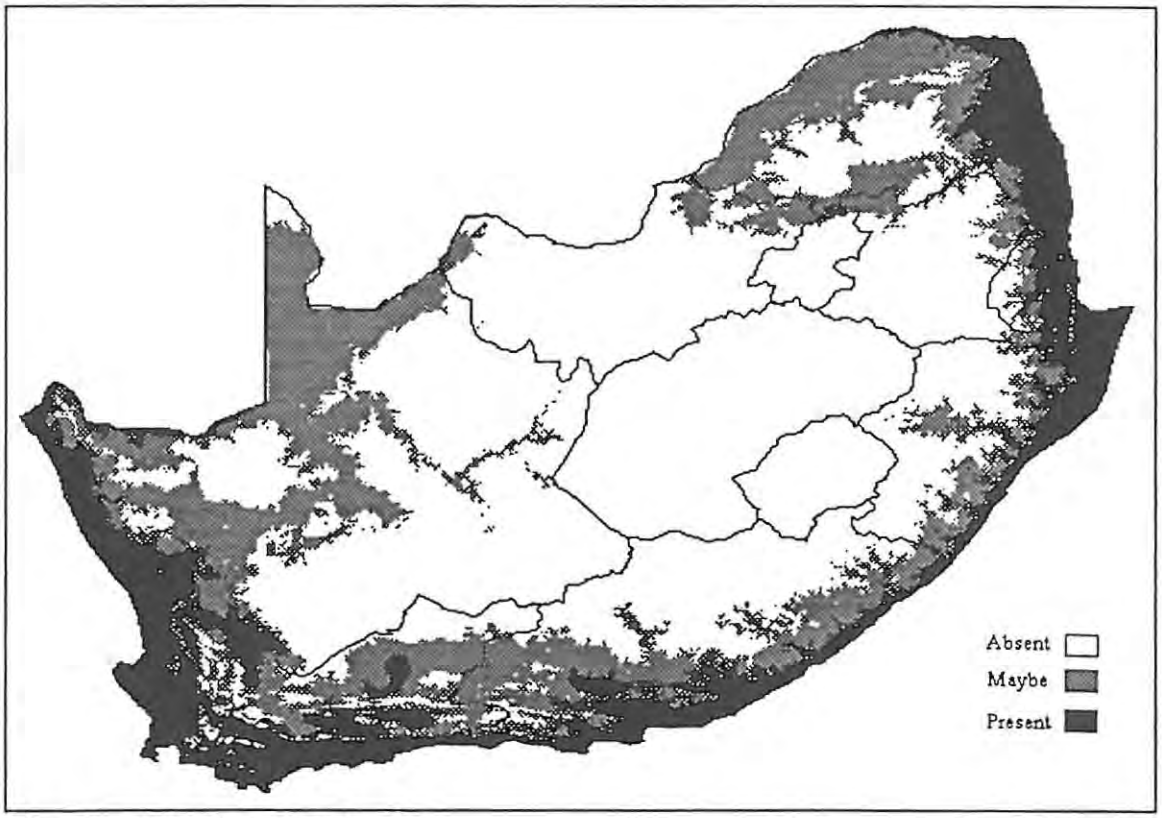
e) *O. ficus-indica*



f) *S. sisymbrifolium*

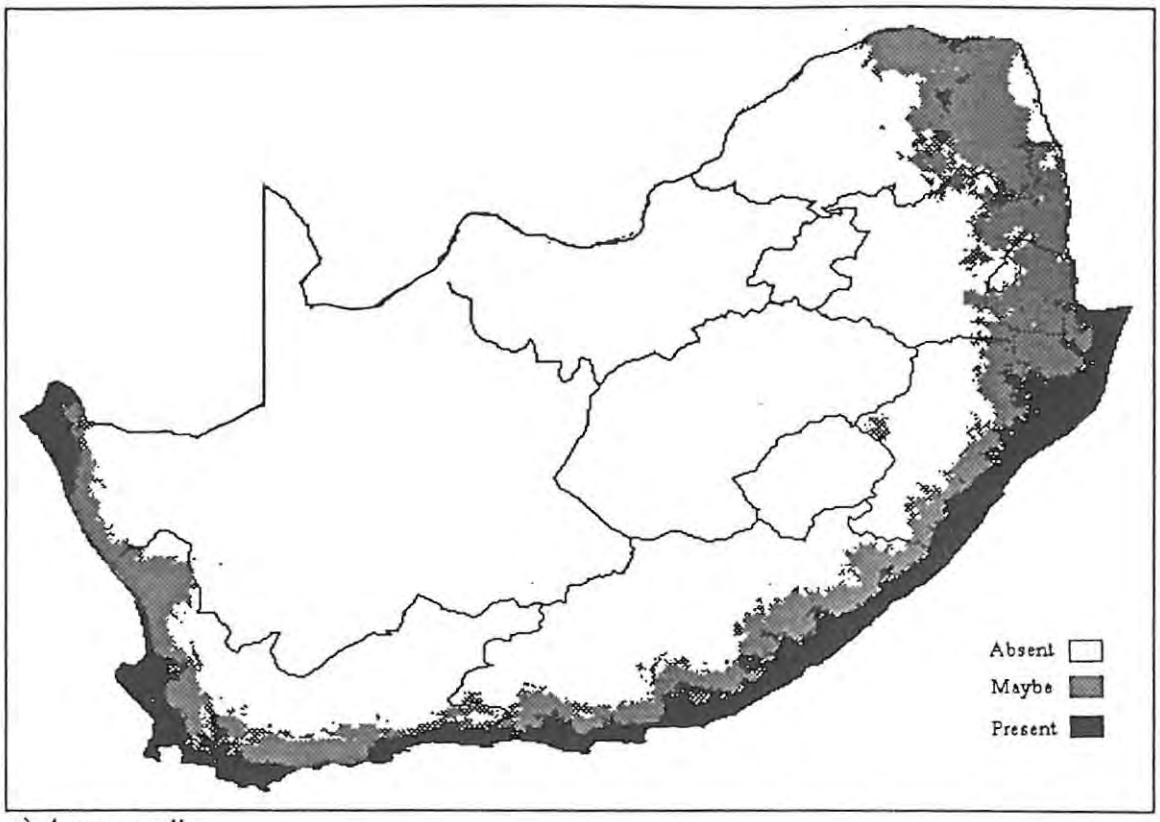


a) *A. longifolia*

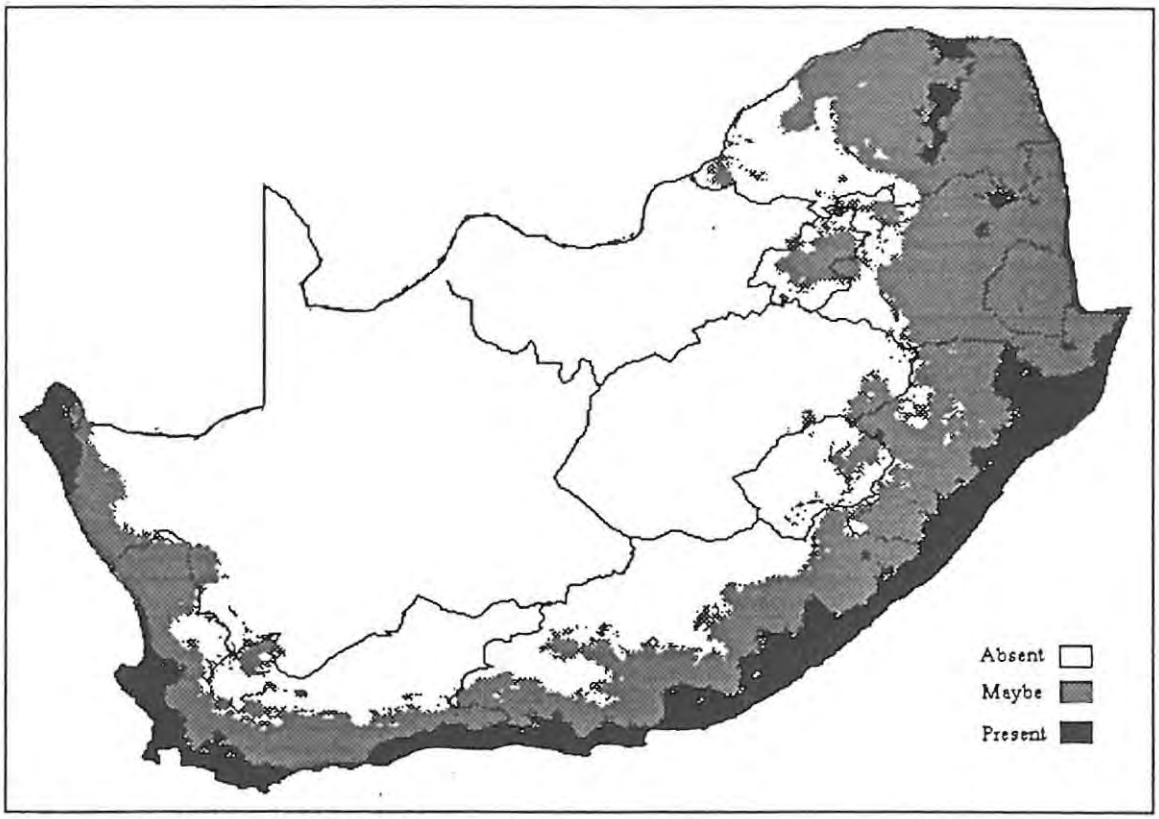


b) *A. longifolia* (extended data set)

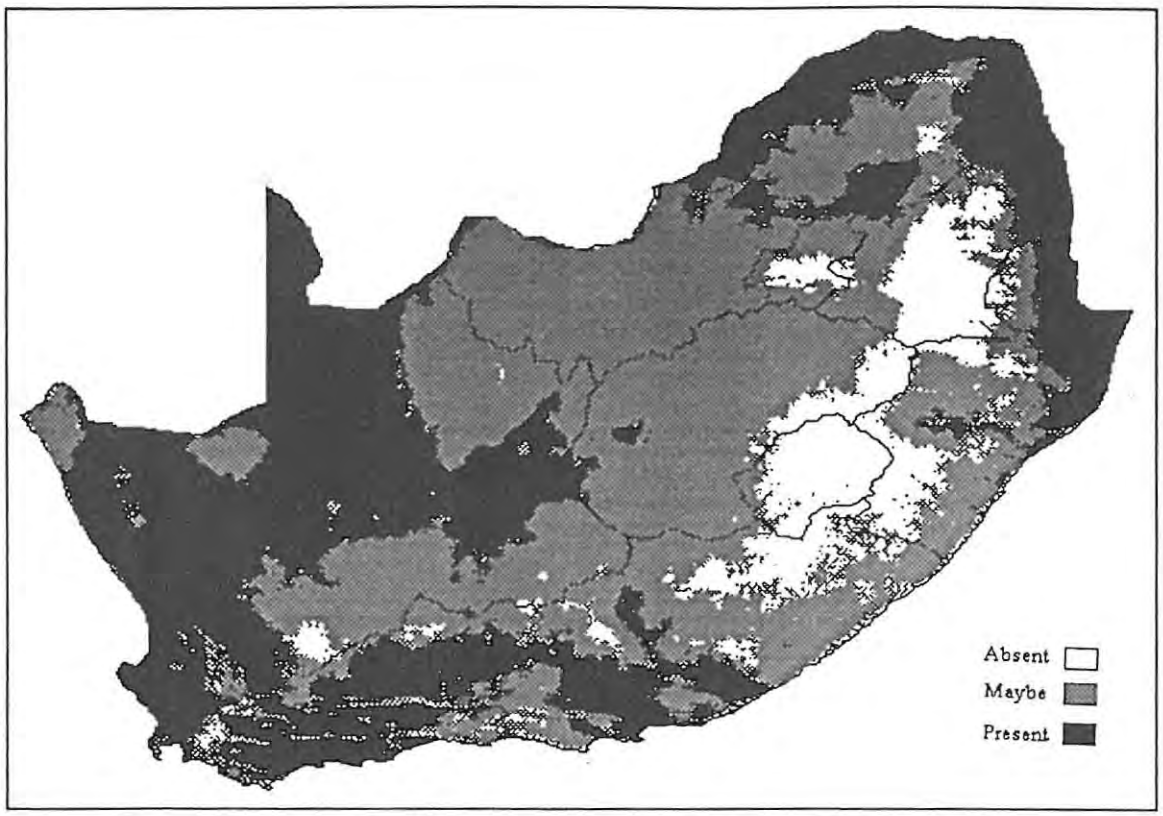
Figure 15. Potential distribution maps derived from the discriminant function analysis, Method 2.



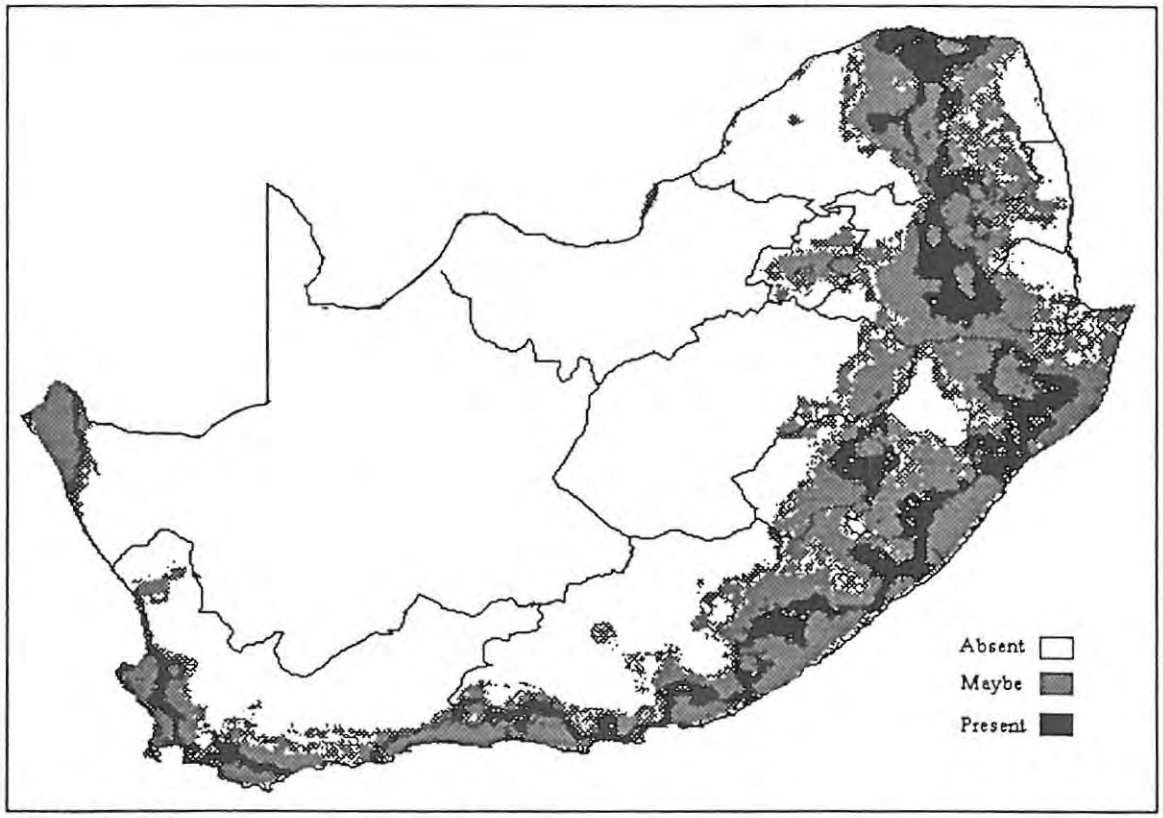
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

Table 19. Chi-squared results and significance levels for Method 1. a) refers to the small data sets and b) to the extended ones.

Species		Chi-squared	Sig. level
A. longifolia	a)	8.53	0.0035
	b)	30.42	0.0000
A. mearnsii	a)	19.43	0.0000
	b)	7.41	0.0065
O. ficus-indica		107.71	0.0000
S. sisymbrifolium		4.84	0.0278

Table 20. Chi-squared results and significance levels for Method 2. a) refers to the small data sets and b) to the extended ones.

Species		Chi-squared	Sig. level
A. longifolia	a)	3.00	0.2231
	b)	6.91	0.0316
A. mearnsii	a)	31.91	0.0000
	b)	95.69	0.0000
O. ficus-indica		251.37	0.0000
S. sisymbrifolium		10.67	0.004

6.4) DISCUSSION

The discriminant analysis procedure produces n-1 axes or functions, where n is the number of groups (Manugistics, 1992); therefore only one discriminant function was produced for each of the four species. This meant that a graph that visually shows the discrimination between the groups of presence and absence for each species could not be constructed as there was only a single axis. However, graphical output was obtained through use of the GIS.

One of the problems with discriminant function analysis is the misclassification of cases into the incorrect groups. Rogers and Williams (1993) term these misclassifications 'false negatives' and 'false positives'. A 'false negative' occurs when absence is incorrectly

predicted (the plant is present but is predicted as absent) and a 'false positive' when presence is incorrectly predicted (a plant is absent but is predicted as present). This misclassification is especially problematic when the differences between the groups are not clearly defined, as the chances of classifying an observation into the wrong group are then greater. The resulting predictive maps can only be as accurate as the original DFA classification. The classification matrices give the percentage of a false negatives and positives likely to occur on the predictive maps for Method 1 (as this is the method used by the software to produce the classification matrices). For Method 2, the false negatives and false positives fall into the 'maybe' category on the predictive maps and therefore give an indication of the areas where most of the misclassification occurs. These areas may be regions where the plants are present in low densities or are constantly invading, but not surviving for long enough to become well established (Rogers & Williams, 1993)

According to the standardised discriminant function co-efficients for *A. longifolia* (table 17), the most important variable in distinguishing between presence and absence is elevation. This species is native to coastal areas in Australia (Stirton, 1987) and does not appear to enjoy high elevations. Examination of the coverages produced for this species (figures 14a, b and 15a, b) also indicate that elevation is important. This can be seen in the artifacts left by the ELV coverage in the form of areas of predicted presence extending inland along rivers and drainage basins and along the Orange River. The Cape Fold mountains are also clearly picked out by the ELV surface.

Maximum and minimum temperatures are the variables given the greatest weight in determining the presence or absence of *A. mearnsii*. While this species will tolerate some frost, it prefers temperate climates (De Beer, 1987). The temperature coverages do not appear to have left many artifacts on the predictive maps. Artifacts on the MAXT and MINT coverages are usually visible as circles produced by the interpolation process.

Elevation appears to be one of the determining factors between presence and absence for *O. ficus-indica*. However, elevation is strongly correlated with COV and MAXT (table 8), and the weights for the first principal component (table 7) suggests that there may be some redundancy between the three, i.e. that they measure similar things (Manly, 1986). It is therefore unlikely that the distribution of *O. ficus-indica* is solely determined by elevation, but

rather by a combination of ELV, COV and MAXT. This would also explain the absence of any obvious artifacts from one of the input coverages (figures 14e and 15e).

The most important distinguishing variables for *S. sisymbrifolium* are MAR and MAXT (table 17). COV is not a prominent factor in determining the distribution of *S. sisymbrifolium* (table 17); it may be that this species is able to survive in areas of uncertain rainfall due to its extensive underground root system. If this is the case, then large parts of the country will be susceptible to invasion. The influence of the MAR coverage is particularly noticeable in the predictive map produced for Method 2 (figure 15f) in the form of the bands and lines of predicted presence.

With regards to the predicted distribution of the species using Method 1, *A. longifolia, A. mearnsii* and *S. sisymbrifolium* all show a preference for the eastern parts of the country, especially the coastal areas. These areas tend to receive more rainfall (and more reliable rainfall) than the western parts of the country. According to this method, the areas suitable for invasion for *O. ficus-indica* are extensive, covering most of the country with the exception of the higher mountain ranges like the Drakensberg and some areas in the Karoo and the Northern Province. *Opuntia ficus-indica* is able to tolerate drier conditions than the other three species due to its succulent nature (Moran & Zimmermann, 1984).

For Method 2, *A. longifolia* exhibits a similar distribution to the results achieved with Method 1, but shows areas of uncertainty in the northern Cape and Kalahari, as well as up along the South African - Zimbabwean border (figure 15a, b). These areas are likely to be areas that were misclassified as present (i.e. false positives) by the DFA. Both of these areas are dry and are therefore unlikely to be favourable to the species which prefers temperate coastal areas (Stirton, 1987). However, it may be that the plant is constantly invading these areas but that the climate is too harsh for it to establish itself. It should also be borne in mind that the distribution of this species may be as a consequence of its cultivation by man. The areas potentially suitable for *O. ficus-indica* (figure 15e) again cover large areas of the country, however many of these areas are now regions of uncertainty. This species shows a preference for the western half of the country and the northern tip. If this prediction is correct, it would make *O. ficus-indica* the only species in this study to show a potential for invading the very arid regions of the Kalahari. Most of the species show a preference for the regions of higher

rainfall i.e. the coastal and eastern parts of South Africa. The exception is *O. ficus-indica* which seems to be suited to the more arid areas. Depending on perspective, *O. ficus-indica* could be a pestilent invader in these sensitive arid areas, or a valuable crop plant that will survive the harsh conditions.

The fynbos region is an area that is already heavily invaded (MacDonald, 1984); according to the predictive maps, it is climatically suited to most of the species and should be managed as a 'high risk' zone. However, it should be borne in mind that while the area may appear to be suited to the plants in terms of rainfall, this study did not take into account the season in which the plants customarily received rainfall in their native regions. The winter rainfall that the fynbos area receives may not be ideal for these four invasive species and they may struggle to establish themselves.

Coastal areas also appear particularly at risk from invasion, and this is especially important in terms of water, as many of our rivers drain from the escarpment down to the coast. *Acacia mearnsii* is particularly a problem along watercourses (DWAF, 1996a). Unfortunately, South Africa's coastal regions are also popular tourist destinations, and development in this zone could further open the way for invasion. These coastal zones should therefore be regarded as high risk areas.

With regards to the validation of the predictive maps, the chi-squared test results indicate that all of the maps produced by Method 1 were statistically significant (table 19); judging by the validation results, the map for *S. sisymbrifolium* is a significantly poor predictor of this species' distribution. The other maps all appear to be good predictors of presence (table 18).

The chi-squared results for Method 2 (table 20) indicate that all the predictive coverages are statistically significant predictors of distribution except for the one produced for the small data set for *A. longifolia*. As for Method 1, the coverage for *S. sisymbrifolium* appears to be a significantly poor predictor of distribution. Examination of the coverage for the extended data set for *A. longifolia* indicates that the DFA is predicting a number of areas of false positives, particularly in the Northern Cape (figure 15b).

A comparison of the two methods reveals that Method 2 shows better prediction with regards to the validation results than Method 1 for areas of presence for *A. mearnsii* and *S. sisymbrifolium*, but slightly worse prediction for the other two species. It is interesting to note that *A. mearnsii* and *S. sisymbrifolium* show the worst classifications with regards to the DFA while the classifications for *A. longifolia* and *O. ficus-indica* are more accurate, perhaps indicating a better defined split between presence and absence. It may be that division into groups based on the midpoint of the centroid (Method 1) produces better results if the split between the groups is well defined. Method 2 on the other hand demonstrates superior results where there are areas of uncertainty between true presence and true absence as it allows false negatives and positives to be taken into account.

Method 1 is quicker to calculate than Method 2, but works best if the differences between the groups are clear cut. When the thresholds between the groups are not clearly defined as for these plants (i.e. many areas of false positives and negatives), this method may not be very accurate. Method 2 has the advantage of discerning areas of uncertainty of presence or absence. However, the thresholds between areas of definite absence or presence and areas of uncertainty cannot be exactly defined and estimates must be used.

The validation results indicate that in terms of the number of plants correctly predicted as present, neither of the methods was consistently better than the other. In terms of the chi-squared results, Method 1 produced one more statistically significant map than did Method 2.

With regards to sample sizes, an increase in sample size for *A. longifolia* improved the success rate of the predictive maps for both methods, but increasing the sample size for *A. mearnsii* resulted in poorer predictions. This may be due to the fact that the discrimination between presence and absence for *A. longifolia* was better defined than for *A. mearnsii*, leading to more accurate classifications for the former.

In general, prediction of distribution using discriminant function analysis appears to be quite promising, except for *S. sisymbrifolium*. There may be several reasons for this. The factors affecting plant distribution are many, and only a few were considered in this study; it is possible that some parameter not considered here, for example, soil type or human impact,

may markedly affect the distribution of this species. The sample size for *S. sisymbrifolium* was also very small (n = 9) and therefore possibly not representative of the species' distribution.

6.5) CONCLUSIONS

For *A. longifolia* the most important variable used by the DFA to distinguish between presence and absence is ELV; for *A. mearnsii*, MAXT and MINT; for *O. ficus-indica*, a combination of ELV, COV and MAXT, and for *S. sisymbrifolium*, MAR and MAXT. The influence of these input converges is indicated on the predictive maps in the form of artifacts and is most marked on the maps for the extended data set for *A. longifolia*.

With the exception of *S. sisymbrifolium*, predictive discriminant function analysis appears to be a good predictor of the plant's distribution, with the validation results for the other three species having a 57% to 95% success rate. In terms of the percentage of presence sites correctly predicted, neither method appears consistently superior to the other and each has its own advantages and disadvantages. With regard to the chi-squared test results, Method 1 produced one more statistically significant map than Method 2.

Areas most suited to invasion appear to be the fynbos and coastal regions, as well as parts of KwaZulu-Natal, Mpumalanga and the Northern Province. These areas should be considered as 'high-risk' regions and control and preventative measures should be implemented.

CHAPTER 7
ARTIFICIAL NEURAL NETWORKS

7.1) INTRODUCTION

Fausett (1994: 430) defines artificial neural networks (ANNs) as "information processing systems, inspired by biological neural systems but not limited to modelling such systems. Neural networks consist of many simple processing elements joined by weighted connection paths. A neural net produces an output signal in response to an input pattern; the output is determined by the values of the weights."

Neural networks can perform a mapping function between three-dimensional inputs and one-dimensional outputs. The neural network must learn the association between the input vectors and the output; this is accomplished by supervised or unsupervised training of the network using examples. Once the neural net has learned the association between the two, it is tested. Testing usually involves giving the neural net data from the same data set that was used to train it, but which was reserved from the original training data; or by using an independent testing data set, as was the case in this study.

There are a number of types of ANNs, but all have a common structure (figure 16):

i) an input layer
ii) one or more hidden layers
iii) an output layer

The way in which these layers and the connections between them are arranged is termed the net architecture (Fausett, 1994).

Some of the advantages that ANNs can offer for modelling and prediction purposes is their nonlinearity, adaptivity and fault tolerance (Haykin, 1994). Adaptivity refers to the ability of the network to modify the weights to adapt to a change in the environment (Wasserman, 1989) while fault tolerance allows the network to degrade gracefully i.e. if the network encounters faults in the data, it does not break down suddenly but degenerates gradually. Artificial neural networks can also generalize which allows them to ignore minor variations in the input data and therefore cope with imperfection (Wasserman, 1989).

The genesis of neural networks is usually traced back to the 1940's with the construction of the McCulloch-Pitts neuron and Hebb's learning rule (Fausett, 1994). The following two decades saw the development of perceptrons (single layer networks), adaline networks and the advent of the delta rule (which is a precursor to the backpropagation network). However, the excitement surrounding the ability of perceptrons soon turned to disillusionment as their inability to solve simple problems was discovered (Wasserman, 1989). After the disappointment with the capabilities of the earlier neural networks to cope with simple problems, the 1970's showed a decrease in artificial neural network research and it was only with the advent of the back-propagational network in the 1980's, which allowed multi-layer networks to be produced that there was a resurgence of interest in the neural network field (Fausett, 1994).

Neural networks and their potential applications have attracted a great deal of attention and not just in the scientific field. Entrepreneurs from other sectors of the economy have realised their potential worth and the uses to which they have been put are many and varied. For example, in the medical field they have been used to make diagnoses based on the symptoms of the patient (Fausett, 1994) and to analyze electroencephalograms to detect various neurological states such as tiredness (Wasserman, 1989). Financial institutions have found them potentially useful in predicting which loan applicants are likely to default on loan repayments (Kinoshita, 1988) while telecommunications companies have used them to dampen noises on telephone lines and to recognize handwritten postal codes (Fausett, 1994). Sejnowski and Rosenberg (1988) taught their neural network to read English text and to convert it to speech.

Prediction using neural networks can be very successful depending on how well they have trained and the quality of the training vectors. Anon (1996) used a backpropagation net to predict tortoise density based on the number of tortoise droppings found along particular transects.

A back-propagational network uses the generalized delta rule to adjust the weights to give the smallest mean squared error (Fausett, 1994). It trains through supervised classification, which means that it learns from a set of data for which one already know the correct outputs. Briefly, training involves feeding forward the training vectors, calculating the error between

the desired output and the net output, propagating this error back and then adjusting the weights to minimize the error (Fausett, 1994).

In more detail, training proceeds as follows: each of the units in the input layer receive an input signal (i.e. a case from the training data); this signal weighted and fed forward to each of the units in the hidden layer. These hidden layer units sum all the signals they receive then apply an activation function to them (to squash the values to within a certain range) and feed them forward to the units in the output layer. The output layer units also weight and sum the signals they receive and apply an activation function. The output unit then compares this value to the target output pattern and determines the error for that pattern. Based on that error, the output unit calculates weight and bias corrections and sends this information back to the hidden layer. The hidden layer again sums the inputs and multiplies by the derivative of its activation function to calculate error information and weight/bias corrections. The weights and biases are then updated simultaneously and the process repeats itself until a stopping condition is reached. A stopping condition may be a number of runs through the training data set (termed epochs), the reaching of a global minimum of error, the reaching of a desired RMS error or when the network has learned all the input-output examples. An activation function is used to ensure that the net outputs are within a certain range. A common activation function for backpropagational networks is the binary or logistic sigmoid (Demuth & Beale, 1994) where the target outputs are binary (between 0 and 1).

Some of the advantages of back-propagational networks are that they are fairly simple to operate, they allow a multiple layer network to be constructed and they work well. Their limitations are that they have to be trained using supervised training (i.e. one must have some input patterns for which one already knows the output) and that to train well, the net needs a large number of input-output examples (Hecht-Nielson, 1988).

Weights are the values associated with a connective path between two neurons (figure 16) and are used to modify the strength of a signal (Fausett, 1994). Weights should be chosen to minimize the difference between outputs and target outputs. The weights determine the mapping between the input and output vectors and training is basically finding the right weights to make the mapping.

Training may be either supervised or unsupervised. Supervised training requires the input of a training pair (an input vector and a corresponding output vector) where the relationship between the input and output vectors is already known. Unsupervised training does not require prior knowledge of the association between inputs and outputs; the network trains by grouping similar vectors together (Fausett, 1994).

It is important to choose the initial weights carefully. If small weights are chosen then the net input to the hidden or output layer may be close to zero and the network will have trouble learning as the change to weights will be very slight as the activation function or derivative may be zero. On the other hand, overly large weights result in the input signal to the next layer falling into the saturation region of the sigmoid function i.e. where the derivative of the sigmoid function has a very small value (Fausett, 1994). Initial weights are usually small random numbers between two values such as -0.5 and 0.5. The initial weights used by the network for this study were between -3 and 3.

Residual mean of the squares (RMS) error is often used to determine the stopping conditions for training. Training usually continues for as long as the error for the training patterns decreases. As soon as the RMS error stops decreasing then overtraining has occurred; i.e. the net has overtrained on the data and while it will fit the training data very well, it will not fit the testing data well (or subsequently the data you want to make predictions for). In other words, the network has learned each of the input-output patterns but then finds it difficult to generalize.

Too many hidden layers may lead to overtraining of the network; Fausett (1994: 299) considers one hidden layer sufficient "for a backpropagation net to approximate any continuous mapping from the input patterns to the output patterns to an arbitrary degree of accuracy. However, two hidden layers may make training easier in some situations." The number of nodes in the hidden layers are also important, too few nodes and the network struggles to learn, too many nodes and the network learns the training patterns very well but then does not generalize well (Fausett, 1994).

The software package used in this study for the neural network offers a number of net outputs, including graphs of pattern outputs (comparison of actual and net outputs), pattern

errors (on which patterns the network is making errors) and sensitivity (what input variables the network is particularly sensitive to).

7.2) METHODS

The software chosen for this study was a backpropagational neural network programme called WinNN version 0.96 (Danon, 1995) that runs in the Windows environment.

The network was trained to be able to predict where, given certain environmental variables, the four invader species were likely to be present or absent. The relevant environmental value at each presence and absence site for each species was extracted from the five environmental coverages (MAR, COV, MAXT, MINT and ELV). The input data set consists of the environmental variables and a classifying variable (1 for presence and 0 for absence) which tells the net what its output should be, given the environmental variables that go with the classifying variable.

The data were normalised before training and training continued until the net stopped automatically (when the percentage of good patterns = 100) or until it had reached what was hopefully a global minimum. Once the net was trained, the weights and net architecture were saved. These weights were then used in IDRISI to produce a predictive map following the structure of the network (see appendix A for a detailed set of steps and a corresponding network diagram).

The architecture of the networks consisted of an input layer with five nodes, one for each environmental variable, a hidden layer with two nodes and an output layer with one node, either presence or absence (figure 16). A weight noise of 0.004 and an input noise of 0.02 were added to each network to ensure that the net did not learn the training patterns so well that it would not be able to generalise. A sigmoid function was chosen for each network to act as a non-linear function that would squash the values to between 0 and 1.

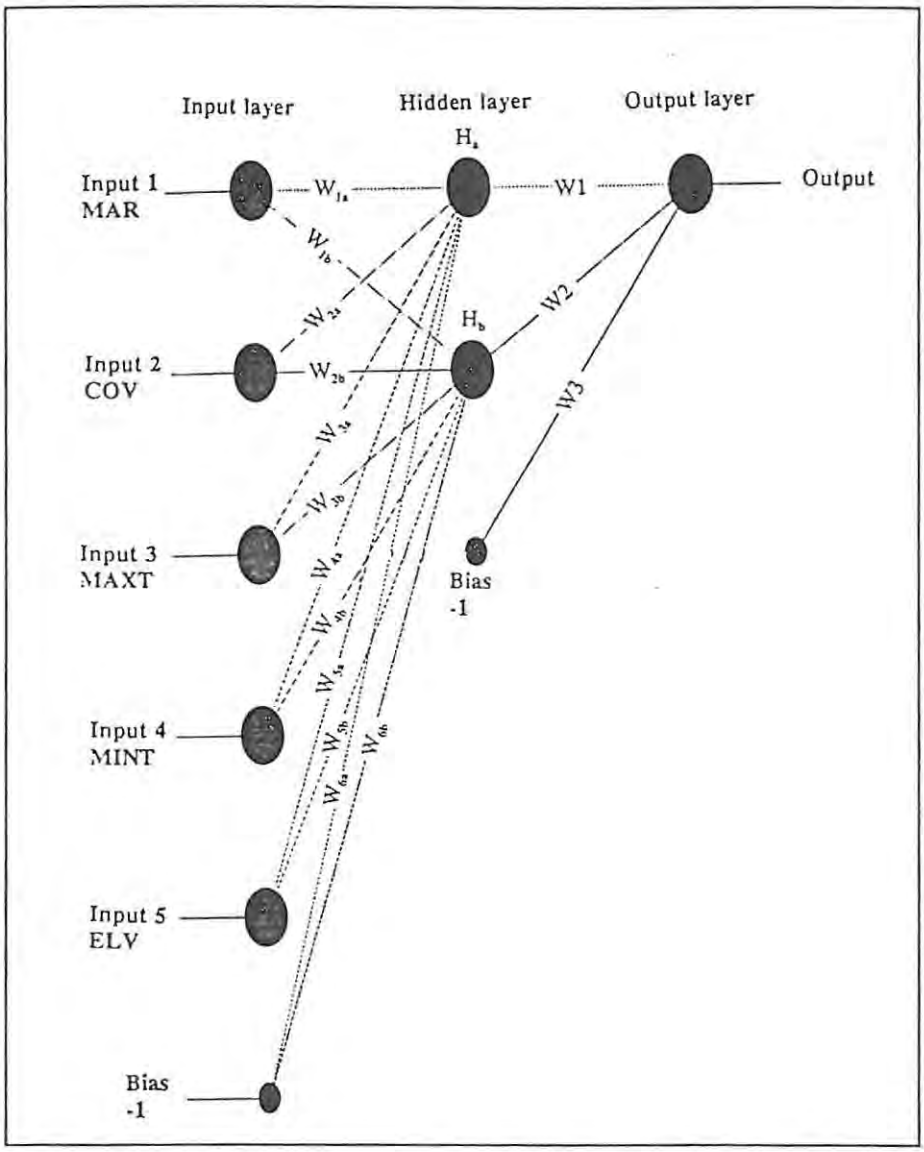


Figure 16. The architecture of the neural networks.

The nets did not train well on the original data set as the discrimination between presence and absence was not very clear i.e. the data was 'dirty' (this is borne out by the DFA classifications, which did not discriminate very cleanly between presence and absence). Therefore another data set for absence was constructed from areas where the plant was definitely known to be absent, such as in the very arid desert regions of South Africa and the net was trained on these data instead.

Once the networks had trained, predictive maps were produced using IDRISI by following the net architecture using the trained weights. See appendix A for the list of steps followed. Briefly, the process proceeded as follows: the input coverages were standardized, multiplied by their respective weights, summed, the bias (multiplied by its weight) added, and the result put through a log-sigmoidal activation function. The result for each hidden layer node was multiplied by weight, the two were summed, and the bias added, to yield the final output.

Testing of the network was carried out to see how well it has trained. Usually the net is tested with some of the data reserved from the training set; but an independent data set (the quarter degrees) was used for testing purposes.

One way chi-squared tests were calculated to determine whether any of the predictive maps were significant departures from randomness. A two way test was not constructed as required testing the net with absence data, which was not readily available (the absence data used to train the nets can not also be used to test them).

7.3) RESULTS

The weights generated by each network (tables 21 to 24) are best interpreted in terms of figure 16 which shows the weighted connection paths between the layers. The second and third columns of the weight tables show the weights given to each input variable for the two hidden layers. The last column is not related to the input variables, but shows the weight output from hidden node 1, hidden node 2 and the bias respectively. The RMS error at which training of the networks stopped is shown in table 25.

Table 21. Weights generated by the ANN for *A. longifolia*. a) refers to weights for the small data set and b) to weights for the extended data set.

a)

	Weights to hidden node 1 ($W_i a$)	Weights to hidden node 2 ($W_j b$)	Weights to output layer (W)
MAR	0.299	2.424	-10.460 (hidden node 1)
COV	1.986	2.935	10.199 (hidden node 2)
MAXT	2.652	0.995	
MINT	-0.706	3.922	
ELV	-1.217	-2.255	
Bias	1.351	-0.309	-4.988

b)

	Weights to hidden node 1 ($W_i a$)	Weights to hidden node 2 ($W_j b$)	Weights to output layer (W)
MAR	-1.111	-0.679	1.994 (hidden node 1)
COV	-2.768	0.256	-9.964 (hidden node 2)
MAXT	-3.395	1.698	
MINT	0.861	-3.531	
ELV	0.691	4.532	
Bias	1.196	1.408	3.334

Table 22. Weights generated by the ANN for *A. mearnsii*. a) refers to the weights for the small data set and b) to the weights for the extended data set.

a)

	Weights to hidden node 1 ($W_i a$)	Weights to hidden node 2 ($W_j b$)	Weights to output layer (W)
MAR	-2.220	-3.061	-2.023 (hidden node 1)
COV	2.606	0.551	-10.590 (hidden node 2)
MAXT	3.278	3.188	
MINT	1.755	-0.539	
ELV	-0.092	2.555	
Bias	2.615	2.386	5.628

b)

	Weights to hidden node 1 ($W_i a$)	Weights to hidden node 2 ($W_j b$)	Weights to output layer
MAR	-2.457	-0.239	-8.925 (hidden node 1)
COV	1.704	2.107	-6.694 (hidden node 2)
MAXT	1.671	6.081	
MINT	0.896	-0.145	
ELV	3.662	-1.692	
Bias	2.864	2.142	5.236

Table 23. Weights generated by the ANN for *O. ficus-indica*.

	Weights to hidden node 1 ($W_i a$)	Weights to hidden node 2 ($W_j b$)	Weights to output layer
MAR	-3.094	3.191	-2.429 (hidden node 1)
COV	-1.511	-4.094	10.431 (hidden node 2)
MAXT	1.544	-1.079	
MINT	-2.294	0.849	
ELV	0.791	1.036	
Bias	1.220	2.318	-3.075

Table 24. Weights generated by the ANN for *S. sisymbrifolium*.

	Weights to hidden node 1 ($W_i a$)	Weights to hidden node 2 ($W_j b$)	Weights to output layer
MAR	-1.915	2.225	-0.749 (hidden node 1)
COV	1.211	-3.103	10.586 (hidden node 2)
MAXT	-2.093	-1.898	
MINT	-1.962	0.656	
ELV	1.947	-0.421	
Bias	-1.034	-3.013	-4.939

Table 25. RMS error at which training of the nets stopped.

		RMS error
A. longifolia	a)	0.0037
	b)	0.0046
A. mearnsii	a)	0.0032
	b)	0.0059
O. ficus-indica		0.0049
S. sisymbrifolium		0.0047

The network output graphs that WinNN generates have set display characteristics and cannot be altered to produce more uniform graphs. The pattern outputs (figures 17a - 22a) show how well the net trained on the input-output patterns. Each training pair is given a number and these are shown on the x axis. The solid line indicates the outputs from the training data (with 1 representing presence and 0 absence on the y axis). The dotted line indicates the net predictions. The closer the two lines conform, the better the network has learned the training patterns.

WinNN (Danon, 1995) produces pattern error graphs that show which training pairs the network made mistakes on (figures 17b - 22b). From the input-output pattern numbers on the x axis, one can tell which of these patterns were responsible for the training mistakes and by how much.

The x axis for the graph of sensitivity refers to the five input variables, ranked as they were put into the network i.e. MAR, COV, MAXT, MINT and ELV (figures 17c - 22c). Points of inflection on the graph indicate the sensitivity at that point for a particular input. For example, in figure 17, the network shows a sensitivity of just above 0.3 to the input of MAR, a sensitivity of 0.7 to COV, 0.5 to MAXT, around 1 to MINT and a sensitivity of approximately 0.35 to ELV.

The predictive coverages (figure 23a - f) produced by the networks have values that range between 0 and 1 and represent the likelihood of each plant species being present (1) or absent (0). This range is represented as continuous shades of grey ranging from white through to

black. The closer the value is to one (black), the greater the likelihood that the plant is present.

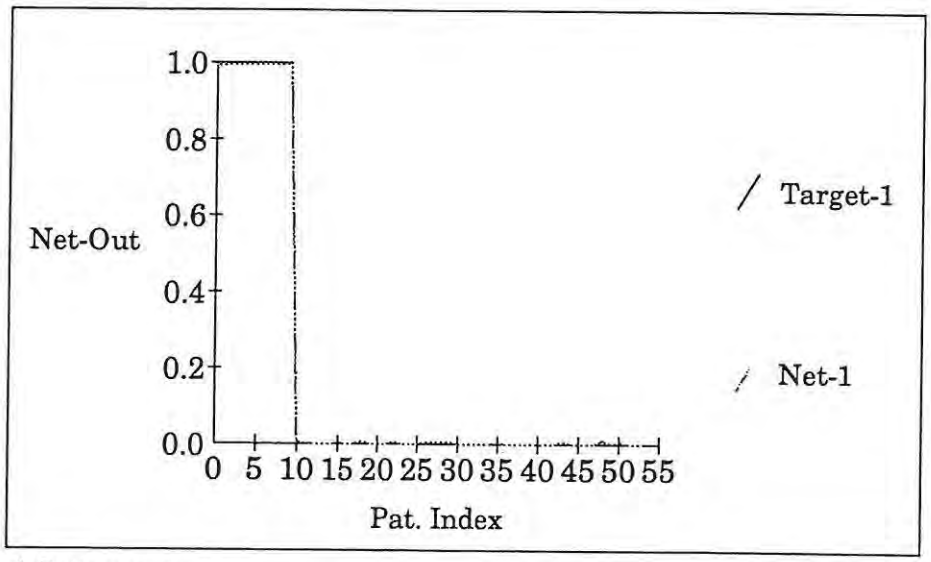
Validation of the networks was carried out by running a test file on the trained nets. The results are given in the form of the percentage of good patterns, i.e. patterns correctly predicted and the RMS error (table 26). Chi-squared tests were also calculated to determine the statistical significance of the maps. Any chi-squared result with a significance level greater than 0.05 was considered to indicate a significant departure from randomness (table 27).

Table 26. Results from running the trained nets on the test files. a) refers to the small data sets and b) to the extended data sets.

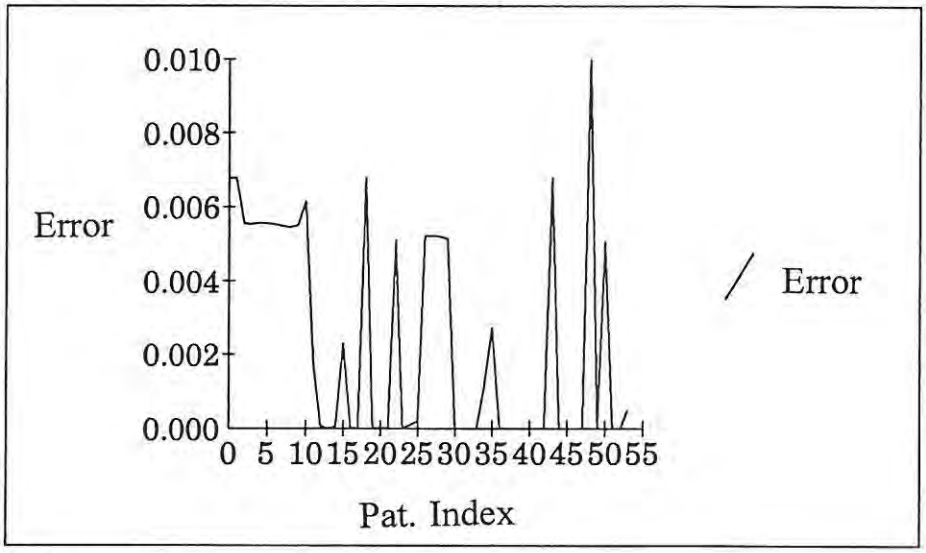
		RMS error	percentage of good patterns
A. longifolia	a)	0.22	72
	b)	0.10	88
A. mearnsii	a)	0.07	74
	b)	0.10	81
O. ficus-indica		0.15	80
S. sisymbrifolium		0.15	54

Table 27. Chi-squared test results and significance levels. a) refers to the small data sets and b) to the extended data sets.

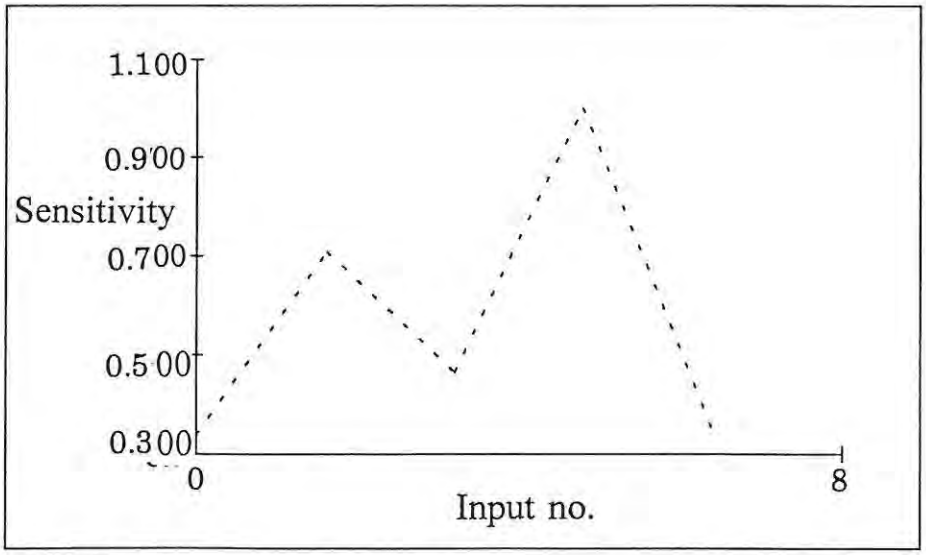
		Chi-squared result	Sig. level
A. longifolia	a)	9.68	0.0019
	b)	28.88	0.0000
A. mearnsii	a)	58.87	0.0000
	b)	98.37	0.0000
O. ficus-indica		151.25	0.0000
S. sisymbrifolium		0.15	0.7003



a) Pattern outputs

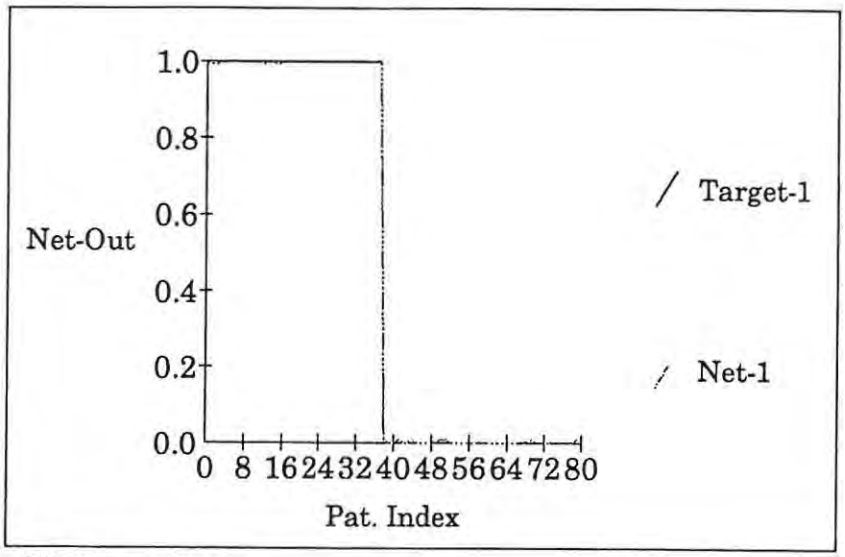


b) Pattern errors

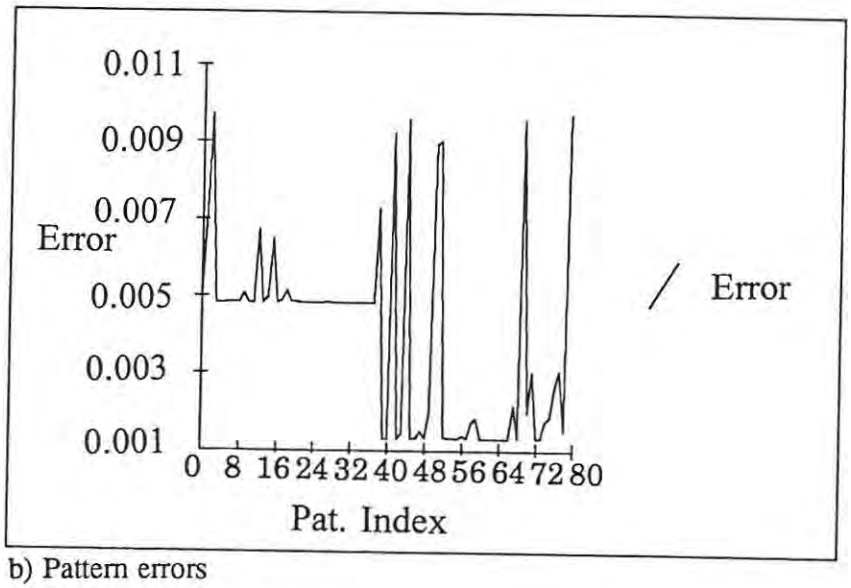


c) Sensitivity

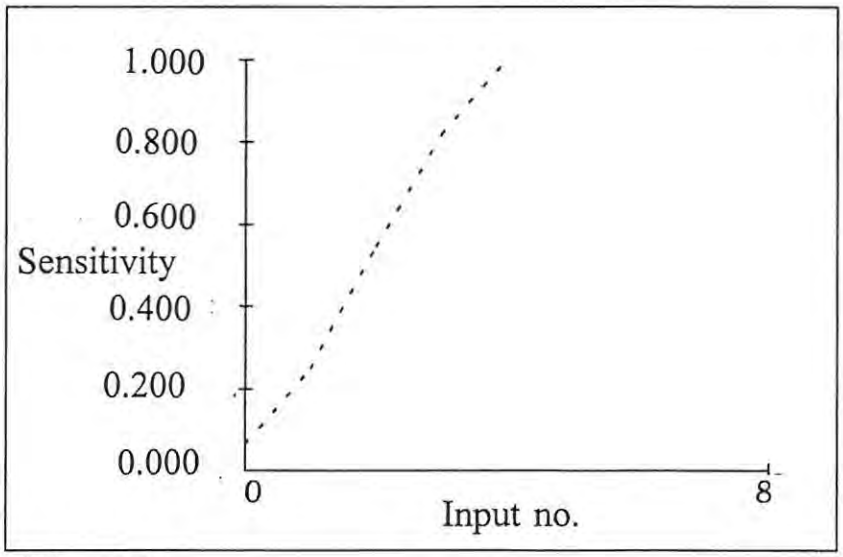
Figure 17. The neural network outputs for *A. longifolia*.



a) Pattern outputs

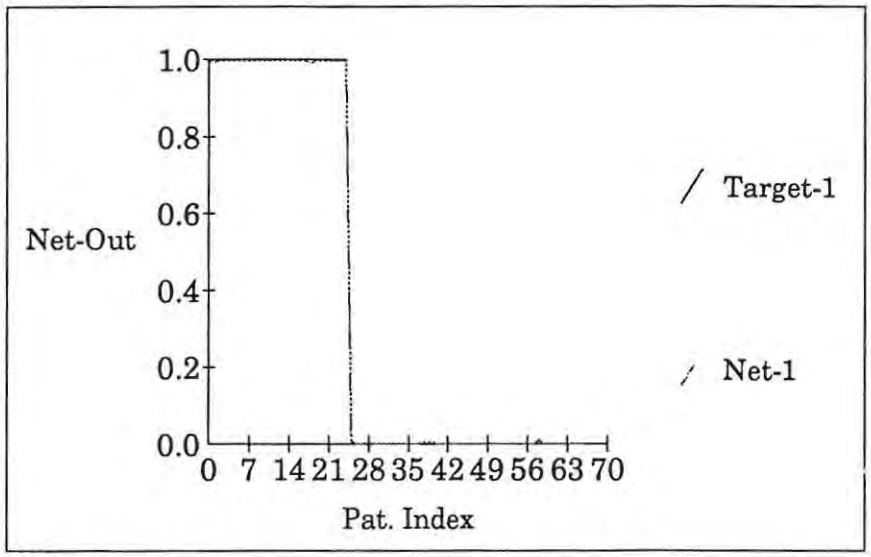


b) Pattern errors

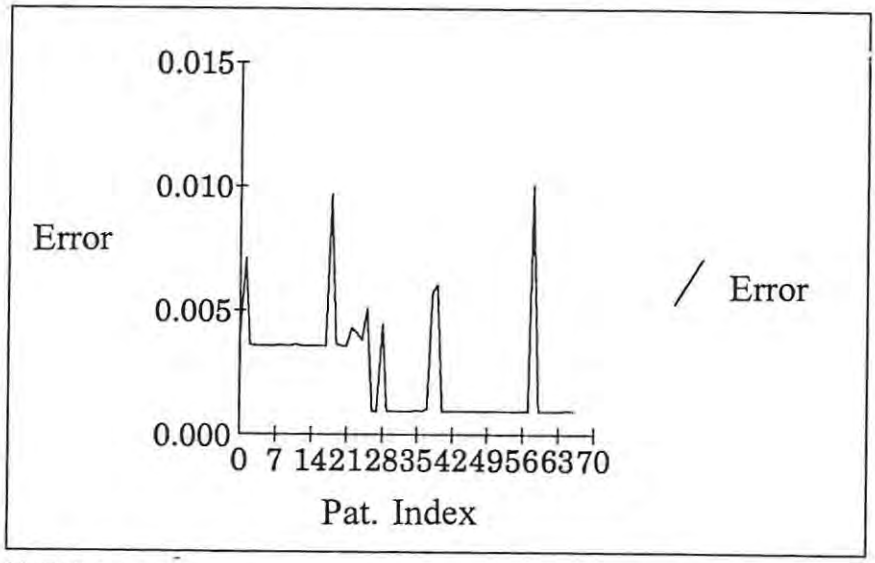


c) Sensitivity

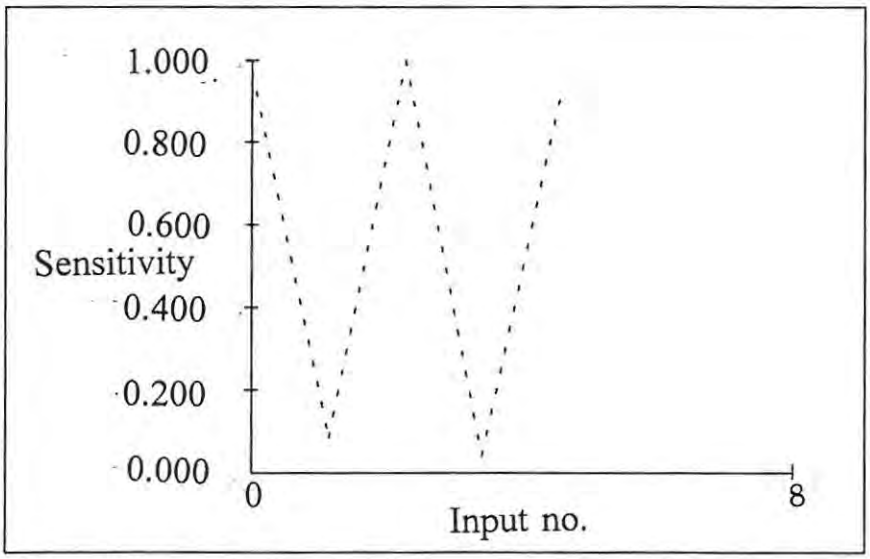
Figure 18. The neural network outputs for *A. longifolia* (extended data set).



a) Pattern outputs

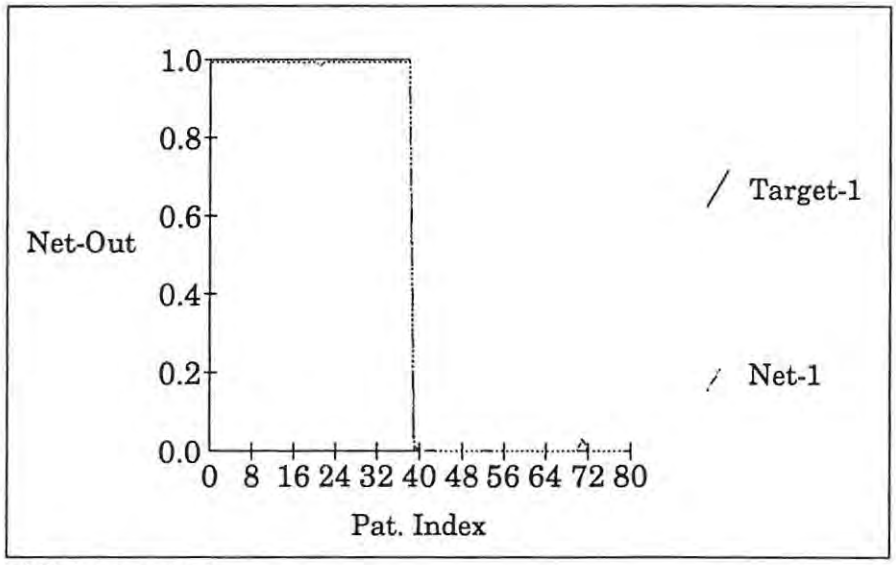


b) Pattern errors

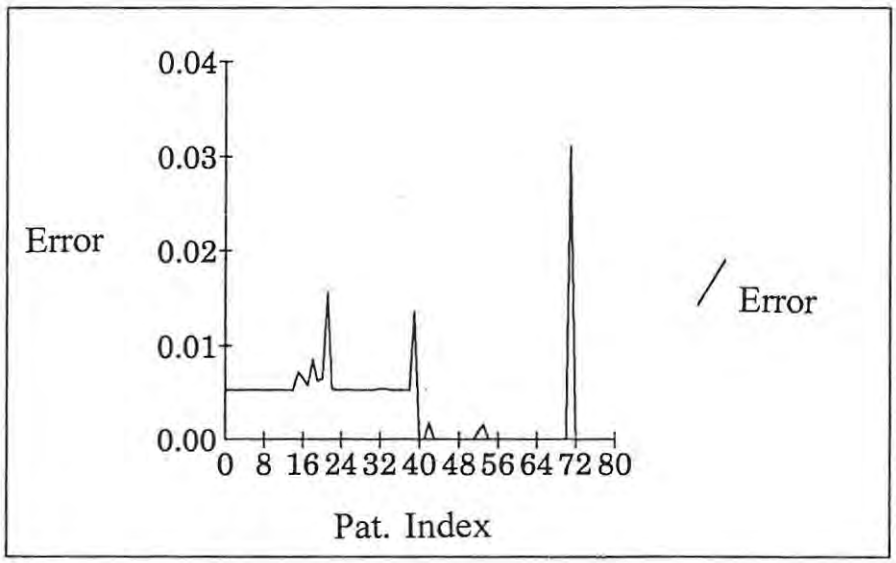


c) Sensitivity

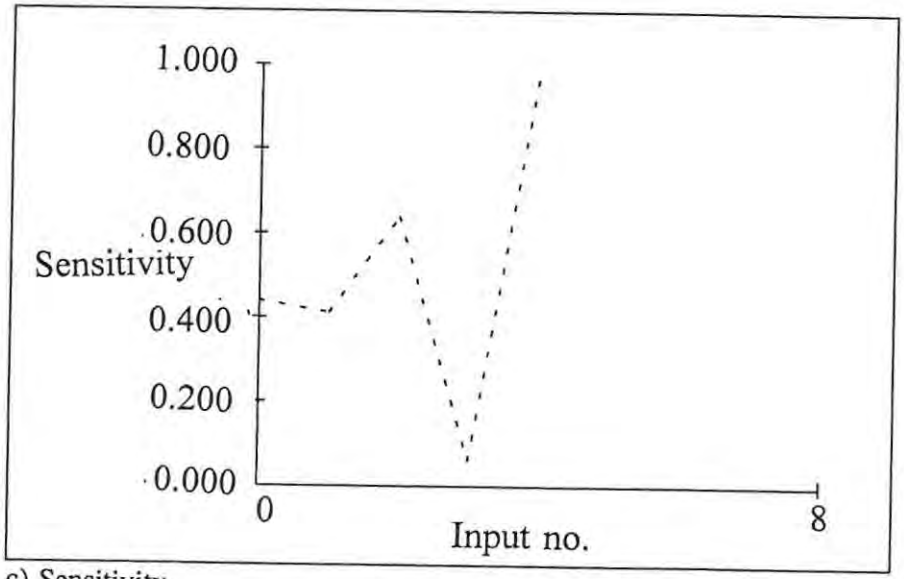
Figure 19. The neural network outputs for *A. mearnsii*.



a) Pattern outputs

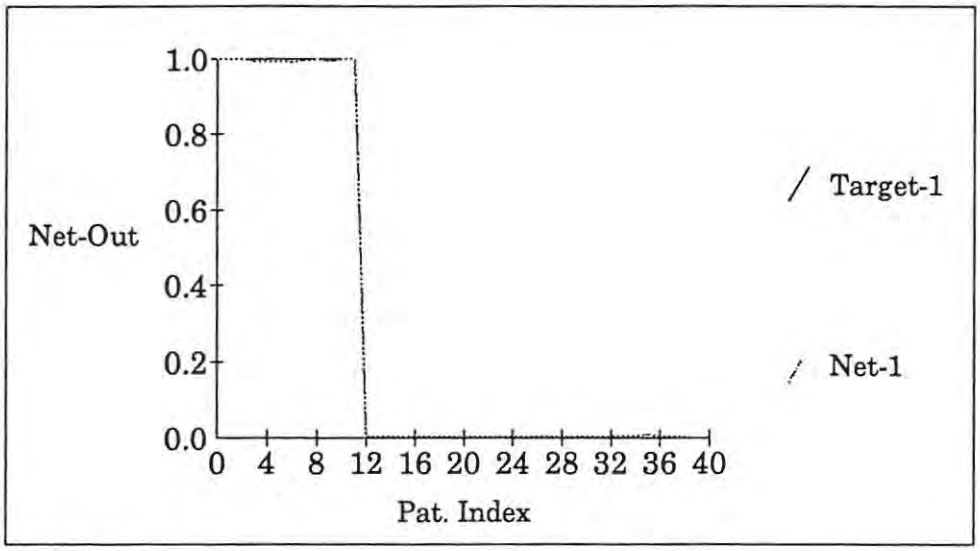


b) Pattern errors

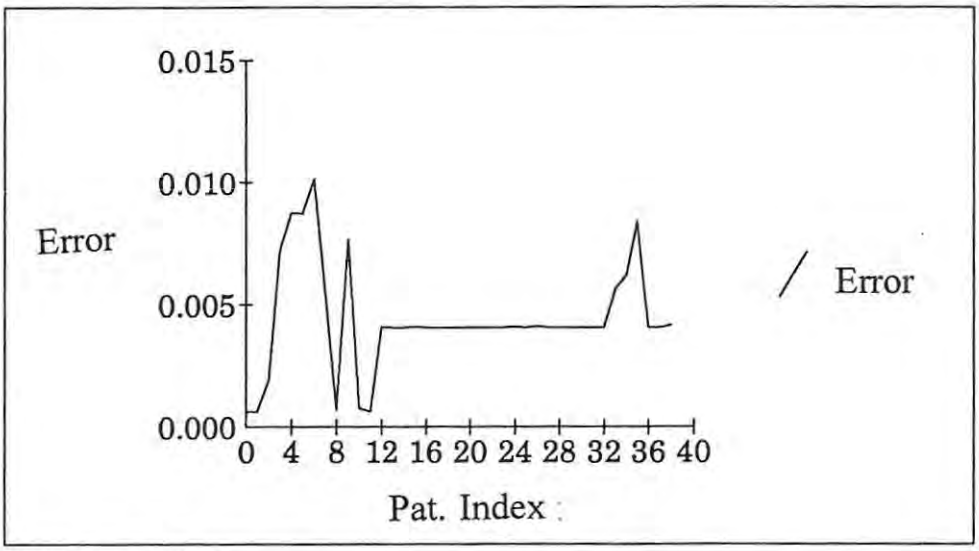


c) Sensitivity

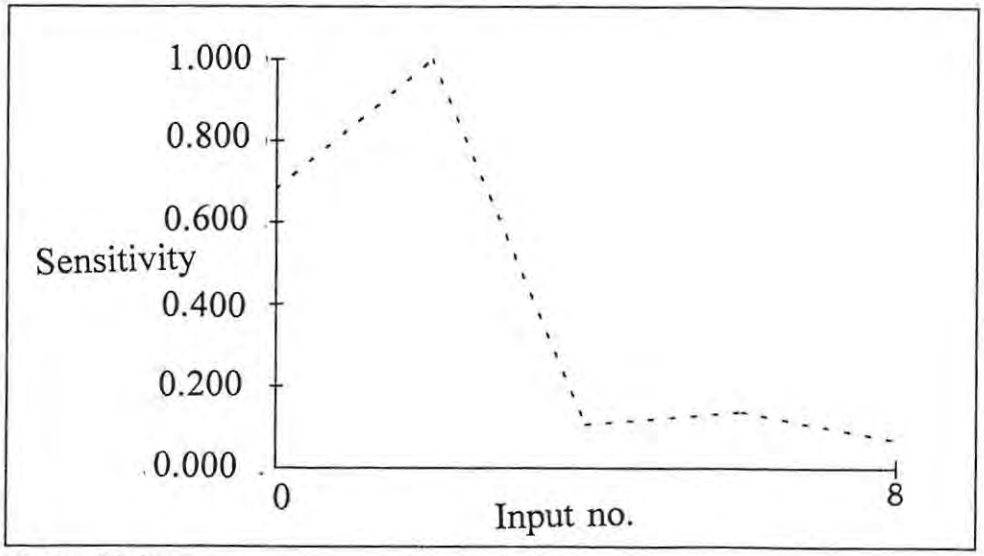
Figure 20. The neural network outputs for *A. mearnsii* (extended data set).



a) Pattern outputs

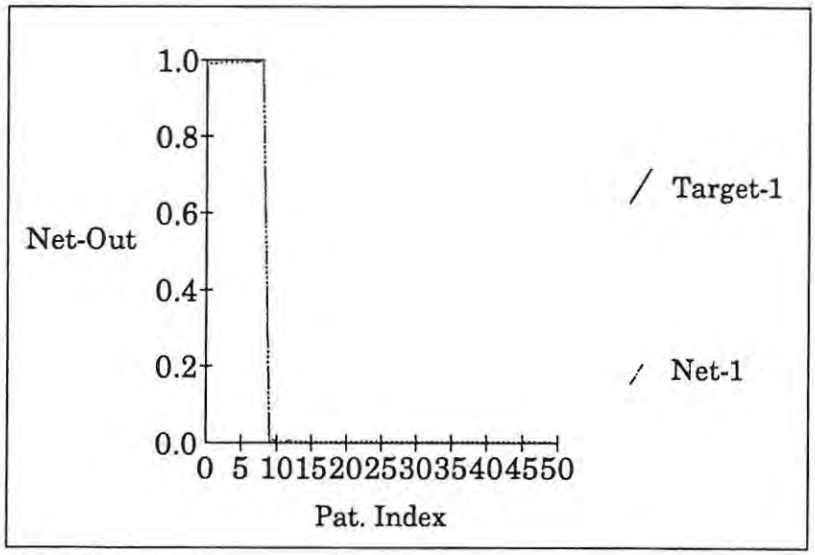


b) Pattern errors

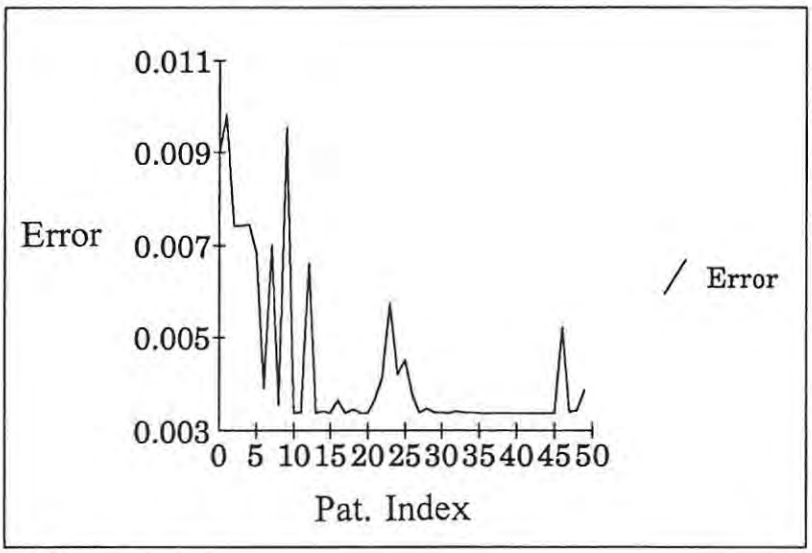


c) Sensitivity

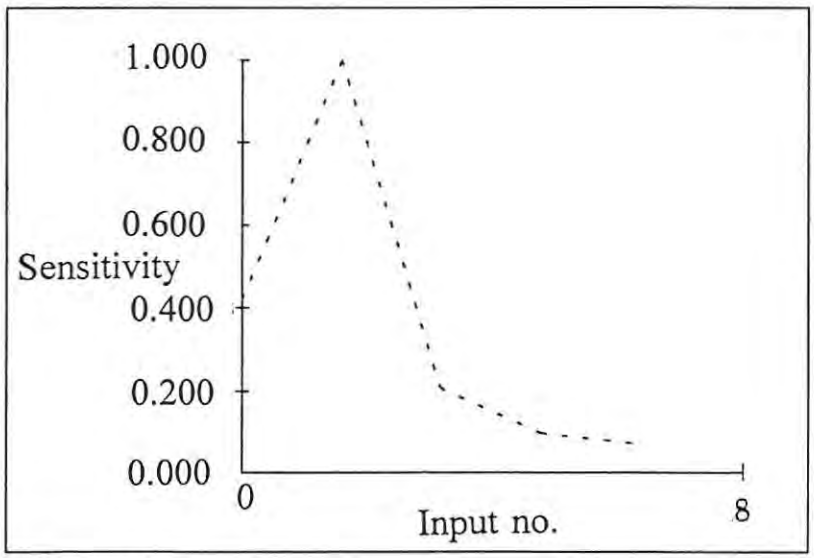
Figure 21. The neural network outputs for *O. ficus-indica*.



a) Pattern outputs

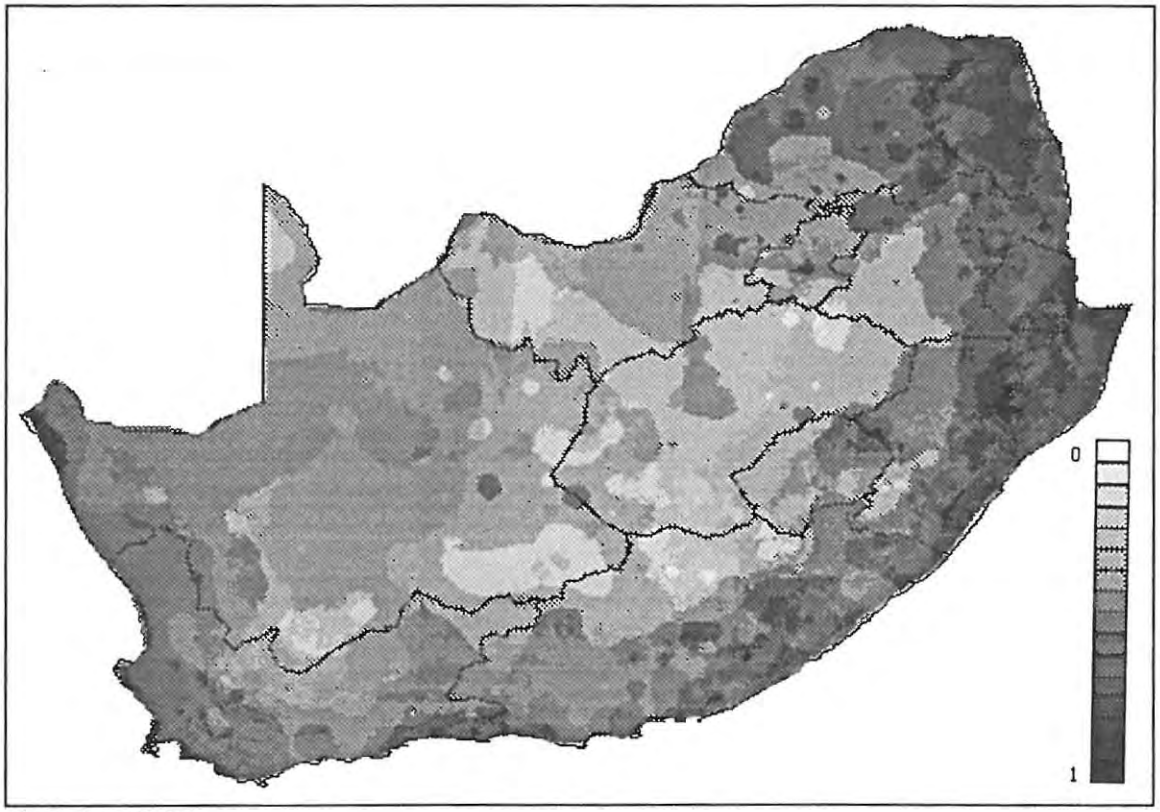


b) Pattern errors

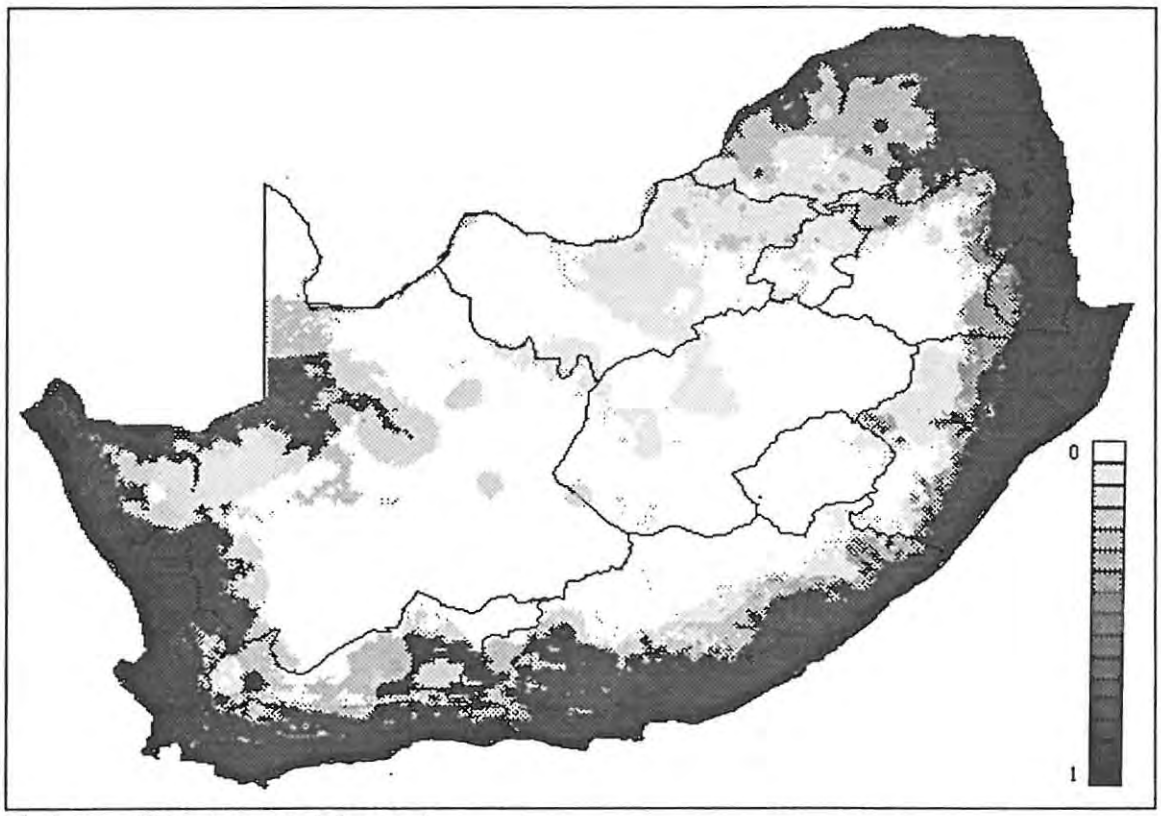


c) Sensitivity

Figure 22. The neural network outputs for *S. sisymbrifolium*.

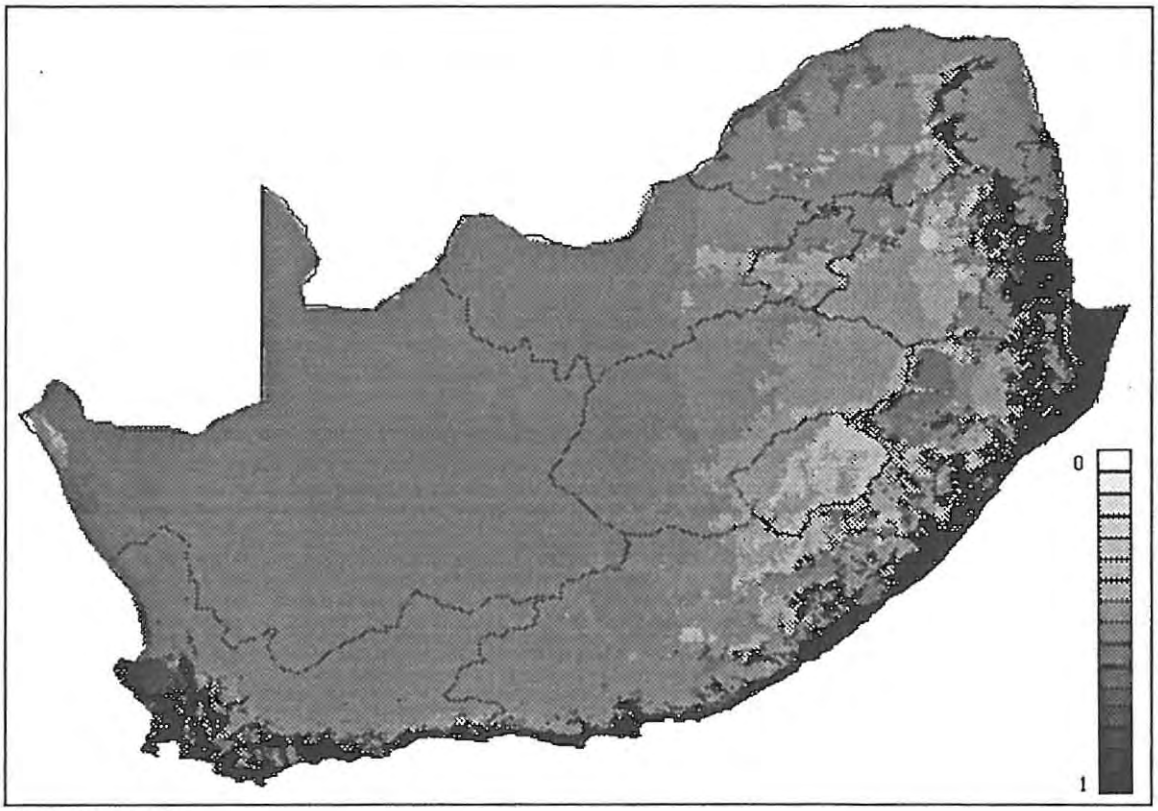


a) *A. longifolia*

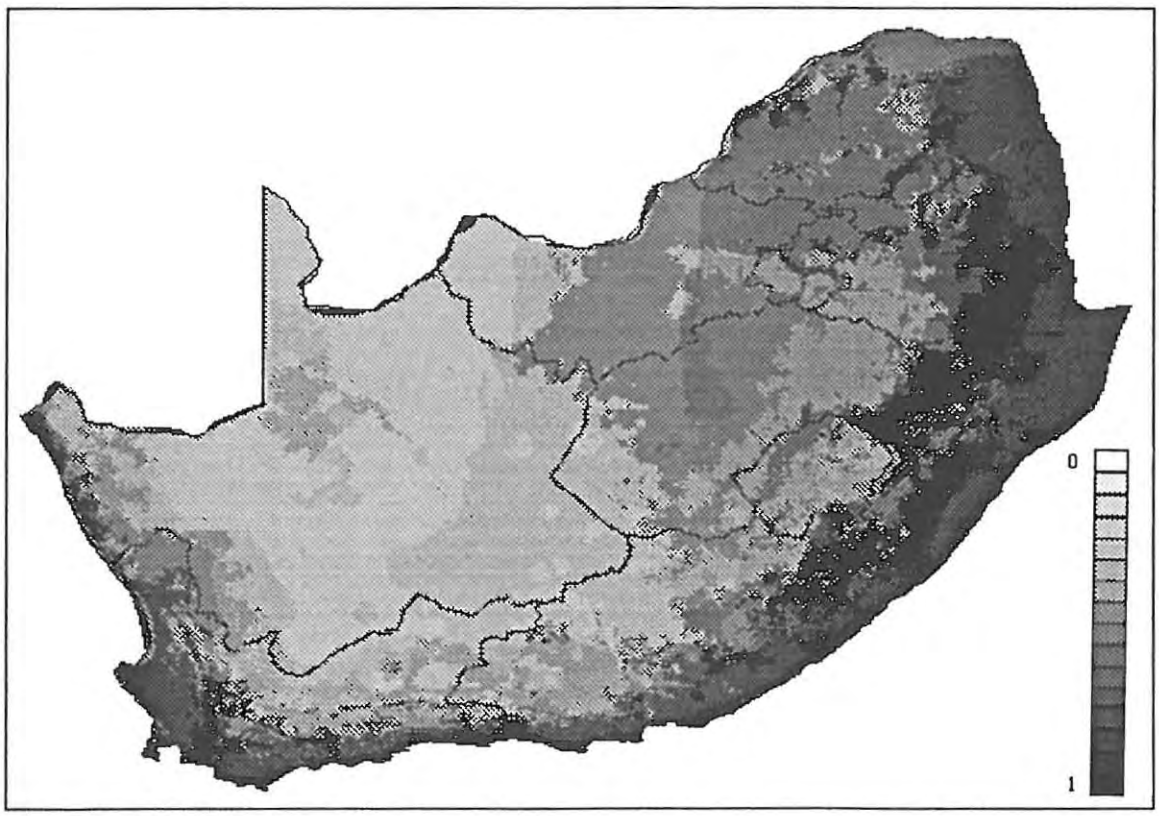


b) *A. longifolia* (extended data set)

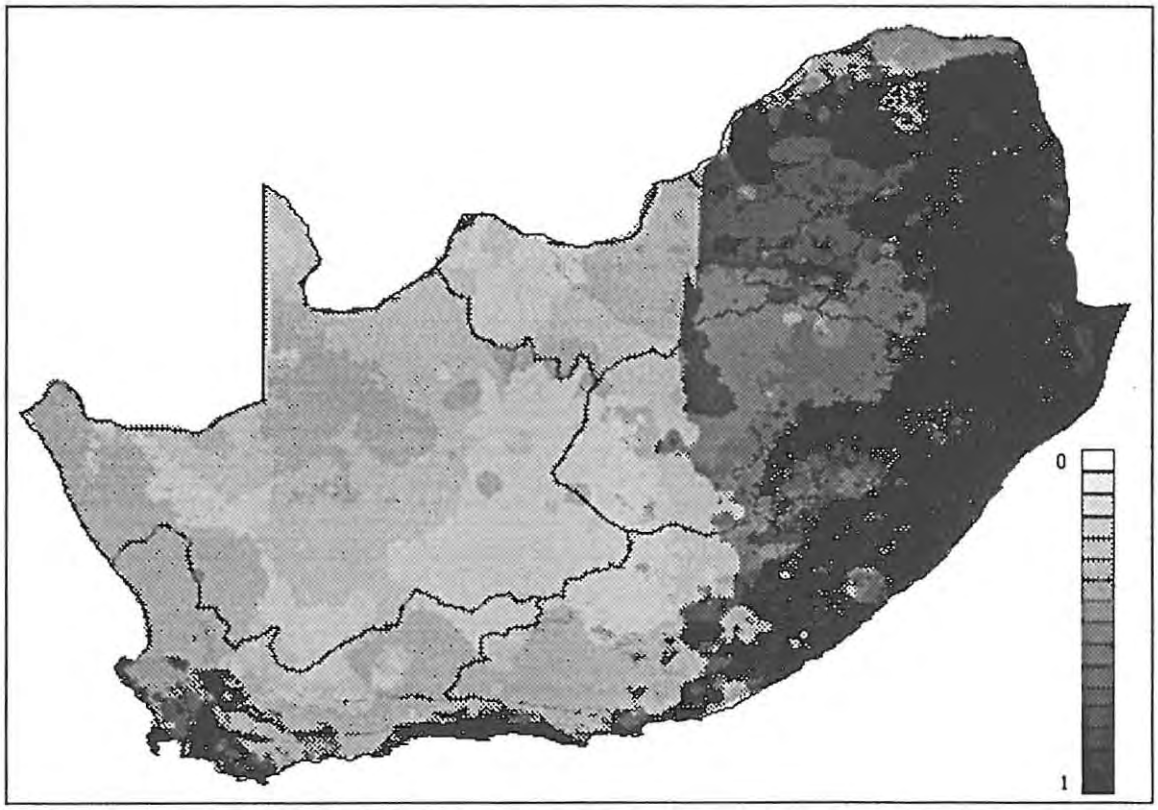
Figure 23. Potential distribution maps derived from the artificial neural networks.



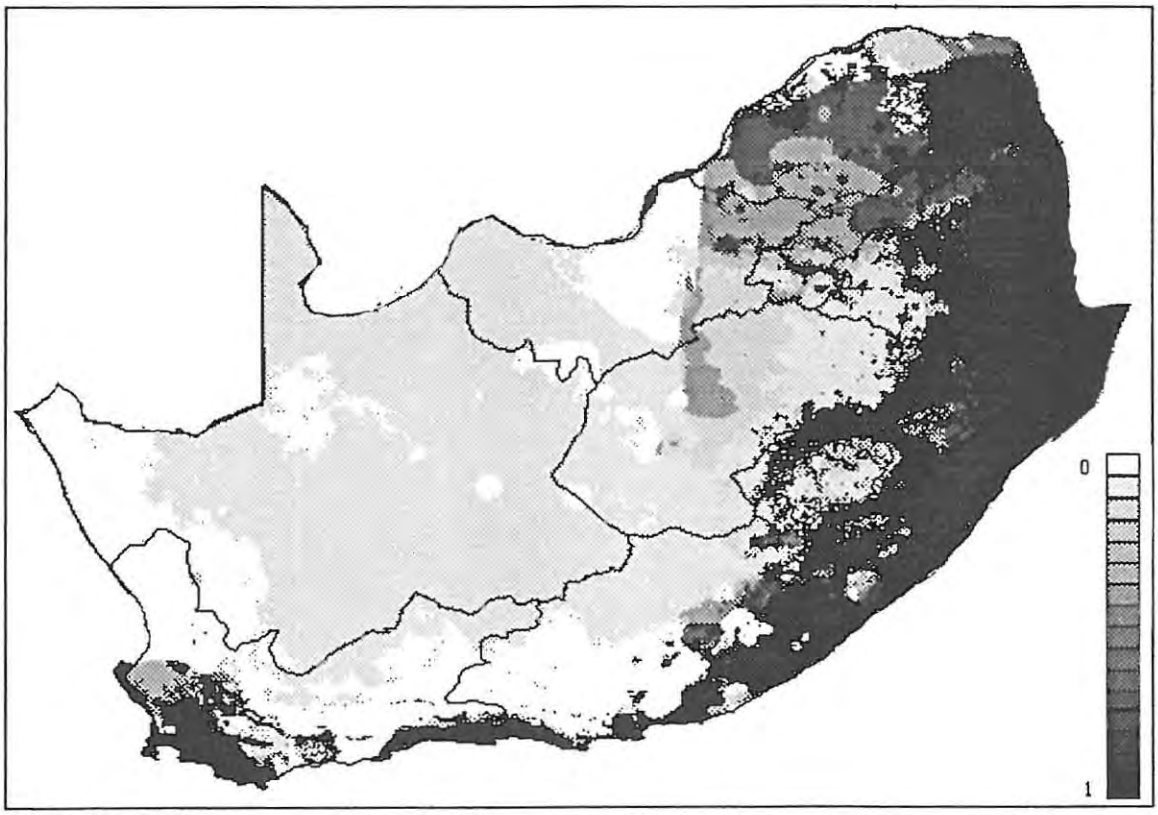
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

7.4) DISCUSSION

The function chosen for a network is important and depends on the end results desired. The function acts to squash input values into a certain range. WiNN (Danon, 1995) offers a choice of three functions; sigmoidal, linear and tan. The sigmoidal function was chosen as it is non-linear and would squash the end results from plus to minus infinity into a range between 0 and 1 (Demuth & Beale, 1994). This resulted in predictive coverages that had values ranging between 0 (plant absent) and 1 (plant present).

The coverages produced for *A. mearnsii* and *O. ficus-indica* also produced good validation results (table 26) and the chi-squared results indicate that the maps produced for these two species and for *A. longifolia* are all significantly different from maps produced by random. However, examination of the coverages (figure 23 a - f) suggest in many cases that the network is predicting some areas of presence where the plants should be absent, such as in the Kalahari Desert.

These areas of unlikely predicted presence, for example in the Kalahari, are probably a result of misclassification by the network of absence data points. In other words, the network incorrectly classified some of the absence points as areas of presence which is why the plant is shown to occur in an area of known absence. It is easy to pick out misclassified areas in this case as one knows that the chance of the species occurring in the Kalahari is very slight. However with absence data recorded in the same study areas as the presence data, it may be more difficult to pick up such misclassification errors.

It is interesting to note that some of the coverages produced by this network method show similar patterns between species in parts of the country. For example figure 23a, b, d and f shows an identically shaped patch of distribution in the centre of the Free State. Figure 23b, d and f demonstrate similar areas of distribution near the Kalahari. Closer analysis of the input coverages indicates that the patches along the Orange River and in the Kalahari Desert could be artifacts from the elevation coverage. The coverages leaving the most artifacts on the images appear to be linked to which inputs the network is sensitive to. For example, for the extended data set for *A. longifolia*, the network is most sensitive to ELV (table 21b); examination of the predictive map (figure 23b) shows that most of the artifacts are from the

original input ELV coverage. Similarly, the predictive maps for *O. ficus-indica* and *S. sisymbrifolium* show artifacts from the COV image, to which the network was the most sensitive for these two species (figure 23e and f).

The weights generated by the network give an indication of which input requires the greatest weighting. For *A. longifolia*, MINT was given the greatest weighting for the small data set and ELV for the extended data set. MAXT was the most heavily weighted input for both data sets for *A. mearnsii* and COV for both *O. ficus-indica* and *S. sisymbrifolium*.

The sensitivity graphs also give an indication of which input variable is important in distinguishing between presence and absence. For *A. longifolia* the network appeared to be the most sensitive to the input variables of MINT and ELV (for the small and extended data sets respectively) the same as for the weights. This species comes from temperate coastal regions of Australia (Stirton, 1987) which is probably why these two variables are of importance in distinguishing between areas of presence and absence, as the plant is unlikely to occur in high or cold areas. The PCA analysis also indicated that MINT is an important determinant of presence or absence for this species.

The sensitivity analysis for the small data set for *A. mearnsii* (figure 19) indicated that MINT was an important variable in distinguishing between presence and absence (as it was for this species in the PCA). For the extended data set, ELV (figure 20) was the input that the network was most sensitive to. The weightings suggest that MAXT is also important (table 22a and b)

For both *O. ficus-indica* and *S. sisymbrifolium*, COV is the input variable that the network is the most sensitive to in making distinctions between presence and absence. This was also the case for the weightings.

With regards to the validation results, three of the four species showed a high rate of success in using the neural network to predict distribution, with the success rate ranging between 72% and 88% (table 26). Only the distribution for *S. sisymbrifolium* was not very accurately predicted. This may be because the distribution of *S. sisymbrifolium* may depend on an environmental variable not used here, because the sample used was too small or because this

species is a relatively recent invader in South Africa.

The network appeared to predict the distribution of the two *Acacia* species more accurately with larger sample sizes, with both species showing an improvement in the number of sites correctly predicted as present using the extended data sets. However, the result for *O. ficus-indica* (80%) was excellent, despite the small sample size for the species. The result for *S. sisymbrifolium* is not as good as the results for the other three species, as has been the case with the other techniques.

7.5) CONCLUSIONS

To train effectively, the backpropagational network required training pairs that showed a clear distinction between the two outputs desired (i.e. presence or absence). Misclassification of presence or absence during training resulted in the network predicting a few areas of presence on the coverages where the absence of data was drawn from such as in the Kalahari Desert. As such, it is easy to pick out these areas of incorrect predicted presence; however, in cases where the presence and absence data is not so clearly separated (e.g. drawn from the same area) then these areas of misclassification would be much more difficult to pick out.

According to the weights and sensitivity analysis, the input variables that were important in distinguishing between presence and absence for *A. longifolia* were MINT and ELV. COV was the determining variable for both *O. ficus-indica* and *S. sisymbrifolium*. The important distinguishing variables for *A. mearnsii* differed slightly, with the weights indicating that MAXT was important and the sensitivity analysis that MINT and ELV were.

By showing which inputs the network is most sensitive to, the sensitivity analysis also indicates which input variables leave the most artifacts on the predictive coverages.

This technique produced good results (72% to 88%) for three of the four species. All of the maps were significant departures from randomness except for the one produced for *S. sisymbrifolium*. It may be that the number of training patterns for this species were not sufficient to allow the network to make good generalizations. Although the network achieved

good prediction for three of the four species, it proved to be a complex and somewhat time-consuming technique given the current software and methods used.

CHAPTER 8
FUZZY LOGIC

8.1) INTRODUCTION

One of the underlying principles in the conventional mathematical theory of sets is the law of the 'excluded middle'. According to this law, elements either belong to a set or not; statements are either true or false (Kosko & Isaka, 1993). In other words, there is a very crisp distinction between sets and elements either belong or do not; there is no middle ground. Unfortunately, in reality, there are seldom such clear-cut definitions between sets or groups of objects. Many variables are continuous rather than categorical, and there are often areas of uncertainty or 'greyness' when defining sets for particular objects. For example, we may say that a plant enjoys a wet environment, but it is extremely difficult to define what a 'wet' environment is; what is regarded as 'wet' in an arid area is obviously different to what is regarded as 'wet' in a tropical rainforest. So, while the plant belongs to the set of plants enjoying a wet environment, we cannot precisely define that set; one cannot say that at, for example, 800 mm of rainfall a year an environment becomes suddenly and categorically wet.

One of the ways of dealing with this problem is through use of fuzzy sets and fuzzy logic (Kosko & Isaka, 1993). Fuzzy set theory was first introduced in 1965 by Zadeh (Zadeh, 1987), and is designed to cope, in a mathematically precise way, with uncertainty or fuzziness. With regards to our example, fuzzy logic would allow us to have differing degrees of 'wetness', such as 'very wet', 'not so wet' or 'slightly wet'. Instead of either belonging to a set or not, an object may have varying grades of membership (Zadeh, 1987). However, these grades of membership must still be defined and, with reference to the example, one would have to decide how to define the fuzzy functions for 'very wet' and 'not so wet'. Definition of the fuzzy functions is usually done through means of expert opinion (Kosko & Isaka, 1993). This can provide an opportunity for local people who are affected by the variables being studied to contribute to the study as they are often the people with the expert knowledge necessary to define the fuzzy sets (Thomas & Sun, 1995).

Being able to define grades of membership has important implications particularly for ecological modelling. Often ecological data roughly approximate a normal distribution i.e.

there is a range that is suitable for the organism (Putman & Wratten, 1984). Above or below that range, the conditions are unfavourable for the organism; but in between these extremes are varying degrees of habitat suitability. Conventional set theory does not allow for the concept that some parts of the range are more suitable than others, whereas fuzzy logic can.

This ability to model non-linear systems is the main reason why fuzzy logic was chosen as one of the predictive techniques. Plant responses to their environment are usually assumed to be non-linear i.e. they are assumed to have gaussian response curves and not linear ones (Putman & Wratten, 1985). Most conventional statistical modelling techniques such as DFA and PCA assume linear relationships. In theory a non-linear system should be better able to model plant distribution according to environment than the linear systems. Thomas and Sun (1995) state that fuzzy techniques are also able to cope with incomplete data sets and crude data.

Much of the current research in fuzzy logic is on producing 'smart appliances' (Cole, 1995). These are appliances such as washing machines that adjust wash and rinse cycles according to how dirty the clothes are and toasters that determine how long to toast a slice of bread for by judging its thickness, how fresh it is and whether it has been frozen or not (Cole, 1995). Fuzzy logic has also been used to programme systems to control subways in Japan and to draw up optimum health plans for employees based on their present health (Kosko & Isaka, 1993). Fuzzy techniques do not, as yet, appear to have been utilised extensively in the ecological field. Thomas and Sun (1995) used fuzzy sets for rangeland assessment to help them predict how rainfall and soil variables affected water availability to plants. Their fuzzy technique produced a 15% to 20% improvement in their prediction rate, which they attribute to the fact that the fuzzy classification took interactions between their variables into account.

8.2) METHODS

IDRISI has a module for fuzzy classification termed FUZZY. To operate, FUZZY requires four control points (figure 24):

a, where the membership function begins to rise above zero

b, where the membership function reaches one

c, where the membership function begins to descend below one

d, where the membership function approaches zero

Points a, b, c and d were determined by constructing frequency histograms for each environmental variable for the four species from their presence data. These four control points were used to define the fuzzy set function control points for each environmental variable. A fuzzy coverage for each environmental variable was produced and these coverage were overlaid using the minimum operator, which chooses the pixel with the minimum value for corresponding coverages (Eastman, 1994). The end result is a fuzzy map that shows the degrees of possibility that the plant will be present (this ranges between 0 and 1, the closer the value is to 1, the greater the possibility that the plant is present).

The function that is chosen for the fuzzy set is of great importance. The shape of the function affects the grade of membership. A good general rule is that the 'cleaner' the data, the sharper the function can afford to be (Burton, 1996 pers com). The term 'clean' data is applied to data that allows boundaries to be cleanly defined; for example if a particular plant species is known to have a very restricted habitat and the environmental conditions it will tolerate are well known, then we can define with more certainty its membership function. If, on the other hand, a plant species tolerates a wide range of environments then it is more difficult to define a function clearly. FUZZY offers a choice of three functions: sigmoidal, j-shaped or linear. In keeping with the assumption of a non-linear response by the plant to its environment, a smooth, non-linear sigmoidal function was chosen for use in the fuzzy classification.

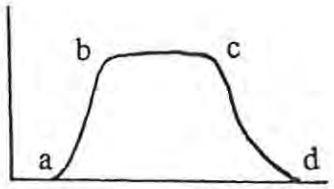


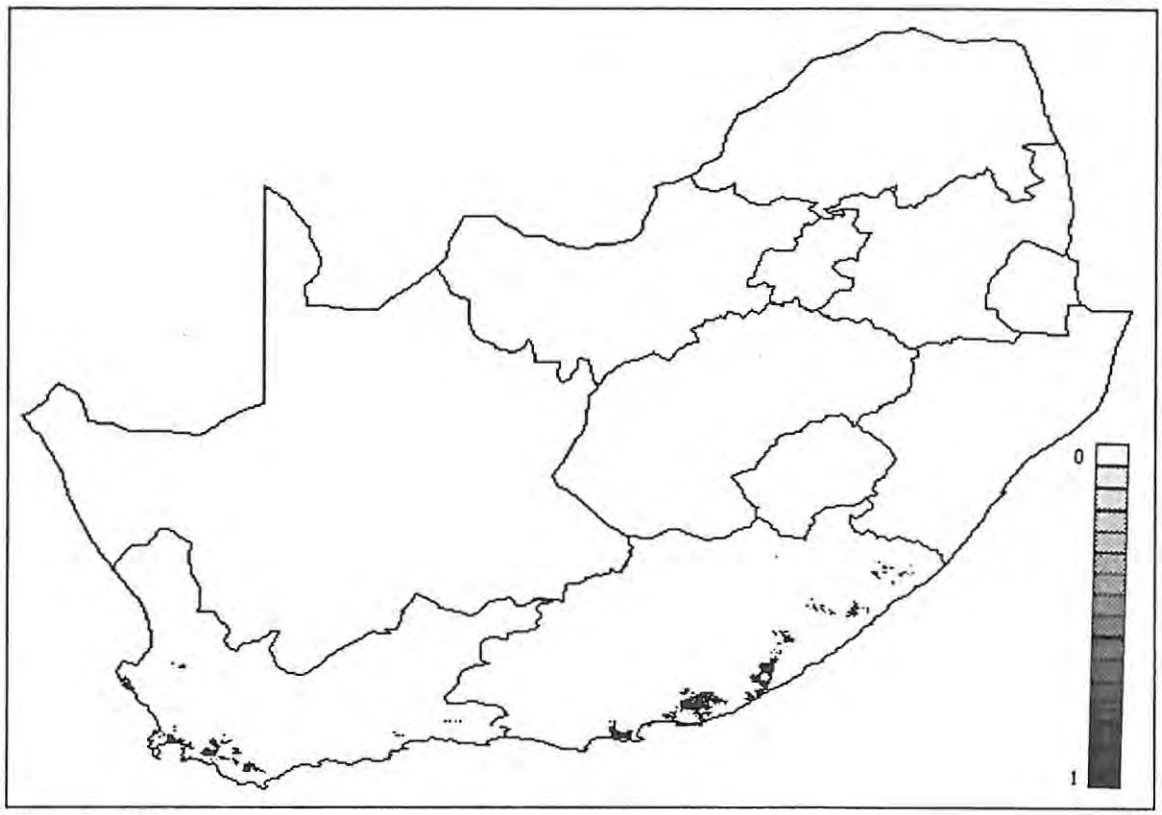
Figure 24. The four control points required by FUZZY for the sigmoidal function.

Validation of the predictive fuzzy maps was carried out by extracting the value of each pixel on the predictive map from the validation data set. All sites that showed a greater than 0.5 possibility of the plant being present were counted as correct predictions of presence. A two by two contingency table with Yates' correction was constructed to determine the statistical significance of the predictions.

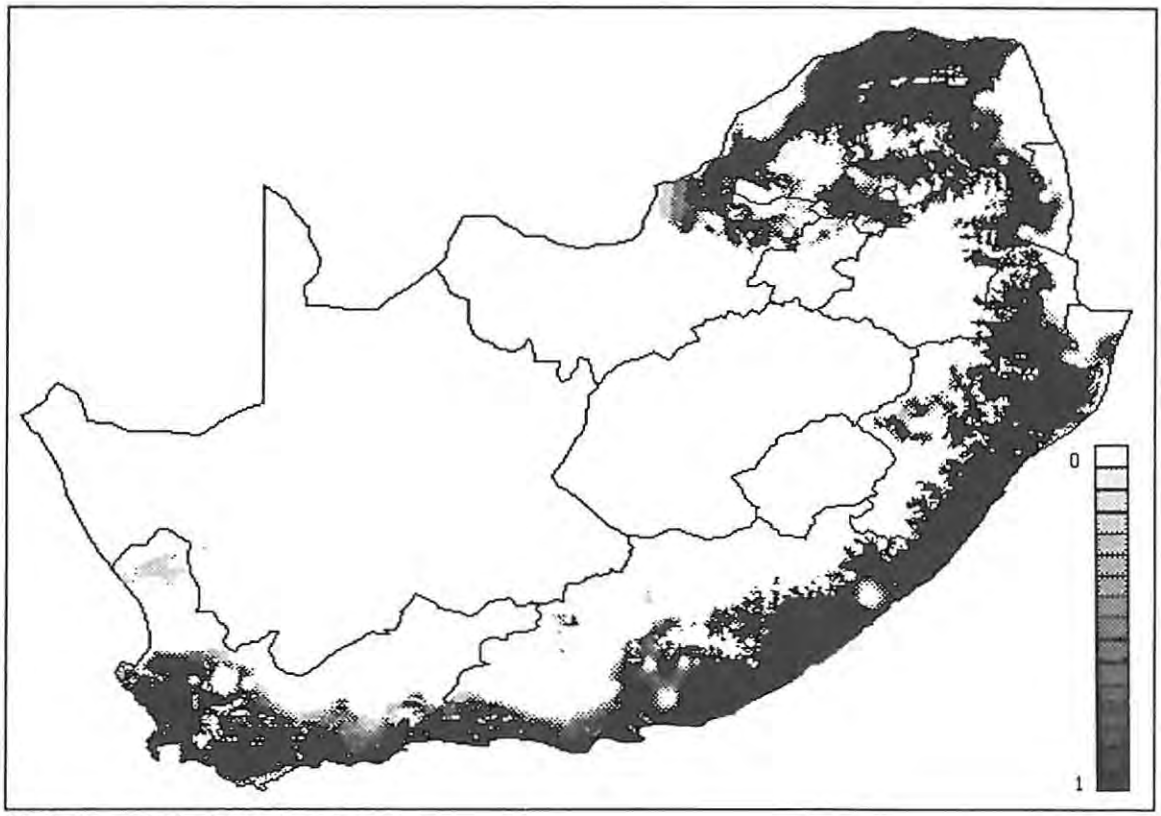
8.3) RESULTS

The coverages show the possibility of the plants being present. The possibility values range between 0 and 1 and are represented as shades of grey; the fuzzier the shade, the lower the possibility of the plant being present (figure 25a - f).

The validation results are expressed as the percentage of sites correctly predicted as present, where any site with a predicted possibility of presence of 0.5 or higher was counted as a correct prediction of presence (table 28).

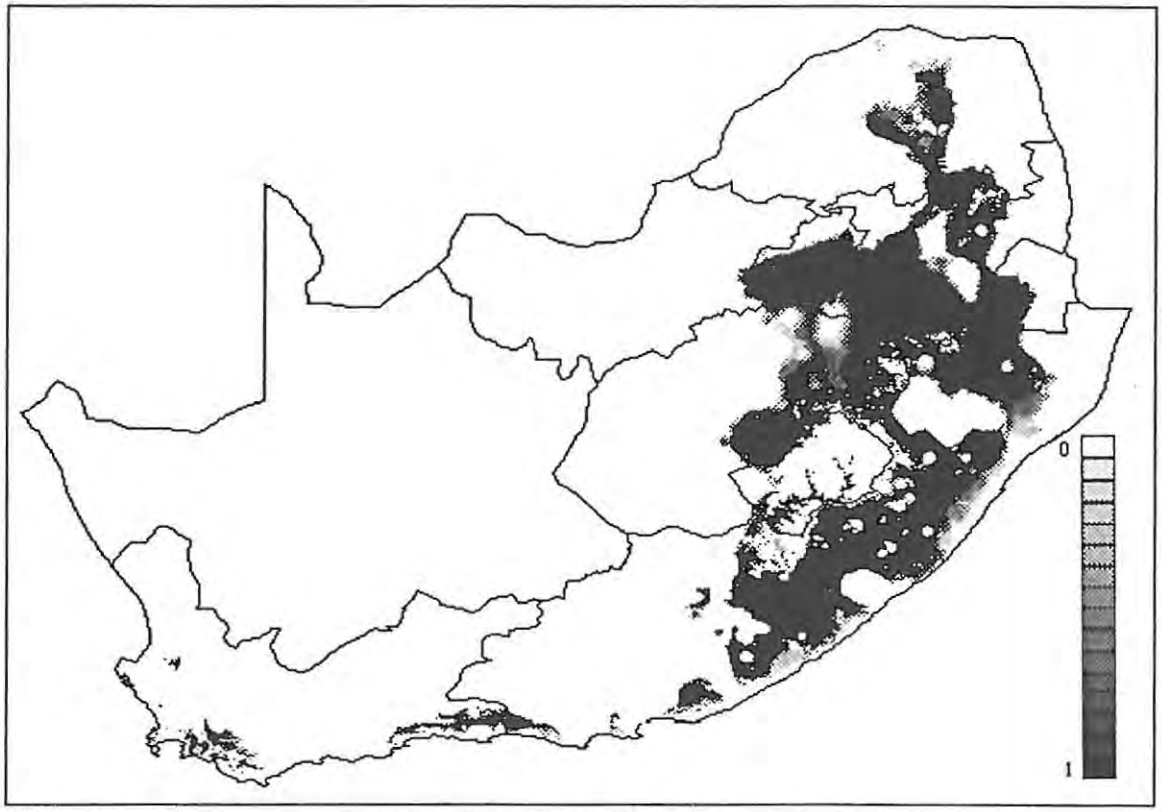


a) *A. longifolia*

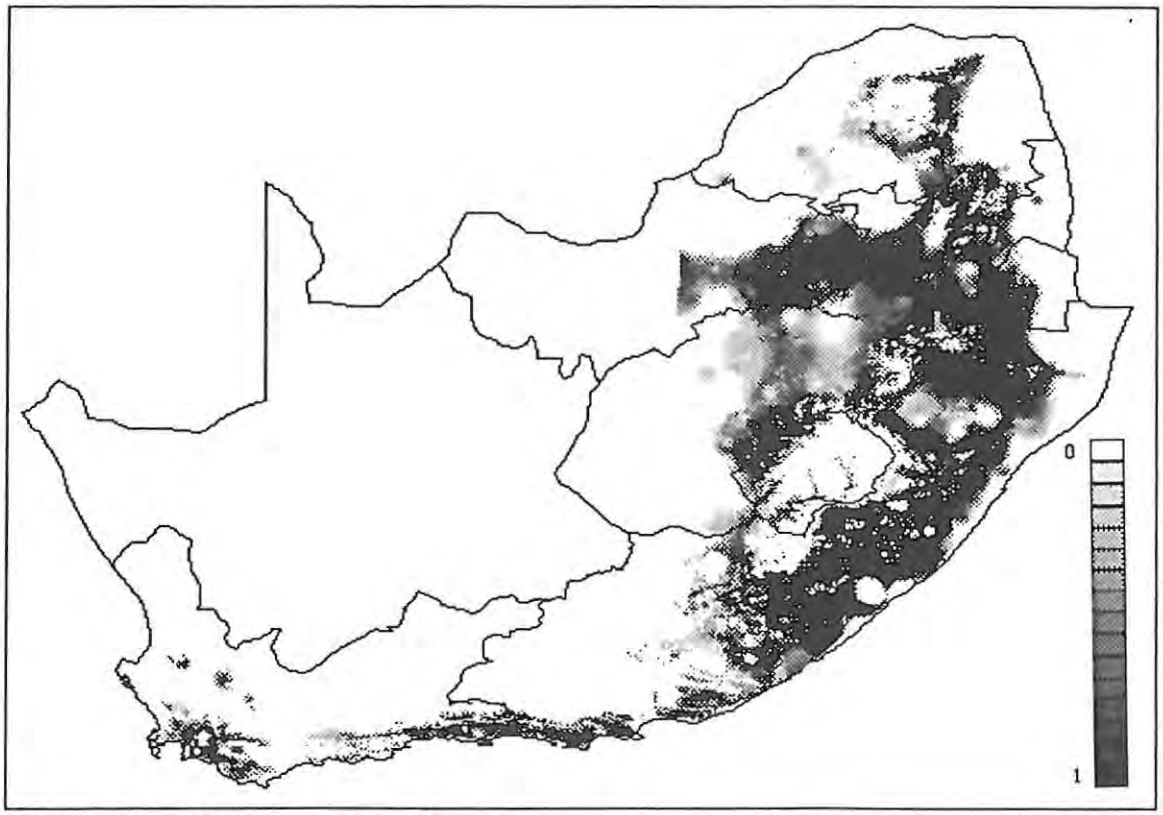


b) *A. longifolia* (extended data set)

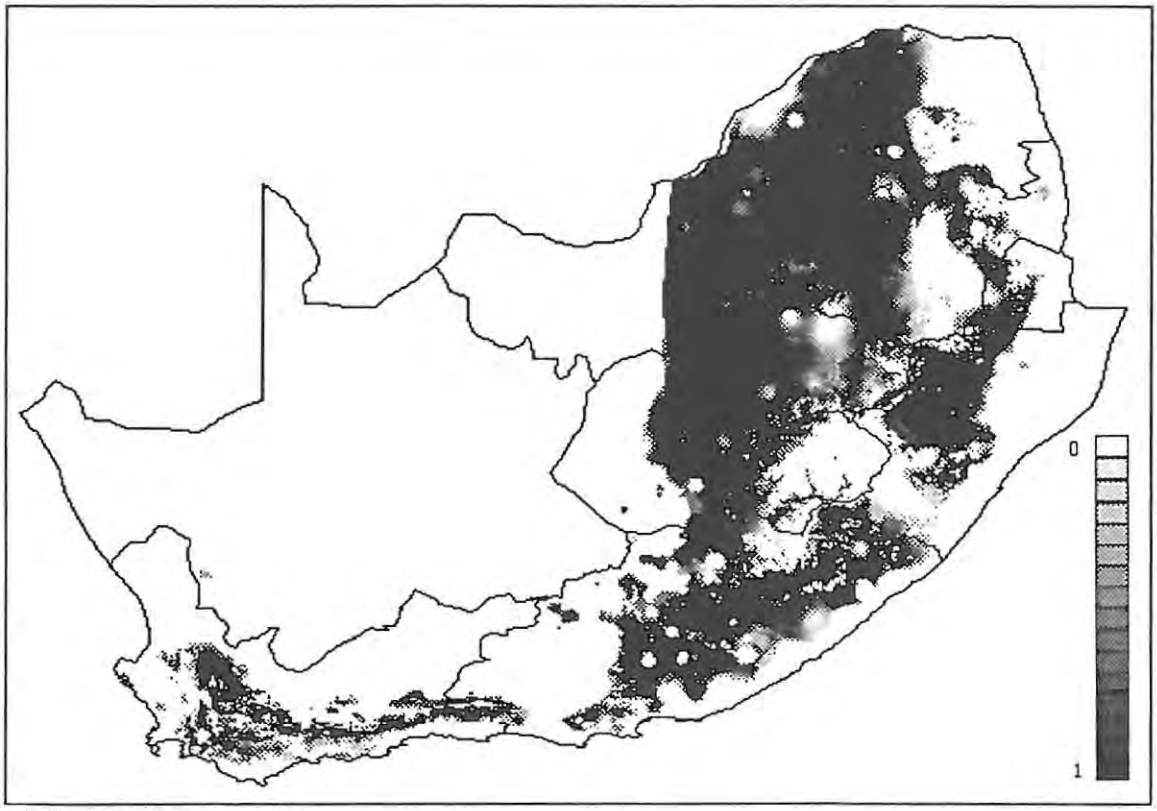
Figure 25. Potential distribution maps produced using fuzzy classification.



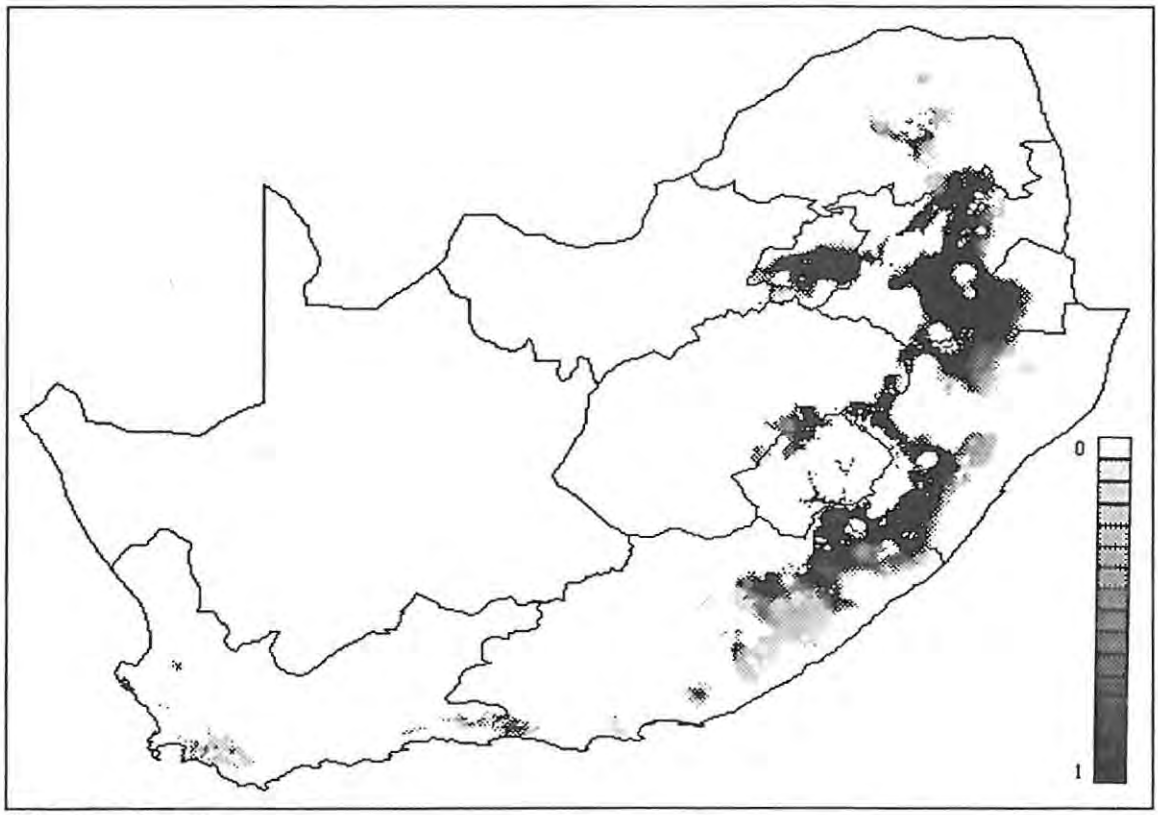
c) *A. mearnsii*



d) *A. mearnsii* (extended data set)



e) *O. ficus-indica*



f) *S. sisymbrifolium*

Table 28. The percentage of sites correctly predicted as present. a) refers to the small data sets and b) to the extended data sets.

		percentage correctly predicted
A. longifolia	a)	10.00
	b)	64.00
A. mearnsii	a)	49.41
	b)	59.92
O. ficus-indica		48.99
S. sisymbrifolium		25.93

Chi-squared tests performed on the contingency tables (table 29) indicate which of the predictive maps show a significant departure from randomness i.e. which of the predictive coverages owe their accuracy to more than chance. Statistically significant maps were those that had a significance level of 0.05 or less.

Table 29. Chi-squared test results and significance levels for each predictive coverage. a) refers to the small data sets and b) to the extended data sets.

		Chi-squared	Sig. level
A. longifolia	a)	17.50	0.0000
	b)	58.00	0.0000
A. mearnsii	a)	12.39	0.0004
	b)	1.93	0.1651
O. ficus-indica		0.30	0.5819
S. sisymbrifolium		0.00	0.9255

8.4) DISCUSSION

One of the greatest advantages of fuzzy logic is that it allows for responses other than linear ones. This is particularly important with ecological data whose distributions are often assumed to approximate a normal distribution (i.e. there are some parts of the range that are more suitable for the plant than others). Ecological data is also seldom 'clean' (Williams, 1983), and therefore the linear and J-shaped functions also offered by IDRISI were not chosen.

Instead the sigmoid function was selected to allow for the environmental ranges of the plant to be defined less sharply and to approximate the assumed non-linear responses of the plants to their environment.

The fuzzy classification as used in this study can perhaps be regarded as a non-linear refinement on the range and interquartile range techniques. The control points a and d, especially for the linear function, are essentially the extremes (range) tolerated by the plant, while the control points b and c could be then regarded as the range between two percentiles, similar to Lindenmayer's *et al* (1991) core distribution.

Unfortunately, the results produced are similar to those produced by the range and interquartile range. Even by counting all the sites where the likelihood of the plant being present was 0.5 or greater as correct predictions of presence, the validation results were disappointing. Only two of the results showed a greater than 50% success rate (table 28), and the chi-squared tests indicate that only one of these (that for the extended data set for *A. longifolia*) is statistically significant (table 29). Two other statistically significant maps were produced, one for the small data set for *A. longifolia* and one for the small data set for *A. mearnsii*, but the validation results indicate that these maps are probably statistically significantly poor predictors of distribution. The chi-squared test results indicate that the maps for *O. ficus-indica* and *S. sisymbrifolium* were not significant departures from randomness.

As with the range and interquartile range techniques, an increase in sample size improved the prediction rates for both of the *Acacia* species, indicating that perhaps this technique requires large enough sample sizes so that the environmental factors are adequately defined. The improved prediction rate with increased sample size may also be noted on the predictive coverages. An extended data set for *A. longifolia* resulted in a marked increase of areas predicted as suitable for invasion by this species (figure 25a and b). This marked difference in distribution between the two sample sizes is not so apparent for the predicted distributions for *A. mearnsii*. While the extended data set for *A. longifolia* predicted an extensive increase of environmentally suitable areas, especially along the coast and in the north-eastern parts of the country; the two distribution maps for *A. mearnsii* show a similar core distribution, with the extended data set serving mainly to increase the range of this core (figure 25c and d). The predictive map for *O. ficus-indica* (figure 25e) indicates a high possibility of the plant

occurring in the northern provinces, sections of the Free State and large tracts of the Eastern Cape. The predicted distribution of *S. sisymbrifolium* (figure 25f) appears very restricted. As this species is a relatively recent invader in South Africa (being introduced in the 1900's; Nel, 1988), it is likely that its range is not yet fully established and therefore the samples are not representative of the full range of environments that the plant could invade.

Also of interest on the predictive coverages are the small circles and other artifacts. These artifacts appear to be related to the importance of the original input coverages on the particular species. For example, from the PCA and ANN techniques, it was established that COV is an important variable for *O. ficus-indica*; most of the circles and the band of predicted presence running through the Free State and North West Province are artifacts from the original COV surface. For the extended data set for *A. longifolia*, it appears that ELV is important in distinguishing between presence and absence, the artifacts can be seen where the areas of predicted presence extend inland, and they indicate the lower-lying rivers and drainage basins. The important surfaces for *A. mearnsii* seem to be a combination of MAXT and ELV. It is difficult to tell which coverages may be leaving artifacts for *S. sisymbrifolium* and for the small data set for *A. longifolia*.

8.5) CONCLUSIONS

Some of the advantages that fuzzy classification can offer are that it allows predictions of presence to be expressed in terms of possibilities of occurrence; it can be non-linear (depending on which function is chosen) and it does not require an absence data set as the absence functions are simply inverse functions of the presence functions. This makes fuzzy classification a useful technique to use if no absence data are available. Thomas and Sun (1995) state that fuzzy techniques are also useful as they can deal with incomplete data sets and crude data.

However, despite these advantages, this technique produced disappointing results, with only one statistically significant map with a greater than 60% success rate being produced, for the extended data set for *A. longifolia*. None of the other predictive maps produced statistically

good results. This poor performance may be related to the small sample sizes, as it was for the range and interquartile ranges (of which this technique could be considered a non-linear refinement). An increase in sample size resulted in an improvement in the number of sites correctly predicted as present for both of the *Acacia* species. Increased sample size would allow the plant's range to be better ascertained, providing that the plants had been in their host country long enough to establish within their range and that the sample sites were representative of the population. It may also be that the response of the plants to the environment is not bell-shaped or that this technique is simply not useful for the purposes of this study.

Artifacts present in the predictive coverages appear to give an indication of which original input variable is important in determining presence of a particular species. This is particularly clear for the coverage for *O. ficus-indica* where there are many artifacts from the COV surface and for the map of the extended data set for *A. longifolia* where the ELV coverage has left artifacts.

SECTION III

CHAPTER 9
DISCUSSION AND CONCLUSIONS

The predictive techniques can be evaluated by several criteria, one of which is the accuracy with which they can predict distribution. Accuracy can be judged quantitatively by the number of sites correctly predicted as present and by the statistical significance of the predictions. At times a technique may show a good validation result, but analysis with chi-squared tests and qualitative analysis (comparing the maps visually) may indicate that many areas of false presence have been predicted i.e. that the technique is predicting the species to occur in areas where it is not present. Qualitative analysis may be accomplished by placing the transparent overlays of the validation sites onto the maps to see whether these validation sites fall into areas of predicted presence. A point to consider when examining the validation results is the degree of accuracy required. A statistically significant predictive success rate of over 80% could be considered excellent and further refinement to the model in an attempt to improve the results may prove unfruitful.

A summary of the validation results from the predictive techniques is available in table 30. All of the results are given as the percentage of sites falling into areas of predicted presence, except for the range and interquartile range techniques which give the percentage of sites falling into areas of maximum suitability.

The chi-squared results for the techniques (table 31) indicate which of the coverages produced are statistically significant ($p < 0.05$) as well as giving an indication of whether the techniques are predicting many areas of false presence (i.e. areas where the plant is actually absent but is predicted to be present). A statistically significant result can either indicate a map that is a significantly good predictor of presence or a significantly poor predictor of presence.

Table 30. Comparison of the validation results of the techniques. a) and b) refer to the small and extended data sets respectively. The numbers are expressed as percentages.

	A. longifolia		*A. mearnsii*		*O. ficus-indica*	*S. sisymbrifolium*
	(a)	(b)	(a)	(b)		
Range	2	54	35	60	35	30
Interquartile range	2	6	5	6	0	4
PC 1	72	78	68	77	69	96
PC 2	48	86	72	93	81	44
PC 3	34	86	72	80	76	64
DFA Method 1	74	95	64	59	76	28
DFA Method 2	57	74	84	74	68	37
ANN	72	88	74	81	80	54
Fuzzy classification	10	64	49	60	49	26

Table 31. Chi-squared results for all of the predictive techniques. The first row for each technique indicates the chi-squared test result and the second row the significance level. a) and b) refer to the small and extended data sets respectively.

	A. longifolia		*A. mearnsii*		*O. ficus-indica*	*S. sisymbrifolium*
	(a)	(b)	(a)	(b)		
Range	48.08	0.32	23.07	10.93	41.80	4.48
	0.0000	0.5716	0.0000	0.0009	0.0000	0.0343
Interquartile range	46.08	38.72	207.63	196.98	404.60	23.15
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PC 1	8.18	2.14	40.44	54.80	6.22	0.01
	0.0427	0.3429	0.0000	0.0000	0.1013	0.9950
PC 2	86.14	78.09	10.87	30.31	10.79	0.07
	0.0000	0.0000	0.0124	0.0000	0.0129	0.9956
PC 3	2.13	4.56	67.43	48.37	2.43	4.05
	0.5456	0.2068	0.0000	0.0000	0.4880	0.2557
DFA Method 1	8.53	30.42	19.43	7.41	107.71	4.84
	0.0035	0.0000	0.0000	0.0065	0.0000	0.0278
DFA Method 2	3.00	6.91	31.91	95.69	251.37	10.67
	0.2231	0.0361	0.0000	0.0000	0.0000	0.0048
ANN	9.68	28.88	58.87	98.37	151.25	0.15
	0.0019	0.0000	0.0000	0.0000	0.0000	0.7003
Fuzzy classification	17.50	58.00	12.39	1.93	0.30	0.00
	0.0000	0.0000	0.0004	0.1651	0.5819	0.9255

The most accurate predictive techniques judging by the validation and chi-squared results are those derived from the DFA and artificial neural networks. These two techniques are also the only ones to use both presence and absence data. The use of absence data may help to cut down on the number of false presences predicted. The least accurate techniques are the range, interquartile range and fuzzy classification, with the chi-squared results indicating that most of the coverages produced are significantly poor predictors of presence. The difference in predictive success between the two DFA methods appears to be negligible. Although the validation results for the PCA indicate that this technique predicts presence well, the chi-squared results and the coverages suggest that this method predicts a lot of false presences.

The three most accurate predictive techniques for the small data set for *A. longifolia* in terms of being statistically significant and having good validation results were: DFA 1 (74% correctly classified), PC 1 (72%) and ANN (72%). For the extended data set the top three techniques were: DFA 1 (95%), ANN (88%) and PC 2 (86%). For the small data set for *A. mearnsii*, the most accurate predictive techniques were: DFA 2 (84%), ANN (74%) and PC 2 (72%) and for the extended data set: ANN (81%), DFA 2 (74%) and Range (60%). ANN (80%), DFA 1 (76%) and DFA 2 (68%) were the three techniques that produced the best predictive results for *O. ficus-indica*. The three most accurate techniques for *S. sisymbrifolium* all showed significantly poor predictions and none of the validation results were over 40%. The techniques that produced the greatest percentage of sites correctly predicted as present were DFA 2 (37%), followed by the Range (30%) and DFA 1 (28%).

With regards to the effect of sample sizes on the predictive qualities of the techniques, an increase in sample size improved the validation results for all of the techniques except for *A. mearnsii* for the two DFA methods, where an increased sample size resulted in slightly less successful predictions of presence. This suggests that the models could perhaps be improved with larger data sets, except for the DFA. The decrease in model performance for *A. mearnsii* for the DFA may be due to the slightly less accurate classification by the DFA for this species with larger sample sizes.

Accuracy is not the only criteria that can be used to evaluate the success of the predictive techniques. Simplicity, ease of use, any special advantages offered and the potential of the technique to be taken further were also judged to be important criteria to the success of the techniques.

With regards to simplicity and ease of use, the range and interquartile range are the simplest and easiest to understand of the five techniques. While this simplicity allows them to be easily incorporated into other techniques, such as the range was in the PCA, it tends to limit further development of the technique as such. While the DFA and PCA techniques are more complex to understand and implement than the range, their use is greatly facilitated by the number of software programmes readily available for their calculation. The ANN is a complex technique, especially to implement. It also requires a degree of experience to operate and can be time-consuming depending on the speed of the computer the programmes are run on and the skill of the operator. Specialized software is required to train the network whereas most statistical programmes will perform DFA and PCA operations. If regarded as a non-linear refinement on the range and interquartile range, fuzzy classification is fairly simple to understand. The fuzzy classification module in IDRISI is easy to run, but requires expert knowledge to define the plants' preferences for each environmental variable. This may present problems if the shape of the response curve of the plants to the environment is not known, as it will affect the function chosen for the fuzzy classification.

Some of the techniques offer advantages additional to predicting potential distribution. PCA, for example, is particularly useful for picking out which variables or combinations thereof are important for determining the distribution of each species. By examining the amount of variance explained by the first principal component, one can also determine whether an appropriate number of variables have been chosen to model with. If the percentage of variance explained by the first component is high, then there may be a lot of redundancy in the data, if the percentage variance explained by the first component is very low, then too few variables may have been chosen to model with (Manly, 1986). The percentage variance explained by the first principal component for this study suggests that a sufficient number of variables were chosen to model with. The ANN and DFA also give an indication (through means of the weights given to each of the variables) of which of the predictor variables may be important in determining the distribution of the four species. Fuzzy classification could prove to be a

useful tool in studies that need to integrate local knowledge, as local people often have the expert knowledge necessary to define the fuzzy sets for the variables being studied in their particular area (Thomas & Sun, 1995).

The non-linearity of the ANN and Fuzzy classification techniques did not appear to offer a significant advantage over the linear techniques used. Although the ANN performed well, it did not appear to give significantly better results over the DFA, which is a linear technique. It may be that the responses of the plants to the environmental variables were not bell-shaped as they were assumed to be for the non-linear techniques. This would help to explain the poor results for the fuzzy classification, as the classification function chosen for this technique assumed that the plants' response to their environment was normally distributed.

Some of the techniques resulted in artifacts on the predictive coverages. These artifacts appeared to be related to the input coverage that was the best determinant of the distribution of a particular species. The techniques that resulted in the coverages with the most artifacts were fuzzy classification, artificial neural networks and Method 2 for the DFA. This suggests that these techniques are useful in determining the variables best used to predict the distribution of any particular species.

With regards to the potential of the techniques to be taken further, PCA offers numerous avenues for further study. It would be interesting to add in more variables, particularly the co-ordinates of the presence/absence sites of the plants to link the plant data more closely with the PCA. It may also be worthwhile to extract more principal components to determine what other variables or combinations thereof may affect the distribution of the plants. The range and interquartile range techniques do not appear to offer much scope for further development, except perhaps as part of other techniques. It may be useful to use other percentiles, such as the 10th and the 90th instead of the interquartile range as Lindenmayer *et al* (1991) did. The DFA technique already predicts very well but might be improved by the addition of more variables. The two methods might also be refined to yield more classes to give an indication of the likelihood of a plant being present or absent. The ANN also predicts well, but could perhaps be further improved (or worsened) by experimenting with changes in the learning parameters. It may be of benefit to train a network using unsupervised instead of supervised classification. Unsupervised classification does not require prior knowledge of whether the

plant is present or absent, but instead trains by grouping similar vectors together (Fausett, 1994). The addition of extra variables to this technique, though while possible, will result in an increase in training time for the network and the complexity of producing the predictive maps. Fuzzy classification may be well suited for use as a risk-analysis tool as the output is in the form of range of possibilities and not just a binary output.

Discriminant function analysis as used here, linked to a GIS, appears to be the best overall predictive technique in terms of accuracy of predictions and ease of use. The ANN also produces excellent statistically significant results, but is more complex and time-consuming to use than the DFA. However, both of these techniques require absence and presence data sets, whereas the remaining techniques only require presence data. The PCA perhaps offers the most scope for further development and is worth refining for use as a predictive technique. The two non-linear techniques, the ANN and fuzzy classification, did not appear to offer significantly better results than the linear techniques. In fact, fuzzy classification produced fewer good predictive maps than the PCA or DFA. However, fuzzy classification is worth further study for use as a risk-assessment tool.

The ultimate aim of the techniques was to produce predictive models that are accurate and easy to use. To be of value then, the maps produced by the predictive techniques need to be used as indicators of the areas climatically suitable for invasion by the four alien plant species, and not just as indicators of how accurate the techniques were.

With regards to the distribution of the invasive species, *A. longifolia* shows a predicted preference for the coastal areas, both east and west, seldom extending inland except along some of the rivers. It is also predicted to occur in the northern parts of the country, along the borders with Mozambique and Zimbabwe. This may necessitate a co-operative effort between the three countries to effectively control this species. At present, control of *A. longifolia* is in the form of a biological control agent, the gall-forming wasp *Trichilogaster acaciaelongifoliae*. This agent appears to be successful in large parts of the range of this plant (Dennill, 1987), although there have been some suggestions that it is not very effective in controlling *A. longifolia* in hotter inland valleys (Dennill & Gordon, 1990). It is in these areas where the biological control agent may not be effective that other control measures should be implemented.

The most important variable influencing the distribution of this species according to the PCA and ANN techniques was minimum temperature. The DFA and ANN also indicated that elevation was important, which is likely as elevation and minimum temperature are inversely related.

Minimum temperature is also an important determinant of the distribution of *A. mearnsii* according to the PCA, DFA and ANN techniques, along with maximum temperature (DFA and ANN) and to a lesser extent, elevation (ANN). Both of the *Acacia* species are natives of the temperate areas of South-East Australia (Stirton, 1987) which is why their distribution in South Africa may be affected by minimum and maximum temperatures. They do not appear to thrive far inland where temperatures are not moderated by proximity to the sea. *Acacia mearnsii* may be slightly less suspect to extreme temperatures than *A. longifolia* as it shows a similar, but more extensive distribution to the latter, especially inland in KwaZulu-Natal and Mpumalanga, where it is already extensively commercially cultivated.

Most of the country, except for some coastal areas such as the west coast, appears suitable for the establishment of *O. ficus-indica*. The most favourable areas appear to be the northern and eastern parts, the Cape fold mountains and the Eastern Cape.

The co-efficient of variation for rainfall was indicated by the ANN as the best predictor variable for this species. The DFA indicated a combination of MAXT, COV and ELV as being important. However, the most important determinant of the distribution of this species is likely to be its biological control agents, which have effectively reduced much of its previous range.

The areas predicted as suitable for *S. sisymbrifolium* include parts of the Northern Province, Gauteng, Mpumalanga, continuing as a mostly continuous band along the KwaZulu-Natal coastal plateau to Cape Town. However, in the light of the poor validation results, the potential distribution of this species should not be considered accurate. The range of *S. sisymbrifolium* may also be effectively curtailed by its recently released biological control agent, *Gratiana spadicea*.

According to the ANN technique, the best predictor variable for this species is COV; however, in light of the poor prediction results, small sample size and contradiction from the DFA, it is unlikely that this is the best predictor variable, and that some other variable not used in this study is a better determinant of the distribution of this species.

All of the predictive models for *S. sisymbrifolium* tended to produce disappointing results. This may be due to the small sample size, but the good model results for *O. ficus-indica* which also had a small sample size, suggest that the problem may rather lie in the length of time which the invader species has had to establish itself in its host country. *Opuntia ficus-indica* was introduced to South Africa over 250 years ago (Moran & Zimmerman, 1991) and has thus had sufficient time to migrate to the limits of its niche. *Solanum sisymbrifolium* on the other hand, is a relatively recent invader to South Africa, having been introduced in the 1900's (Nel, 1988) and may not have had time enough to establish within its bioclimatic range in the country.

With regards to potential use of the predictive maps, there is much scope for further development. One could use the most successful predictive techniques to produce a predictive atlas for invasive alien plants in South Africa; or one could model the distribution of other organisms entirely. While there is scope to refine some of the techniques, it would be interesting to see if the DFA and ANN as used here produce good results for other species or if the success is limited to the plant species modelled here. It would also be interesting to map the distributions of the biological control agents for some of the alien species modelled here to determine if there are any areas in South Africa where the alien plants are predicted to survive, but the biological control agents are not i.e. to pinpoint areas where biological control may not be successful. The linkage of the predictive techniques to a GIS, as developed in this study, greatly facilitates further work of this nature. The GIS database is easily and rapidly updated and offers superior mapping and data analysis capabilities to the predictive techniques used on their own. It is to be hoped that the modelling techniques as developed in this study, particularly the DFA, ANN and PCA, are to be taken further as they show great potential as predictive modelling techniques.

References

Abrams, R.W. (1985) Environmental determinants of pelagic seabird distribution in the African sector of the Southern Ocean. *Journal of Biogeography*, 12: 473-492.

Accone, T. (1992) *The mapping of potential pineapple producing areas and comparison of these areas with land presently under cultivation, in the Eastern Cape*. Unpublished third year Geography project; Rhodes University: Grahamstown.

Anon (1996) *A neural network approach for interpolating species density patterns from remotely sensed & GIS data: an example using the desert tortoise*. Internet site, http://ice.gis.uiuc.edu/Neural/tort.html#tort_scat

Aronoff, S. (1991) *GIS: a management perspective*. WDL Publications: Ottawa.

Austin, M.P., Nicholls, A.O. & Margules, C.R. (1990) Measurement of the realised qualitative niche: environmental niches of five *Eucalyptus* species. *Ecological Monographs*, 60: 161-177.

Bauer, I.E, McMorrow, J. & Yalden, D.W. (1994) The historic ranges of three equid species in north-east Africa: a quantitative comparison of environmental tolerances. *Journal of Biogeography*, 21: 169-182.

Boynton, R.J. (1989) *A spatial analysis of the distribution of the genus Psammophis (grass and sand snakes) in southern Africa*. Unpublished Honours thesis: Rhodes University, Grahamstown.

Box, E.O., Crumpacker, D.W. & Hardin, E.D. (1993) A climatic model for location of plant species in Florida, U.S.A. *Journal of Biogeography*, 20: 629-644.

Brunsdon, C. & Openshaw, S. (1993) Simulating the effects of error in GIS. In Mather, P.M (Ed) *Geographical information handling - research and applications*, pp 47-61. John Wiley & Sons: Chichester.

Buckland, S.T. & Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, 30: 478-495.

Cactus Pear Growers Association (1996) *Information brochure on the production of spineless prickly pears*. Brochure.

Caughley, G., Short, J., Grigg, G.C. & Nix, H. (1987) Kangaroos and climate: an analysis of distribution. *Journal of Animal Ecology*, 56: 751-761.

Chapman, A.D & Busby, J.R. (1994) Linking plant species information to continental biodiversity inventory, climate modelling and environmental monitoring. In Miller, R.I. (Ed) *Mapping the diversity of nature*, pp 179-195. Chapman and Hall: London.

Chrisman, N.R. (1991) The error component in spatial data. In Maguire, D.J., Goodchild, M.F. & Rhind, D.W. (Eds) *Geographical information systems*. Volume 1, Principles, pp 165-174. Longman: Harlow.

Cole, G. (1995) Fuzzy logic, fabulous performance. *Accountancy*, 116: 54-56.

Cowen, D.J. (1988) GIS versus CAD versus DBMS: what are the differences? *Photogrammetric Engineering and Remote Sensing*, 54: 1551-1555

Danon, Y. (1995) WinNN - Windows neural networks. Shareware.

De Beer, H. (1986) Black wattle. *Farming in South Africa. Weeds A.24/1986*. Department of Agriculture and Water Supply: Pretoria.

de Laubenfels, D.J. (1975) *Mapping the world's vegetation: regionalization of formations and flora*. Syracuse University Press: New York.

Demuth, H. & Beale, M. (1994) *Neural network toolbox. For use with MATLAB. User's guide*. The Math Works Inc.: Massachusetts.

Dennill, G.B. (1987) *The biological control of the weed Acacia longifolia by the gall wasp Trichilogaster acaiaelongifoliae, a study of a plant-insect interaction*. Unpublished PhD thesis, University of Cape Town, Cape Town.

Dennill, G.B. & Gordon, A.J. (1990) Climate-related differences in the efficacy of the Australian gall wasp (Hymenoptera: Pteromalidae) released for the control of *Acacia longifolia* in South Africa. *Environmental Entomology*, 19: 130-136.

Dent, M.C., Lynch, S.D. & Schulze, R.E. (1989) *Mapping mean annual and other rainfall statistics in southern Africa*. Department of Agricultural Engineering, University of Natal. ACRU. Report No. 27.

Department of Water Affairs and Forestry (1996a) *How much water do alien plant invaders use*? Working for water programme. Phamphlet: Government Printer.

Department of Water Affairs and Forestry (1996b) *Working for water*. The RDP water conservation programme. Phamphlet: Government Printer.

de Selincourt, K. (1992) South Africa's other bush war. *New Scientist*, 133 Feb 15: 46-49.

Eastman, J.R. (1994) *IDRISI technical reference*. Clark University: Worcester, Massachusetts.

Eastman, J.R. (1995) *IDRISI for Windows technical reference*. Clark University: Worcester, Massachusetts.

Elston, D.A. & Buckland, S.T. (1993) Statistical modelling of regional GIS data: an overview. *Ecological Modelling*, 67: 81-102.

Fabricius, C. & Coetzee, K. (1992) Geographic information systems and artificial intelligence for predicting the presence or absence of mountain reedbuck. *Scientific Journal of Wildlife Research*, 22: 80-86.

Fausett, L. (1994) *Fundamentals of neural networks. Architectures, algorithms, and applications*. Prentice Hall: Engelwood Cliffs, New Jersey.

Fresenmaier, D.R., Goodchild, M.F. & Morrison, S. (1979) The spatial structure of the rural-urban fringe: a multivariate approach. *The Canadian Geographer*, 23: 255-265.

Gibson, D. (1995) *Modelling the distribution of Portulacaria afra in the Eastern and Western Cape Provinces, South Africa, in relation to environmental variables and the normalised difference vegetation index*. Unpublished Honours Thesis: Rhodes University, Grahamstown.

Goodchild, M.F. (1991) Geographic information systems. *Progress in Human Geography*, 15: 194-200.

Gregory, S. (1978) *Statistical methods and the geographer*. Longman: London.

Haykin, S. (1994) *Neural networks. A comprehensive foundation*. MacMillan: New York.

Hecht-Nielson, R. (1988) Neurocomputing: picking the human brain. In Vemuri, V. (Ed) *Artificial neural networks: theoretical concepts*, pp 13-18. Computer Society Press of the IEEE: Washington.

Henderson, L. (1995) *Plant invaders of southern Africa*. Plant Protection Research Institute handbook No. 5, Agricultural Research Centre: Pretoria.

Henderson, M., Fourie, D.M.C., Wells, M.J. & Henderson, L. (1987) *Declared weeds and alien invader plants in South Africa*. Department of Agriculture and Water Supply: Pretoria. Bulletin 413.

Higgins, S.I & Richardson, D.M (1995) A review of models of alien plant spread. *Ecological modelling*, in press.

Hill, M.P. (1994) *Evaluation of Gratiana spadicea (Klug, 1829) and Metriona elatior (Klug, 1829) (Chrysomelidae: Cassidinae) for the biological control of sticky nightshade Solanum sisymbriifolium Lamarck (Solanaceae) in South Africa*. Unpublished PhD Thesis, Rhodes University: Grahamstown.

Huberty, C.J. (1992) *Applied discriminant analysis*. Wiley: New York.

Isaaks, E.H. & Srivastava, R.M. (1989) *An introduction to applied geostatistics*. Oxford University Press: Oxford.

Jackson, B.B. (1983) *Multivariate data analysis. An introduction*. Richard D Irwin: Homewood, Illinois.

Jeffers, J.N.R. (1967) Two case studies in the application of principal component analysis. *Applied Statistics*, 16: 225-236.

Jeffers, J.N.R (1982) *Modelling*. Outline studies in Ecology; Chapman and Hall: London.

Kinoshita, J (1988) Neural networks at work: they watch over factories, credit applicants, sleepy pilots. *Scientific American*, 259: 96-98, May.

Knight, R.S (1986) Interrelationships between fruit types in southern African trees and environmental variables. *Journal of Biogeography*, 13: 99-108.

Knoke, J.D. (1982) Discriminant analysis with discrete and continuous variables. *Biometrics*, 38: 191-200.

Korte, G.B. (1994) *The GIS book*. OnWord Press: Santa Fe.

Kosko, B. & Isaka, S. (1993) Fuzzy logic. *Scientific American*, 269: 62-67, January.

Koutnik, M.A & Padilla, D.K. (1994) Predicting the spatial distribution of *Dreissena polymorpha* (zebra mussel) among inland lakes of Wisconsin: modelling with a GIS. *Canadian Journal of Fisheries and Aquatic Sciences*, 51: 1189-1196.

le Maitre, D.C., Van Wilgen, B.W., Chapman, R.A. & McKelly, D.H (1996) Invasive plants and water resources in the Western Cape Province, South Africa: modelling the consequences of a lack of management. *Journal of Applied Ecology*, 33: 161-172.

Liebhold, A.M, Halverson, J.A & Elmes, G.A. (1992) Gypsy moth invasions in North America: a quantitative analysis. *Journal of Biogeography*, 19: 513-520.

Liebhold, A.M., Rossi, R.E. & Kemp, W.P. (1993) Geostatistics and geographical information systems in applied insect ecology. *Annual Review of Entomology*, 38: 303-327.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P, Hutchinson, M.F. & Tanton, M.T. (1991) The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*, 18: 371-383.

Liu, K. & Lam, N. (1985) Paleovegetational reconstruction based on modern and fossil pollen data: an application of discriminant analysis. *Annals of the Association of American Geographers*, 75(1): 115-130.

Macdonald, I.A.W. (1984) Is the fynbos biome especially susceptible to invasion by alien plants? A re-analysis of available data. *South African Journal of Science*, 80: 369-377.

Macdonald, I.A.W. & Jarman, M.L (Eds) (1985) *Invasive alien plants in the terrestrial ecosystems of Natal, South Africa*. South African National Scientific Programmes Report no. 118. CSIR: Pretoria.

Manly, B.F.J. (1986) *Multivariate statistical methods. A primer*. Chapman and Hall: London.

Manugistics (1992) Statgraphics examples manual. Font Software, Bitstream Inc., Cambridge.

McAllister, D.E., Schueler, F.W., Roberts, C.M. & Hawkins, J.P. (1994) Mapping and GIS analysis of the global distribution of coral reef fishes on an equal-area grid. In Miller, R.I. (Ed) *Mapping the diversity of nature*, pp 155-175. Chapman and Hall: London.

Meadows, M.E. (1985) *Biogeography and ecosystems of South Africa.* Juta & Co.: Cape Town.

Menozzi, P., Piazza, A. & Cavalli-Sforza, L. (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, 201: 786-792.

Michelmore, F. (1994) Keeping elephants on the map: case studies of the application of GIS for conservation. In Miller, R.I. (Ed) *Mapping the diversity of nature*, pp 107-125. Chapman and Hall: London.

Miller, R.I. (1994) Possibilities for the future. In Miller, R.I (Ed) *Mapping the diversity of nature*, pp 199-205. Chapman and Hall: London.

Moran, V.C., Neser, S. & Hoffmann, J.H. (1986) The potential of insect herbivores for the biological control of invasive plants in South Africa. In MacDonald, I.A.W., Kruger, F.J. and Ferrar, A.A. (Eds) *The ecology and management of biological invasions in southern Africa*, pp 261-268. Oxford University Press: Cape Town.

Moran, V.C. & Zimmerman, H.G. (1984) *The biological control of cactus weeds: achievements and prospects.* Biocontrol News and information, review article. Commonwealth Institute of Biological Control, 297-320.

Moss, D., Furse, M.T., Wright, J.F. & Armitage, P.D. (1987) The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. *Freshwater Biology*, 17: 41-52.

Mowforth, M. (1979) *Statistics for geographers.* Harrap: London.

Muir, K. (1986) *Vegetable dyeing in South Africa.* Weaverbird Publications: Cape Town.

Munday, J (1988) *Poisonous plants in South African gardens and parks. A field guide.* Delta books: Craighall.

Nel, C. (1988) A tomato that gobbles up grazing. *Farmer's Weekly*, 10 June, 18-19.

O'Conaill, M.A., Mason, D.C. & Bell, S.B.M. (1994) Spatiotemporal GIS techniques for environmental modelling. In Mather, P.M., *Geographical information handling - research and applications*, John Wiley & Sons: Chichester; pp 103-112.

Osborne, P.E. & Tigar, B.J. (1992) Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. *Journal of Applied Ecology*, 29: 55-62.

Palmer, A.R. (1991) The potential vegetation of the upper Orange River, South Africa: concentration analysis and its application to rangeland assessment. *COENOSES* 6: 131-138.

Palmer, A.R. & Van Staden, J.M. (1992) Predicting the distribution of plant communities using annual rainfall and elevation: an example from southern Africa. *Journal of Vegetation Science*, 3: 261-266.

Panetta, F.D. & Dodd, J. (1987) Bioclimatic prediction of the potential distribution of skeleton weed *Chondrilla juncea* L. in western Australia. *The Journal of the Australian Institute of Agricultural Science*, 53: 11-16.

Putman, R.J. & Wratten, S.D. (1985) *Principles of ecology*. Croom Helm: London.

Ramcharan, C.W., Padilla, D.K. & Dodson, S.I. (1992) Models to predict potential occurrence and density of the zebra mussel, *Dreissena polymorpha. Canadian Journal of Fisheries and Aquatic Sciences*, 49: 2611-2620.

Rebelo, A.G. (1987) Management implications. In Rebelo, A.G. (Ed) *A preliminary synthesis of pollination biology in the Cape flora*, pp 193-211. South African National Scientific Programmes Report no. 141. CSIR: Pretoria.

Richardson, D.M, MacDonald, I.A.W., Holmes, P.M. & Cowling, R.M (1992) Plant and animal invasions. In Cowling, R.M (ed) *The ecology of fynbos, nutrients, fire and diversity*, pp 271-308. Oxford University Press: Oxford.

Richardson, D.M. & McMahon, J.P. (1992) A bioclimatic analysis of *Eucalyptus nitens* to identify potential planting regions in southern Africa. *South African Journal of Science*, 88: 380-387.

Rogers, D.J. & Williams, B.G. (1993) Monitoring trypanosomiasis in space and time. *Parasitology*, 106: S77-S92.

Rowntree, K.M. (1991) An assessment of the potential impact of alien invasive vegetation on the geomorphology of river channels in South Africa. *South African Journal of Aquatic Sciences*, 17: 28-43.

Schoener, T.W. (1988) The ecological niche. In Cherrett, J.M (Ed) *Ecological concepts, the contribution of Ecology to an understanding of the natural world*, pp 79-113. Blackwell Scientific Publications: Oxford.

Scott, J.K & Panetta, F.D. (1993) Predicting the Australian weed status of some southern African plants. *Journal of Biogeography*, 20: 87-93.

Sejnowski, T.J. & Rosenberg, C.R. (1988) NETtalk: a parallel network that learns to read aloud. In Anderson, J.A. & Rosenfeld, E. (Eds) *Neurocomputing: foundations of research*, pp 661-672. The MIT Press: Cambridge, Massachusetts.

Sharov, A. (1996) *Modelling Forest Insect Dynamics*. Internet site http://www/gypsymoth.ento.vt.edu/~sharov/popechome/model/model.html

Smith, R.E. & Bosch, J.M. (1989) A description of the Westfalia experiment to determine the influence of conversion of indigenous forest on water yield. *South African Forestry Journal*, 151: 26-31.

Stirton, C.H.(ed) (1987) *Plant invaders; beautiful but dangerous*. Department of Nature and Environmental Conservation of the Cape Provincial Administration, ABC Press: Cape Town.

Thomas, D.A. & Sun, X. (1995) Rangeland production: use of models incorporating aggregated knowledge and fuzzy construction. *Journal of Arid Environments*, 30: 479-494.

Tivy, J. (1993) *Biogeography. A study of plants in the ecosphere*. Third edition. Longman Scientific & Technical: Harlow.

Vahrmeijer, J. (1981) *Poisonous plants of southern Africa that cause stock losses*. Tafelberg: Cape Town.

van Lil, W.S., Kruger, F.J. & van Wyk, D.B. (1980) The effect of afforestation with *Eucalyptus grandis* Hill ex Maiden and *Pinus patula* Schlecht. Et. Cham. on streamflow from experimental catchments at Mokobulaan, Transvaal. *Journal of Hydrology*, 48: 107-118.

van Wilgen, B.W., Cowling, R.M. & Burgers, C.J. (1996) Valuation of ecosystem services. A case study from South African fynbos ecosystems. *BioScience*, 46: 184-189.

Vermeulen, W.J. (1989) *Guide on the control of Australian acacias*. Department of Environment Affairs, pamphlet 365/8.

von Gadow, K. & van Hensbergen, H.J. (1987) Contributions to forest modelling research. *South African Forestry Journal*, 140: 44-50.

Wadge, G., Wislocki, A., Pearson, E.J & Whittow, J.B. (1993) Mapping natural hazards with spatial modelling systems. In Mather, P.M (Ed) *Geographical information handling - research and applications*, pp 239-250. John Wiley & Sons: Chichester.

Walker, P.A. (1990) Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. *Journal of Biogeography*, 17: 279-289.

Wasserman, P.D. (1989) *Neural computing. Theory and practice*. Van Nostrand Reinhold: New York.

Wells, M.J., Balsinhas, A.A., Joffe, H., Engelbrecht, V.M., Harding, G. & Stirton, C.H. (1986) *A catalogue of problem plants in southern Africa incorporating the National Weed List of South Africa*. Memoirs of the Botanical Survey of South Africa. Botanical Research Institute: Pretoria.

Williams, B.K. (1983) Some observations on the use of discriminant analysis in ecology. *Ecology*, 64: 1283-1291.

Wong, S.T. (1968) A multivariate statistical model for predicting mean annual flood in New England. In Berry, J.L. & Marble, D.F., *Spatial analysis, a reader in statistical geography*, pp 353-367. Prentice-Hall: New Jersey.

Young, J.A.T. (1986) *A U.K. geographic information system for environmental monitoring, resource planning and management capable of integrating and using satellite remotely sensed data.* A Remote Sensing Society Monograph; Remote sensing society, Geography Department, University of Nottingham: Nottingham.

Zadeh, L.A. (1965) Fuzzy sets. In Yager, R.R.; Ovchinnikov, S., Tong, R.M. & Nguyen, H.T. (Eds) (1987) *Fuzzy sets and applications: selected papers by L.A. Zadeh*, pp 29-44. Wiley-interscience: New York.

Zimmermann, H.G. & Moran, V.C. (1991) Biological control of prickly pear, *Opuntia ficus-indca* (Cactaceae), in South Africa. In Hoffmann, J.R. (Ed) *Agriculture, Ecosystems and Environment*, 37: 29-35.

APPENDIX A

Appendix A contains the list of steps followed in the geographical information system for implementing the artificial neural networks and explanatory diagram (figure A1).

Words in square brackets indicate the IDRISI modules used.

For the first hidden node:
1) Standardize inputs (MAR, COV, MAXT, MINT, ELV) [STANDARDIZE]
2) Multiply each input image by its weight [SCALAR, multiply]
3) Sum the weighted images [OVERLAY, add]
4) Multiply the bias by its weight [SCALAR, multiply]
5) Add the weighted bias to the summed images [SCALAR, add]
6) Apply the activation function [OVERLAY, exponentiate; SCALAR, add 1; OVERLAY, ratio]

A sigmoidal activation function was used and took the form of:

$$f = 1/(1 + \exp^{-n})$$

where n is the sum of the weighted inputs for each node

7) Take the output from the activation function and multiply it by its weight [SCALAR, multiply]
8) For hidden node 2 repeat steps 1 to 7
9) Sum the results from step 7 for each hidden node [OVERLAY, add]
10) Multiply the bias by its weight [SCALAR, multiply]
11) Add the weighted bias to the output from step 9 [SCALAR, add]
12) Apply the activation function [OVERLAY, exponentiate; SCALAR, add 1; OVERLAY, ratio]

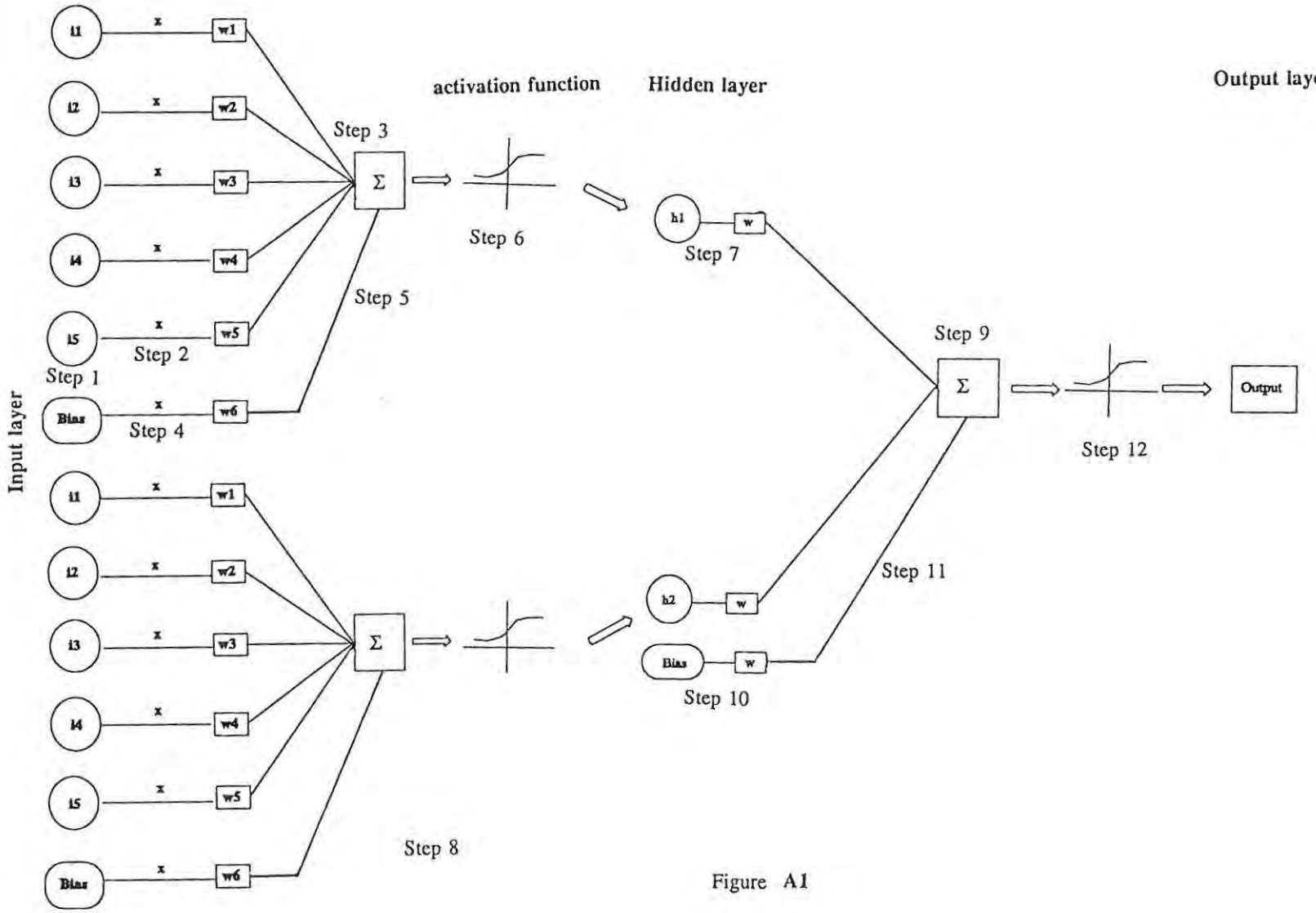

Input layer
activation function
Hidden layer
Output layer
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
Step 9
Step 10
Step 11
Step 12
i1
i2
i3
i4
i5
Bias
w1
w2
w3
w4
w5
w6
x
Σ
h1
h2
w
Output

Figure A1

APPENDIX B

Appendix B is a guide to the use of the validation overlays which are available in the pocket at the back of this thesis. The validation data sets have been reproduced onto transparent overlays. These can be overlayed onto the predictive maps to give a qualtitative assesment of how well the maps are predicting presence. While there is only one validation data set for each species, the data sets for the two *Acacia* species have been duplicated and placed on one sheet so that they conform to the layout of the predictive maps (i.e. the validation data set for each *Acacia* species is the same for both the small and the extended data sets).