UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Integrating Geographical Information Systems and Artificial Neural Networks to improve spatial decision making

By

Sanet Eksteen

Submitted in fulfilment of the requirements for the degree of

Masters of Science (Geoinformatics)

Faculty of Natural and Agricultural Science
University of Pretoria
May 2010

Declaration:
I declare that the thesis/dissertation that I hereby submit for the MSC degree in Geoinformatics at the University of Pretoria has not previously been submitted by me for degree purposes at any other university.

Signature…… Date: …….

The use of Geographical Information Systems and Artificial Neural Networks in Spatial Decision Making

Student: Sanet P Eksteen

Supervisor: Dr Gregory Breetzke

Department: Geography, Geoinformatics and Meteorology

Summary

GIS has been used in Veterinary Science for a couple of year and the application thereof has been growing rapidly. A number of GIS models have been developed to predict the occurrences of certain types of insect species including the *Culicoides* species (spp), the insect vectors responsible for the transmission of the African horse sickness (AHS) virus. AHS is endemic to sub-Saharan Africa and is carried by two midges called *Culicoides Imicola* and *Culicoides Bolitinos*. The disease causes severe illness in horses and has significant economic impact if not dealt with timeously. Although these models had some success in the prediction of possible abundance of the *Culicoides* spp. the complicated nature and high number of variables influencing the abundance of *Culicoides* spp. posed some challenges to these GIS models. This informs the need for models that can accurately predict potential abundance of *Culicoides* spp to prevent unnecessary horse deaths.

This lead the study to the use of a combination of a GIS and an artificial neural networks (ANN) to develop a model that can predict the abundance of *C. Imicola* and *C. Bolitinos*. ANNs are models designed to imitate the human brain and have the ability to learn through examples. ANNs can therefore model extremely complex

features. In addition, using GIS maps to visualise the predictions will make the models more accessible to a wider range of practitioners.

ACKNOWLEDGEMENTS:
Onderstepoort Veterinary Institute, Agricultural Research Council for supplying the *Culicoides* spp counts used in this study
Directorate: Veterinary Services, Department of Agriculture, Forestry and Fisheries for continued support and evaluation.

TABLE OF CONTENT

LIST OF FIGURES

List of Tables:

Annexure:

Chapter 1: Introduction

1. Background

The application of Geographical Information Systems (GIS) in veterinary science dates back to the late 1960s when a Canadian scientist applied GIS to better understand the spread of foot-and-mouth disease in England (Ramirez A, 2004). Since then, the application of GIS in veterinary science has grown rapidly and currently includes models for disease monitoring (Rogers *et al*., 1993), biological risk management (Boone *et al*. 2007), scenario planning (Genchi *et al*., 2005) and animal health surveillance (Ramirez, 2004). A number of GIS models have also been developed to predict occurrences of certain types of insect species including the *Culicoides* species (spp) responsible for the transmission of the African horse sickness (AHS) virus. Baylis *et al.* (1999) for example used GIS to understand the causes of the geographical variation in the quantity of *Culicoides* spp in South Africa whilst Wittmann *et al.* (2001) used climate data to map the potential distribution of *C.Imicola* in Europe. The former study is the only current model developed in South Africa to predict the abundance of *Culicoides* spp. Although these models have had some success in predicting potential Culicoides occurrences, the exact relationships among the different variables causing the occurrence of these species could not be determined. The complicated nature of the study and the high number of variables that influence the abundance of the *Culicoides* spp pose challenges to the development of GIS models. This has led to the development and incorporation of artificial neural networks (ANN) within GIS as an additive tool to improve spatial decision making. ANNs are models designed to imitate the human brain and have the ability to derive meaning from complicated and imprecise data (Thurston, 2002; Stergiou, 1995). Combining GIS

and ANN for decision making has been used by a number of researchers for example in mining applications (Pradan *et al*.,2008), to forecast possible changes in land use (Vafeidis *et al.,* 2007) and to model deforestation (Mas *et al*., 2003). Recent advances in computer technology and associated applications are allowing decision makers to deal with increasing levels of complexity in decision making. While these complex challenges deals with many dimensions and uncertainties it limits the effectiveness of exact methods of analysis. A number of applications displaying these complex characteristics are found in areas such as water quality management (Jiang *et al*, 2008), ecology (Bessa-Gomes *et al,* 2003, Lusk *et al,* 2002) and veterinary science (Cavero *et al*, 2008, Ward *et al*, 2006). Given the importance of agriculture in the South African economy, an application displaying these characteristics from within the veterinary science was selected as a case study. One such problem that is very complex with a strong spatial component is the prediction of abundance of *Culicoides* spp. that causes AHS.

2. Research Aim

This study aims to use GIS and ANN to predict the abundance of two *Culicoides* spp. responsible for transmitting the AHS virus in South Africa. In doing so the study outlines the increasing role and integration of artificial intelligence within mainstream GIS applications.

The secondary aims of the study are:

- to evaluate the process of integrating a GIS and ANN; and

- to demonstrate the integration of GIS and ANN by means of a case study. In this case to predict the abundance of *C. imicola and C. bolitinos* in the Western Cape Province of South Africa.

3. The Case Study

In order to demonstrate the integration of a GIS and ANN an application with the following characteristics was selected:

- A strong spatial component to illustrate the value of a GIS in decision making and allow broader use of the results through visualisation, making it more accessible,
- integration of multiple data sources; and
- significant complexity without exact solutions indicating the need for integration of the two systems.

This case study focuses on the occurrence of *C. imicola* and *C. bolitinos* in the Western Cape Province of South Africa. (See Figure 1) The Western Cape Province is located in the south west of the country and is historically an AHS-free zone even though the vector species occur naturally in the area. Since the first recorded outbreak of AHS in this province, in Stellenbosch in 1999, there have been further outbreaks however, specifically in the Knysna/George area (Lord *et al*., 2005). This is cause for concern since such outbreaks could lead to legislation to restrict the movement of horses – especially race horses – countrywide, and could also impact on the export of horses and the hosting of international events (Lord *et al*. 2002).

Figure 1.1: Location of the Western Cape Province

4. The Structure of the Study

The study is structured into three sections. The first section examines the importance of GIS and ANNs as decision-making tools. GIS and neural networks are two separate, potentially complementary systems that can be used to improve decision-making. GIS is explained in terms of various definitions and applications. A detailed explanation of the processes involved in developing and training an ANN is given. In addition, various applications of GIS and ANNs will be discussed.

The second section uses a case study to illustrate how ANN can enhance the decision-making capabilities of a GIS. By combining the GIS and ANN a model is developed that predicts the abundance of *Culicoides* spp. in the Western Cape Province. A full description of the development of the model is given.

The final section focuses on the results obtained from the developed models. The ANN models are tested to select the best prediction model. The model is subsequently used to predict the potential abundance of the *Culicoides* spp. The results are displayed on a map using a GIS.

5. Chapter Overview

The following table provides an overview of the chapters in this study to illustrate the logical flow of the paper to the reader.

Chapter	Contents
1	An **introduction** to the study with short descriptions for the reasons for undertaking the study, a description of the study area, the research aims and secondary aims.
2	**GIS and ANN** are examined and explained in terms of their definitions, processes and applications. The process of developing and training an ANN is explained.
3	An ANN and GIS is used to **develop a model** to predict the possible abundance of *Culicoides* spp. in. The process followed to develop the GIS and the ANN will be explained in detail
4	The **results obtained from the model are discussed**. The various ANN models will be tested to determine the best prediction model. The final predictions are imported in the GIS to present the final result of the model.

5	The project is **summarised** and a final **conclusion is made to** indicate whether the research aims and secondary aims has been reached.

Chapter 2: Geographical Information Systems and Artificial Intelligence

A GIS can be defined as '*a computer-based system to aid in the collection, maintenance, storage, analysis, output and distribution of spatial data and information'* (Bolstad, 2005, p1). Demers (2000, p7) defines GIS in broader terms as '*a tool that allows for the processing of spatial data into information, generally information tied explicitly to, and used to make decisions about, some portion of the earth*'. A definition that is relevant to this study is given by Davis (2001, p.13) as: '*a computer-based technology and methodology for collecting, managing, analysing, modelling and presenting geographic data for a wide range of applications*'. New applications of GIS are emerging on a continuous basis as a result of their widespread appeal. For example, GIS have been used as an analytical tool to assist in crime analysis (Breetzke, 2006); to monitor wildlife movement (Walker *et al.*,1997); to reduce pollution (McDonald *et al.*, 2000); to cope with natural disasters (Barredo, 2007; to analyse AIDS epidemics (Kalipeni *et al.*, 2008); and to improve public health (All *et al.*, 2008). Over the past few decades the focus of GIS has been on providing knowledge and understanding of spatial data, while their significance as a decision making tool has often been overlooked (Thurston, 2002). GIS combined with artificial intelligence (AI) can make a valuable contribution to the decision making process. Especially with recent advances in software which now make AI more accessible by making it possible to run AI from desktop computers.

6. The use of GIS and Artificial Intelligence

Artificial Intelligence (AI) is a well-developed science with new fields of application emerging rapidly. A broad definition of artificial intelligence is '*the study of how to make computers do things at which, at the moment, people are better'* (Rich, 1983, p1). Another definition by Winston (1992, p5) defines AI as '*the study of the computations that make is possible to perceive, reason, and act'*. Artificial intelligence has also been described as a science that seeks to understand, build and construct intelligent entities (Russel, 1995).

There are a number of applications of the use of AI in GIS including waste-water management (Ha *et al.*, 2003); environmental-health management (Bédard *et al.*, 2003); habitat-suitability prediction (Garzón *et al.*, 2006); vegetation management (Deadman *et al.*, 1997); land-quality assessment (Ochola *et al.*, 2004); evaluation of wildlife habitat (Pope *et al.*, 1998). The complementary use of GIS and AI can make information more valuable for decision making (Thurston, 2002) as well as bring more intelligence to other computer-based technologies (Deadman *et al.*, 1997). AI with decision making functionalities includes ANN, fuzzy logic, decision trees, Markov models and evolutionary computation (Thurston, 2002). This thesis concentrates on the use of ANNs to assist GIS in the decision making process.

7. ANNs

An ANN is a type of artificial intelligence based on how the human brain functions (McCloy, 2006). Explanations of the way in which ANNs operate are moving away from this notion towards an applied mathematical technique which incorporates some biological

terminology (Hewitson *et al.*, 1994). ANNs have incorporated two important characteristics of the human brain: their ability to learn through examples, and their ability to interpolate from incomplete information (Hewitson *et al.*, 1994). As a result of these two characteristics, ANNs can model extremely complex features. ANNs have also emerged as an important tool for classification and is a promising alternative to conventional classifiers (Zhang, 2000). The technique has been applied to a variety of applications in classification tasks including bankruptcy prediction (Lacher *et al.*, 1995); speech recognition (Bourlard *et al.*, 1993); medical diagnosis (Baxt, 1990); and handwriting recognition (Guyon, 1991). Although the use of ANNs requires some heuristic knowledge on the working, structure, training and interpretation of an ANN, the level of knowledge needed to successfully apply ANNs is often much lower than would be the case for many other statistical methods.

The functioning of an ANN is broadly modelled on the brain. Accordingly, an ANN consists of neurons, called processors (nodes), which are connected by weighted links (Hewitson, 1994). The basic elements of an ANN consist of a number of inputs –these may be from the original data set or from the output of other neurons – which are linked to a neuron via weighted links. Each neuron has a transfer function which, together with the weights, determines an output. These basic elements of an ANN are illustrated in Figure 1.2.

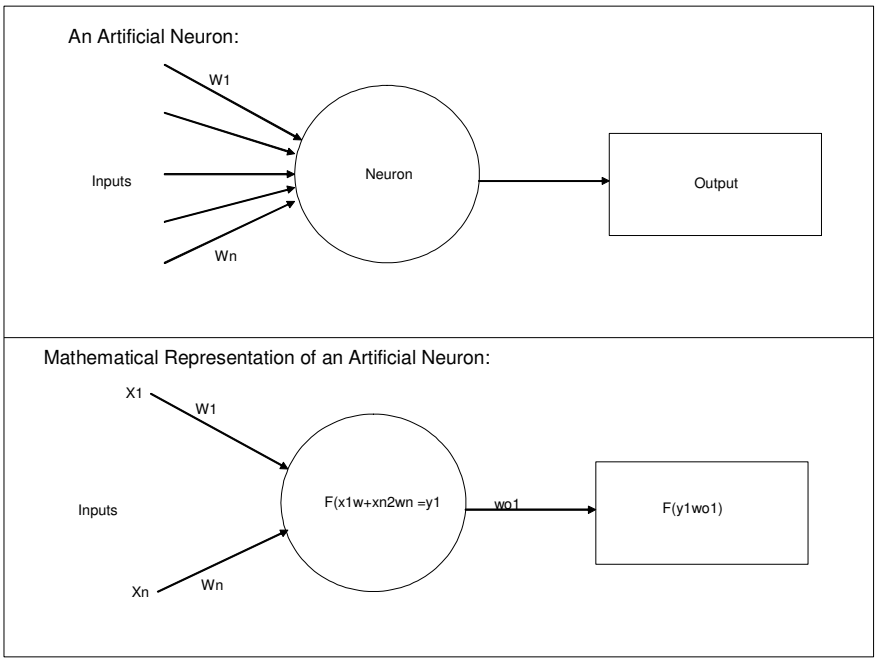


Figure 1.2: Basics Elements of an ANN (Adapted from Saha, 2003)

These basic elements are arranged together to form an ANN. The most generalised type of ANN consists of three separate layers: an input layer, a hidden layer, and an output layer. These layers are shown in Figure 2.2.

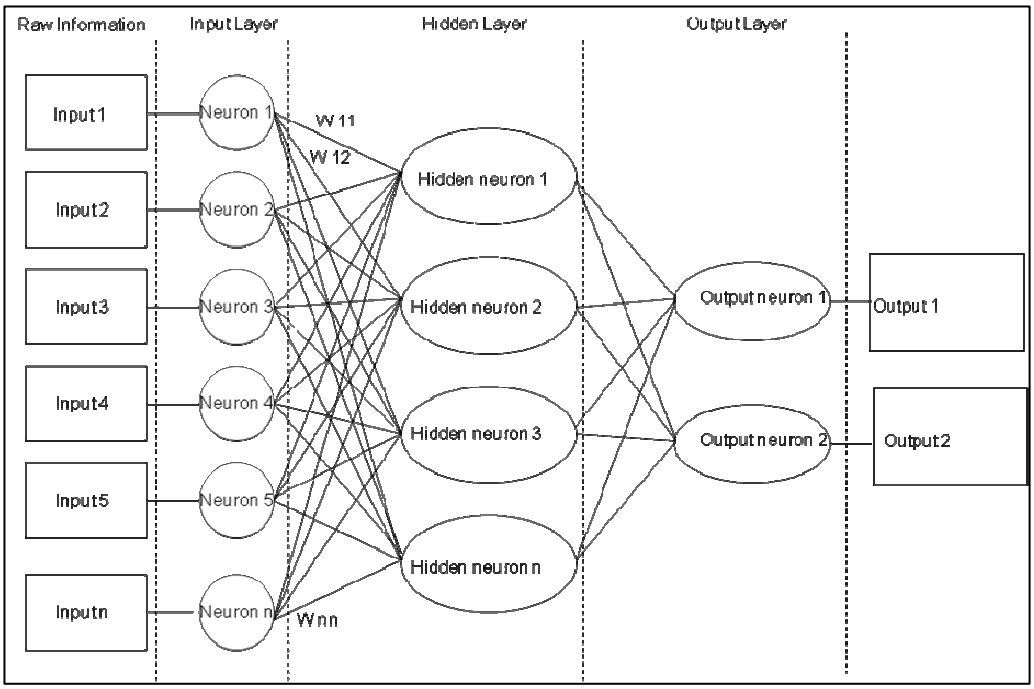


Figure 2.2: Layers of an ANN (Adapted from Saha, 2003)

The input to an ANN consists of raw data. The raw data is linked to the input layer, which consists of neurons which are connected to the neurons in the hidden layers. The hidden neurons are connected to the output neurons in the output layer (Stergiou *et al.*, 1996), each link having a weight associated with it. The links can be negative (i.e. have an inhibitory effect) or positive (i.e. have an excitatory effect). The output neurons in the output layer are linked to the final output (Saha, 2003). Once an ANN has been compiled it can be trained on the existing data to make predictions for unknown cases.

2.1 How to Construct and Run an ANN

When an ANN is used as a classifier to assist in decision-making, a basic process must be followed in order to design and implement it. This process is illustrated in Figure 2.3. The first step in the process is the same as the first step of the GIS process which will be dealt with in the following chapter while the rest of the steps will be explained in more detail in this current chapter.

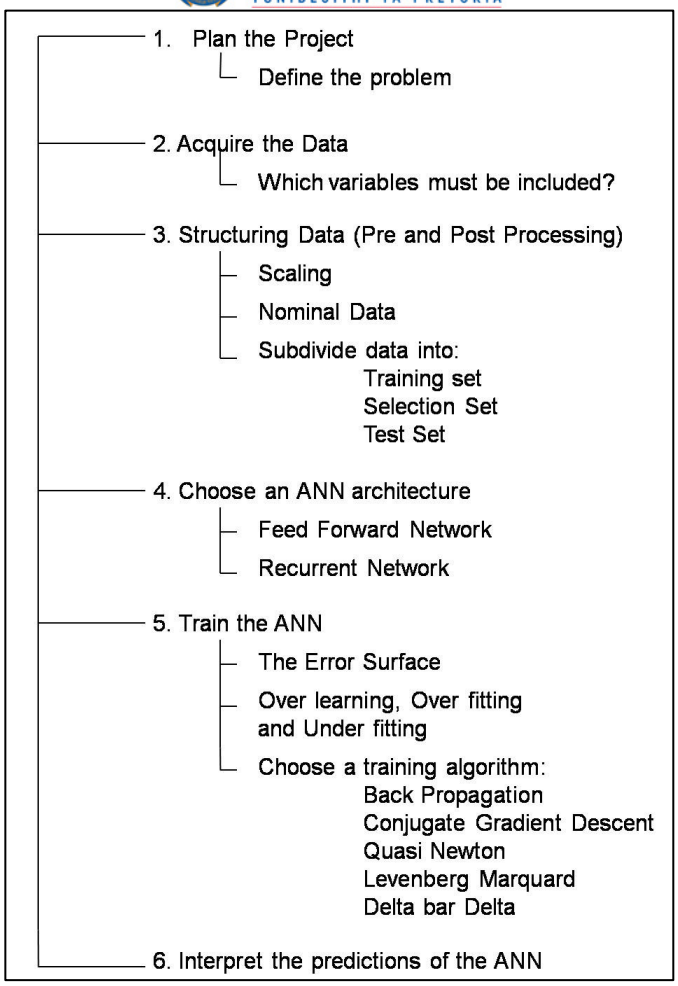


Figure 2.3: Implementing an ANN

2.1.1 Acquiring data for use in an ANN

The acquisition and input of raw data in an ANN is an important first step. The data that is subsequently used for the training of an ANN must typically include a number of cases consisting of input variables and their corresponding output variables (Statsoft, 2008). The most important decision to be made is which input variables to include in an ANN. This choice is guided by the user's intuition and experience in the field. Best practice is typically to include all variables that may influence the output of an ANN and then to narrow down the variables to select the essential ones. According to Statsoft (2008), the following need to be taken into consideration when selecting input variables:

- Since ANNs work on a multidimensional surface, each additional input adds another dimension in which the data cases reside, and can hinder the performance of the ANN;
- Each input variable must be independently assessed for usefulness, so that only the most crucial variables are included in an ANN. In practice, it is seldom possible to do this as the variables and their relationships are not always known;
- It is possible that a number of variables can, to some extent, carry the same information (e.g. the height and build of a person can be an indication of their weight). If the variables are correlated, it may be sufficient to include only some of the variables.

ANNs can handle both numeric and non-numeric (or nominal) values. All non-numeric values must, however, be represented as a numeric value. In some cases the non-numeric values are easy to manipulate (for example, gender, where male can be set as male=1 and female=2). ANNs do not perform well with nominal variables that have a large number of possible values, for instance, date and time variables. Such variables must, therefore, be converted to an offset value from a starting date or time. The next challenge in data acquisition is the problem of missing values, which are interpreted as zero by an ANN. All missing numerical values must be identified and substituted with a statistically calculated value (e.g. mean value) of that variable, or be removed from the data set.

One of the real strengths of ANNs is their ability to handle noisy data, that is, data with corrupted or incorrect values. However, this has its limitations. If there are outliers far outside the normal range, they may skew the data set and bias the training process. So, it is important that outliers are identified and removed from the data sets, or replaced using

some statistical calculation, e.g. a mean, minimum or maximum value. Once the data cases have been collected, some pre-processing is required before the data can be imported into an ANN.

2.1.2 Preparing data for use in an ANN

The preparation of data for use in an ANN is an important step as this determines the success of the training of the ANN. Two issues need to be addressed before the training of a network can begin: the scaling of variables, and the replacement of nominal values. Scaling, which brings two variables closer in term of their numeric value and facilitates the training process, is performed when there is a large difference between the minimum and maximum output variables. Raw variables are usually scaled using linear scaling. However, if a variable is exponentially distributed, non-linear scaling using logarithms, for example, may be necessary. The replacement of nominal variables may be two-state or multi-state. A two-state variable is easily transformed into a numeric value (e.g. dog = 1, cat = 2). Multi-state variables are more difficult to interpret when in a numeric form. An ordinal value can be used (for example, dog = 1, cat = 2, bird = 3; this gives the variables some sense of order, so that dog may be seen as first or as more important than cat or bird). Another approach is known as the one-of-N encoding. In this approach a list of numeric variables is used to represent the nominal value (for example, dog = (1,0,0), cat = (0,1,0), bird = 0,0,1). However, a nominal variable with a large number of states can cause an increase in the network size and complicate the learning process (Statsoft, 2008).

Once all the variables have been scaled and nominal values replaced, the number of cases must be subdivided into three data sets: a training set, a verification set and a test set. These three sets must also be representative of the underlying model and must also be

independently representative of real-world outputs. The training set is used for the initial training of an ANN, while the verification set is used to test the output of the model with real-world outputs and adapt the weights of the links until a minimum error is reached. The test set is used to test the trained network and compare the outputs of the network with the real world before a prediction model is run. It is important that the training data set is carefully selected, otherwise the network cannot be trained properly and the output error will be high. According to Statsoft (2008), the following issues need to be taken into account when selecting a training data set:

- Training data are always selected from a historical data set. If circumstances have changed, some relationships between the variables may no longer be valid.
- All possible occurrences of the variables must be covered by the training data set (e.g. if a maximum rainfall of 100mm is used to train a network, the network cannot be expected to make the correct prediction if the rainfall variable is entered as 200mm). To make correct predictions an ANN must be trained using all minimum and maximum cases that can be anticipated.
- During the training process an ANN tries to minimise overall error. The proportion of different types of data represented in the training set is therefore critical. Assume an ANN is required to determine a 'good' or 'bad' output. If the network is trained using 900 good cases and only 100 bad cases, the trained network will be biased towards the good cases. The best approach is to ensure an even representation of all the different cases

Once the data have been structured into the three data sets, the actual construction of the ANN can start. The first important decision will be the choice of ANN architecture.

2.1.3 Selection of an ANN Architecture.

There are two main types of ANN architecture: feed-forward networks, and recurrent networks. Feed-forward networks have a simple structure and are mostly used when constructing ANNs. The neurons in a feed-forward network have a distinctly layered pattern with the signals in the network travelling in only one direction. A feed-forward network can have any number of hidden layers (Saha, 2003; Statsoft 2008). When the network is executed, the input variables (or raw data) are placed in the input units. The hidden and output layers are then executed progressively. Each layer calculates an activation value by multiplying the weight of the link with the input value. Theses values are then passed through the network to produce the output. This procedure is illustrated below.

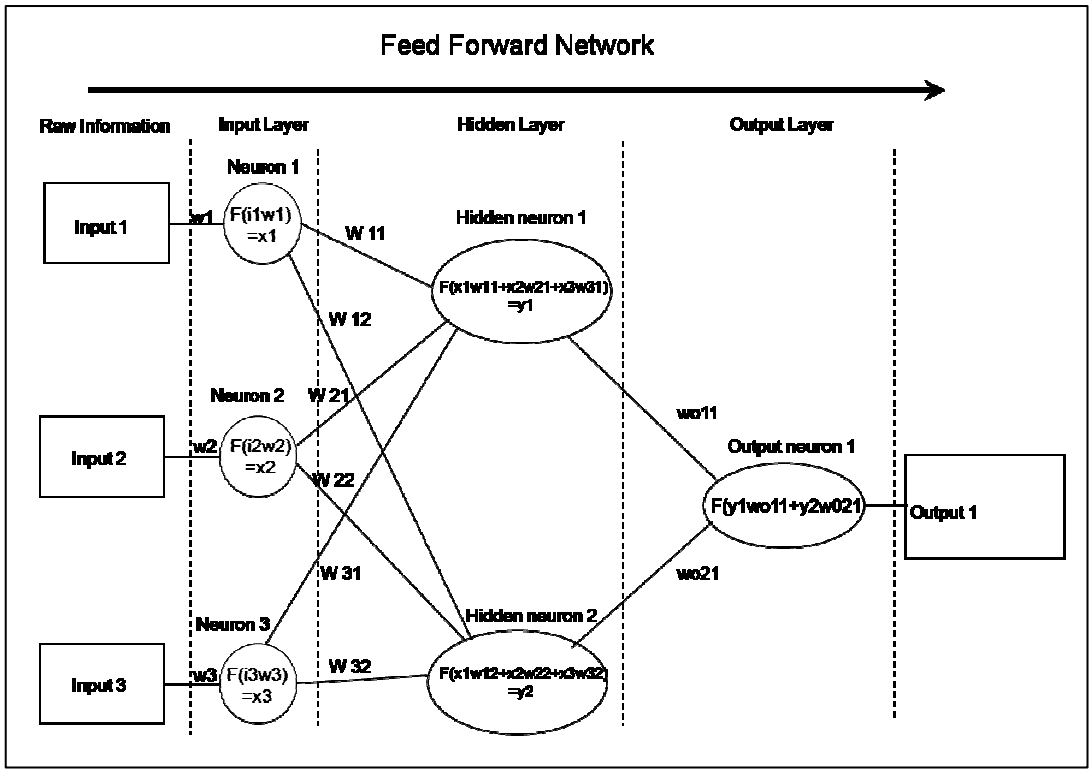


Figure 2.4: An example of a Feed-Forward ANN (From Saha, 2003)

Some connections are not possible in a feed-forward ANN. As illustrated in Figure 2.5, the neurons in one layer cannot be connected to each other on the same layer. Neurons can

only connect to neurons or hidden neurons in the next layer. In addition, links between neurons cannot jump a layer (Saha, 2003).

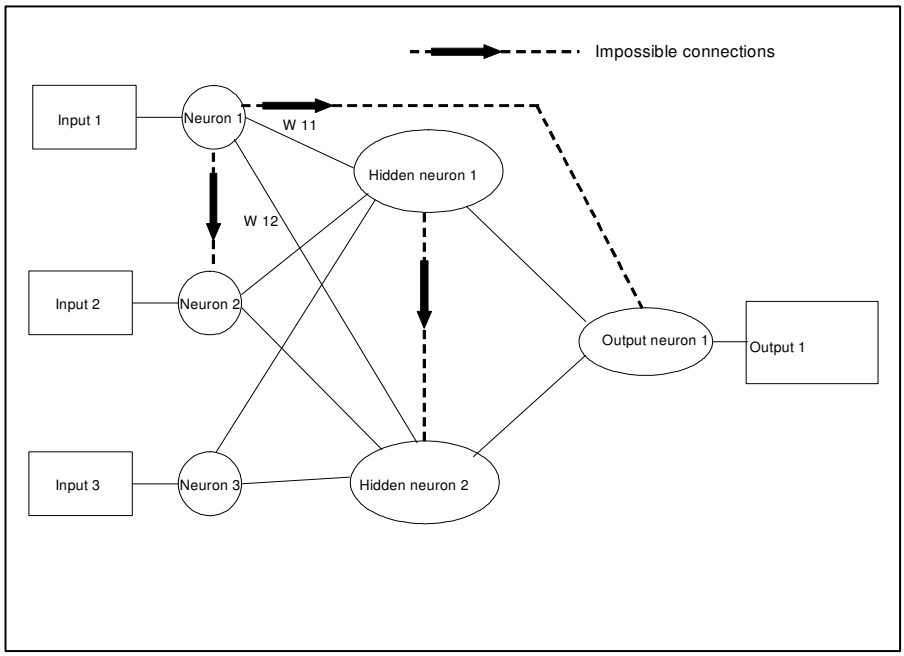


Figure 2.5: Connections not possible in a Feed-Forward ANN (From Saha, 2003)

The second type of ANN is the recurrent network. In contrast with the feed forward network, backward connections are introduced in the recurrent network, i.e. from later to earlier neurons in order for signals to travel in both directions. The network will run in a loop until equilibrium is reached – that is, where the weights are determined and the output error is a minimum. It will stay at this equilibrium point until the input is changed and a new equilibrium needs to be found (Stergiou *et al.*, 1996). An example of a recurrent network is illustrated in Figure 2.6.

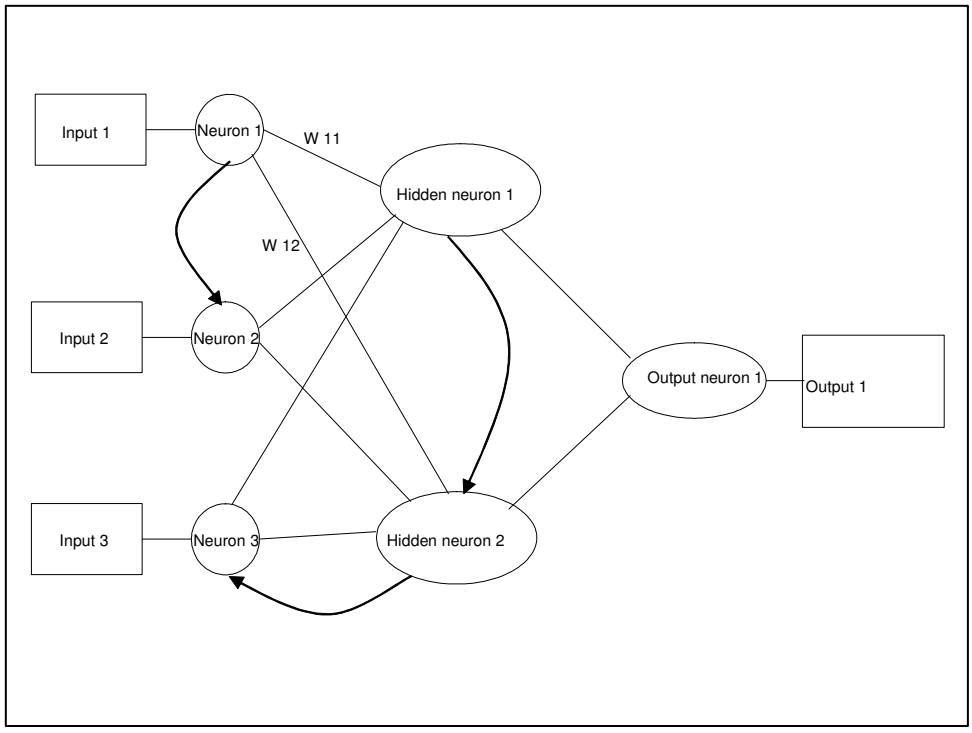


Figure 2.6: An example of a Recurrent Network (From Stergiou *et al*, 1996)

The successful functioning of these two architectural types of ANN depends on the successful training of the networks.

2.1.4 Training an ANN

There is no fixed strategy in the training of an ANN with most of the training being done through trial and error (Saha, 2003). The most commonly used training algorithms for feed-forward ANNs is the back propagation algorithm. Although other algorithms can be used to train an ANN, the back propagation algorithm is the easiest to understand (Statsoft, 2008). (Other algorithms will be mentioned later in this chapter). When training an ANN, the back propagation algorithm progresses iteratively through a number of epochs (an epoch is defined as a single movement through the entire training set followed by testing of the test set). During each epoch, the training cases are submitted to the network and the calculated output of the ANN is compared with the actual output. The error is calculated

and, together with the surface gradient, is used to adjust the weights of the ANN through back propagation (Statsoft, 2008). The whole process is repeated until training is stopped. The training process is basically an exploration of an error surface which is calculated by running all the training cases through the network and calculating an output. This calculated output is then compared with the desired output and the mean square error is calculated (Statsoft, 2008). An error surface is illustrated in Figure 2.7.

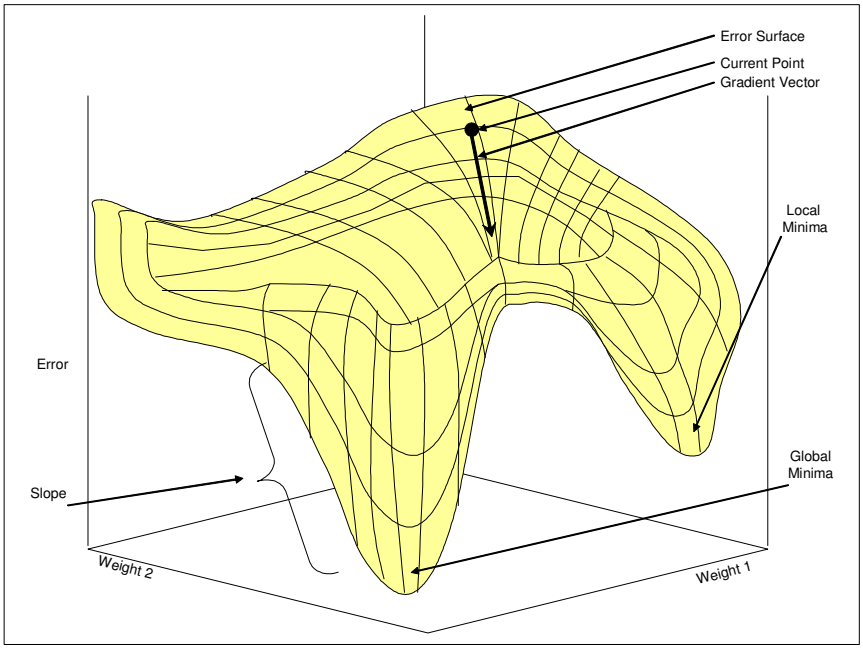


Figure 2.7: A typical error surface (From Saha, 2003)

When using the back propagation algorithm the gradient vector of the error surface is calculated. The gradient vector indicates the line of steepest decline from a current point (See Figure 2.7). The point moves along the gradient vector to find the global minima in order to decrease the output error. The process of calculating the gradient vector and the movement along the gradient are repeated until a minimum error is reached. The movement along the gradient occurs in a series of steps which determine the adjustments in weight sizes, which is the learning rate. The size of the steps determines the direction of movement and how the point moves down a slope. If the steps are too large, the point may overstep the solution or it may bounce from side to side as it moves down a slope. The

network then struggles to find the global minima. If the steps are very small, the current point may move in the right direction but this requires a high number of iterations, slowing the training process. The correct learning rate depends on the application and is typically chosen by experiment. The algorithm also includes momentum, which encourages the movement of the point in a specific direction. If several steps are taken in the same direction, the movement becomes faster over flat spots and gives the network the ability to escape local minima. The training of an ANN stops when a given number of epochs elapse or when the output error reaches an acceptable level. The challenge is that the user usually does not know when a minimum error is met. The best practice is to stop the training if the prediction error increases, or if the overall changes in weights decrease (Saha, 2003). An ANN is not capable of relearning. Consequently, if a different input is added, the whole system needs to be retrained (Thurston, 2002).

A major problem with the training process outlined above is that it does not minimise the expected error made by the network when new cases are submitted. In this regard ANNs can suffer from over-fitting, under-fitting and over-learning. Under-fitting occurs when an ANN is not sufficiently complex to model the problem. Over-fitting occurs when an ANN is too complex for the model and responds to the noise in the data and not to the general signal in the data (Statsoft, 2008). Over-fitting and under-fitting are best illustrated by making use of polynomials (See Figure 2.8 below).

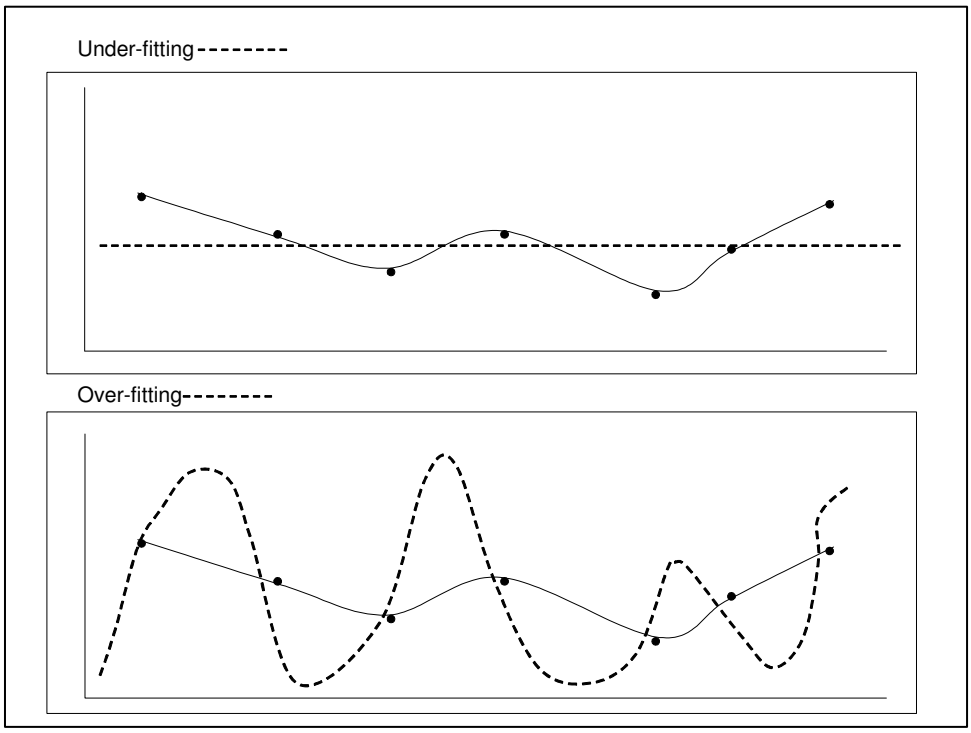


Figure 2.8: Polynomial illustration of Under-Fitting and Over-Fitting (From Statsoft, 2008)

The best way to avoid over- and under-fitting is to test the network against a verification set. An indication of over-fitting is when the verification set stops decreasing and starts increasing; training of the data should be stopped immediately. Over-fitting occurring during the training process is called over-learning: the network is too well-trained for the available data. This will lead to a very small error in the training data, but the network will have poor generalisation power on unseen data (Saha, 2003). In this case it is best to decrease the number of hidden units or hidden layers and restart the training process.

Other more sophisticated training algorithms such as the conjugate gradient descent, quasi Newton, Levenberg Marquard and Delta bar Delta algorithms are also used (Statsoft, 2008). However, in this study the back propagation algorithm is preferred because it is quick and performs better on smaller data sets.

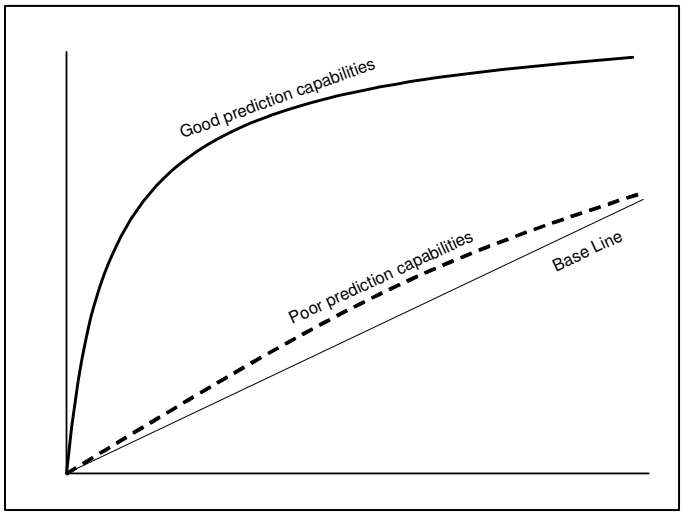


Figure 2.9: Illustration of a typical Lift Chart (From Abouzakhar *et al.*, 2003)

After training an ANN, a lift chart can be created to indicate the effectiveness of the predictive model. The greater the area between the baseline and the lift curve the better the predictive capabilities of the model (Abouzakhar *et al.*, 2003). A typical lift chart is illustrated in Figure 2.9. Once the network has been trained and tested on the test set, the model can be used as a predictive model.

2.1.5 Interpreting the results of an ANN

The output layer of an ANN attempts to assign the value of a specific class. The output of an ANN is a correlation of the input variables and indicates the possibility that the output belongs to a specific class. Correlation is a statistical technique that indicates how strongly two or more variables are related. There are various statistical techniques for determining the correlation between variables. A technique used most often is the Pearson correlation coefficient (r), and values range from –1.00 to +1.00. A correlation coefficient of –1.00 represents a perfect negative correlation, which means that if one variable tends to decrease

another variable tends to increase. A correlation coefficient of +1.00 represents a perfect positive correlation, which means that an increase in the value of one variable will lead to an increase in the value of a related variable. A correlation coefficient of 0 indicates that there is no correlation between the variables. Although correlation coefficients are reported as a value between – 1 and +1, squaring the value and expressing it as a percentage makes it easier to understand: for example, if r = 0.2, the square = 0.4, the decimal is ignored and the correlation coefficient is 40%.

In the case of interpreting an ANN, a value of +1 indicates that the input of the prediction case correlates with the collective, trained inputs of the network related to a specific class. An output value of zero means that there is no correlation between the input of the prediction case and the collective, trained inputs of the network related to that specific class. This means that the input falls 100% outside the class. Values between +1 and 0 are an indication of the probability that the output will fall within a specific class. An output value of 0.75 indicates a 75% probability that a set of inputs will fall inside the class whereas a value of 0.25 indicates a 25% probability that a set of inputs will fall inside the class. A 50% probability indicates that the inputs may or may not fall inside the class.

2.2 Why use ANNs?

Not all problems can be solved using ANNs. ANNs are best suited for cases where there is a known relationship between the variable inputs and outputs, but the exact nature of the relationship is not known. ANNs are indicated in cases where the relationship between the different variables requires a complex mathematical model which has not yet been developed (Deadman *et al.*, 1997). ANNs have the added capability to extract patterns and

trends from data sets too complicated for the human brain to recognise or for conventional computers to calculate (Stergiou *et al.*, 1996). An additional benefit of ANNs is their capability to incorporate uncertainty or noise in the data sets (Yang *et al.*, 2001). Furthermore, ANNs make no assumptions regarding the statistical nature of the data and can integrate nominal and ordinal data. ANNs can be trained using comparatively fewer points than any other statistical model and it is not necessary to choose a data distribution model (German *et al.*, 1997). Trained ANNs can be envisaged as 'experts' in the data on which they have been trained (Stergiou *et al.*, 1996).

2.3 ANNs vs. Exact Classifiers

Classification is common in decision making and can be described as a method whereby an object is assigned to a specific category based on a set of predefined conditions (Bolstad, 20005). One major limitation of traditional classifiers is that the predefined conditions must be known and satisfied in order for the object to be assigned to a specific category. The user must therefore have a good knowledge of the variables influencing the conditions and the correlations between the different variables (Zhang, 2000). In contrast, an ANN relies on statistical methods, including probabilities, to classify variables. An ANN is therefore an example of a probabilistic or statistical classifier. For example an ANN predicts the probability that an outcome will fall into a specific class and does not assign a discrete class. (Stergiou *et al.*, 1996). Since ANNs rely on statistical methods, ANNs can be trained to adapt to any circumstance whereas a GIS is limited to solving problems that the user understands or knows (Stergiou *et al.*, 1996), and only if the steps to solve the problems are known to the user. ANNs have unpredictable outcomes; this is so because they cannot be programmed to perform specific tasks, they need to be trained for each

specific problem, learning by example, and they solve problems on their own. By contrast, the outcomes of GIS are predictable; this is because the algorithm is known. It is important to emphasise that ANNs and exact classifiers are not mutually exclusive but are complementary and can be used together to solve problems. Some tasks are best solved with an algorithmic approach, others with an ANN, yet others with a combination of the two approaches (Stergiou *et al.*, 1996).

2.4 Integration of GIS and ANN

GIS and ANN can be mutually beneficial. Indeed, ANN can benefit from the powerful processing capabilities of spatial data from a GIS, which in turn can provide the ANN with the necessary input data for training of the network. The GIS can also display the results of the ANN in a visual and user friendly format. The ANN on the other hand with its strong decision making capabilities and ability to handle fuzzy data can determine and describe the relationships between the different variables extracted from a GIS. A cursory review of existing literature indicated that the integration of the two systems are not well developed and in most projects the data was transferred from the GIS to the ANN and back (see Rigol-Sanchex *et al*, 2002, Pijanowski *et al*, 2002, Pradhan *et al*, 2008). ANN software that can be used externally to the GIS includes the well known MATLAB (Pradhan *et al*, 2008) and Tiberius (Sarip, 2005). Examples where ANN was developed specifically for use in a GIS include an ANN interface in GRASS GIS (Muttiah, *et al*, nd); and an application in ArcGIS using Visual Basic scripts to specifically predict tunnelling performance in routine tunnel design (Yoo *et al*, 2006). The latter example was specifically developed for use in tunnel prediction and cannot be applied on other examples. The ANN developed in GRASS GIS can be trained for use in any application. However, the effectiveness of the

integration use of the two systems has been proven and various examples of the complementary use of GIS and ANN exist.

8. Applications of GIS and ANN

Applications using a combined GIS/ANN approach include a study to predict human population growth and distribution using historical census data (Graham and Goswani, 2001). In this study a customised graphical user interface (GUI) was developed to allow the extraction of data for analysis using a GIS. An ANN was subsequently developed and trained to use the extracted data to make numerical projections and display these in a GIS. Another example includes the use of an integrative GIS/ANN approach to predict landslide susceptibility based on an analysis of factors such as slope, curvature, soil texture, soil drainage, soil-effective thickness, timber age and timber diameter (Lee *et al.,* 2001). Finally, Pijanowski *et al* (2001) used integrative GIS/ANN approach to predict the location of new urban uses in the year 2020 in the Minneapolis–St. Paul and Detroit metropolitan areas. Other reports of GIS/ANN applications include agricultural land-suitability analysis (Wang,1994); determining the potential location of a road for military land management (Wu *et al*. 2004); developing a system for rapid feedback of potential ecological risks in a flood diversion zone (Ni and Xue, 2003); developing a model to predict water quality (Jiang and Nan, 2006); determining the sedimentology of Gothenburg harbour, a study which shows how an ANN can be used to solve support engineering and harbour management problems (Yang and Rosenbaum, 2001); improving the accuracy of valuations of residential properties by minimising the influence of subjectivity (Sarip, 2005).

In veterinary science, ANNs have been used for three main purposes: for diagnosis; for the determination of species distribution; and for identification of species. The application of ANNs for diagnostic purposes typically involves the identification of diseases. ANNs have been used to identify the bacterial cause of mastitis in dairy herds (Heald *et al.*,2000); to develop a model for early detection of mastitis in cows milked with an automatic milking system (Cavero *et al.,*2008); to detect lameness in milking cows (Pastell *et al.,*2007); to develop a model to assist in the diagnosis of *Ascites* in broilers (Roush *et al.,*1997); and to identify the type of ANN that best predicts susceptibility of chickens to pulmonary-hypertension syndrome (Roush *et al.,* 2001). ANNs have been used to study species distribution of *Colinus virginianus* (Lusk *et al.*, 2002); wolves (Bessa-Gomes *et al.,* 2003); and *Limanda limanda* in UK marine waters (Ward *et al.,* 2006). ANNs have also been used in species identification. In one study, fish species, using parasites as biological tags, were identified through an ANN (and other artificial intelligence techniques including random forests) (Perdiguero-Alonso *et al.,* 2008).

A cursory study of local and international research failed to reveal any known application of ANNs specifically in relation to the prediction of the abundance of *Culicoides* spp. in South Africa or elsewhere. However, two relevant published reports using a GIS were located. In the first study GIS was used to map the potential distribution of C. *imicola* in Europe using climate data (Wittman *et al.,* 2000)[1]. In the other, a GIS was used to model the distribution of *C. imicola* in southern Africa using climate and satellite data (Baylis *et*

[1] Although AHS does not occur in Europe the same vectors carry diseases such as bluetongue (BT). This is of major international concern as it affects the import and export of animals and meat (Wittman *et al.* 2000).

al., 1999). In the forthcoming case study a GIS and an ANN are used to predict the potential abundance of *Culicoides* spp, the insect vectors of the AHS virus.

Chapter 3: Case Study

This chapter demonstrates how a GIS combined with an ANN can be used to predict the abundance of *Culicoides* spp and the areas at risk of AHS outbreaks in the Western Cape Province. What follows is a description of the data requirements and method used to develop just such a GIS/ ANN model. A specific process was followed when integrating GIS and ANN (see Figure 3.1). These steps are specific to this case study and may differ from other applications.

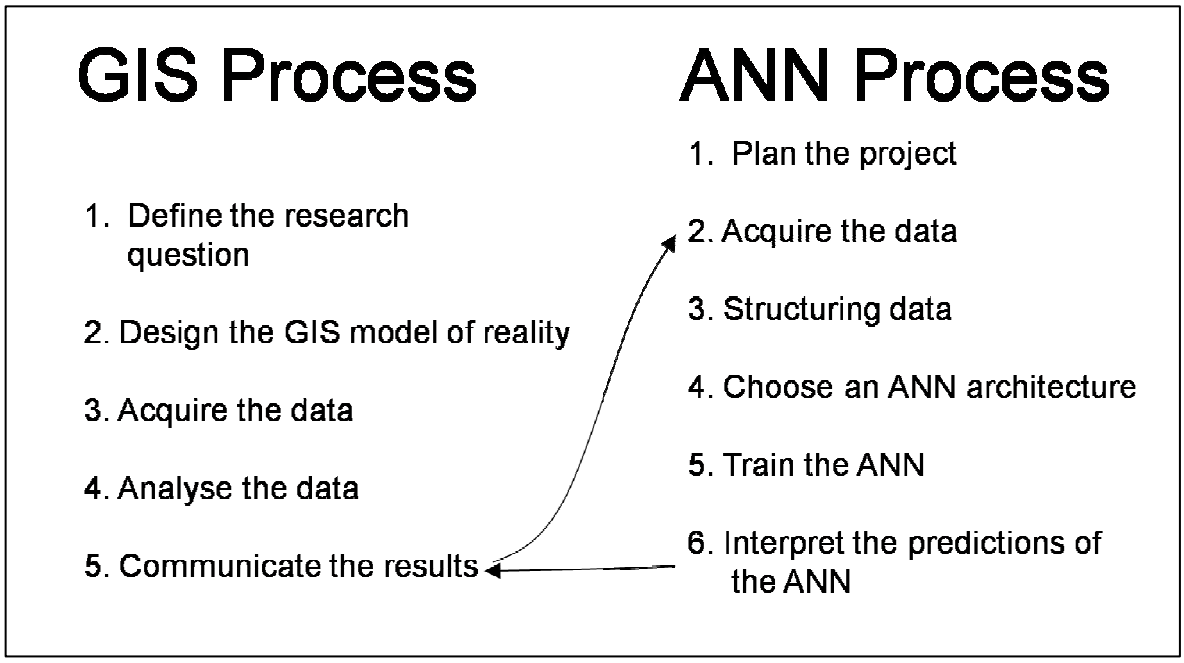


Figure 3.1: Integration of GIS and ANN (Compiled by S P Eksteen, 2009)

9. Aims of the Case Study

The aim of the case study is to develop a GIS incorporating an ANN model that can predict the abundance of *C. imicola* and *C. bolitinos* in the Western Cape Province of South Africa.

The secondary aims of the study are:

- to determine the variables influencing the occurrence of the Culicoides spp.; and
- to predict the abundance of Culicoides at trap points where counts were not made for the particular months during the study period.

10. Data Requirements

The selection of data for the study was based on the research undertaken by Baylis *et al.* (1999) and Wittman *et al.* (2000).The occurrence of the Culicoides spp. in abundance is dependent on various climate factors, the presence of clayey soils, water bodies, livestock density and irrigated fields. The following data sets for the Western Cape for the time period December 2005 to December 2006 were obtained.

(i) Climate Data

Climate data was obtained from the South African Weather Service (SAWS) and the Agricultural Research Council (ARC). The following variables - as monthly averages - were calculated from the climate data:

- total rainfall
- average rainfall
- maximum rainfall
- minimum rainfall
- maximum of the maximum temperatures
- minimum of the maximum temperatures

- average of the maximum temperatures
- maximum of the minimum temperatures
- minimum of the maximum temperatures
- average of the minimum temperatures
- maximum humidity
- minimum humidity
- average humidity

A one-kilometre raster surface was calculated for each of these variables using spatial interpolation.

Long-term monthly minimum and maximum temperatures and rainfall were used to calculate anomalies – deviations from long-term averages – that could possibly favour an upsurge in *Culicoides* spp. population density:

- monthly maximum temperature anomalies
- monthly minimum temperature anomalies
- monthly rainfall anomalies

(ii) Distribution of *Culicoides* spp.

Total daily counts and the geographic distribution of *C. imicola* and *C. bolitinos,* were obtained from the Entomology Division of the Onderstepoort Veterinary Institute as GPS points. These were imported and displayed in the GIS. The Entomology Division relies on the traps being set up by farmers and since this was erratic, counts for most months were incomplete. There were no *Culicoides* counts for June, July and October 2006.

Monthly averages and monthly totals of the *Culicoides* spp. were also calculated for each species and for both species combined. (See Annexure 2 for a map of trap locations in the Western Cape. province) Species counts should be seen not as an accurate count of absolute numbers. Rather they should be seen as an indication of whether or not *Culicoides* spp. occur in abundance (Venter, 2008, personal communication), a *Culicoides* spp. count greater than 1000 *Culicoides* spp. being regarded as an abundant population density (Venter, 2008, personal communication).

(iii) Clay Areas and Water Bodies

The location of potential breeding sites for *Culicoides* spp. (Wittman *et al.*, 2001) – clay areas and water bodies – was obtained in electronic format from the Environmental Potential Atlas as released by the Department of Environmental Affairs and Tourism. Since *Culicoides* spp. can easily spread as much as two kilometres away from their breeding sites (Meiswinkel *et al.,* 2004), a two-kilometre buffer zone was created around all the water bodies and clay areas and converted to a one-kilometre raster layer.

(iv) Normalised Difference Vegetation Index (NDVI) and Land Surface Temperature

Land surface temperature and NDVI data were obtained as one-kilometre grid raster images from the Moderate Resolution Imaging Spectroradiometer (Modis) website (http://modis.gsfc.nasa.gov/). NDVIs were obtained as monthly averages for the time period covered, while images with the lowest possible cloud cover per month for the land surface temperature were used as a raster layers in the GIS.

(v) Altitude

Altitude plays a significant role in the geographic distribution of *Culicoides* spp. (Baylis *et al.*, l999, Wittmann *et al.*, 2001). Accordingly, a one-kilometre digital terrain model (DTM) of South Africa was obtained from the Department of Geography, Geoinformatics and Meteorology, University of Pretoria.

(vii) Livestock and Field Boundaries

The geographic distribution of livestock per magisterial district – indicating the total number of cattle, sheep, poultry and horses in a magisterial district – was obtained from the Directorate: Animal Health of the Department of Agriculture. Livestock density per magisterial district was calculated by dividing the total animal population for a district by its area. These values were then converted to a one-kilometre raster layer for further use in the GIS. This layer is significant since *Culicoides* spp. breed and can survive cold winters in cattle dung, and animals other than horses also serve as hosts for *Culicoides* spp. (Meiswinkel *et al.,* 2004).

Field boundaries were obtained from the Department of Agriculture. No information regarding farming methods was available, so, for the purpose of this study, all cultivated fields were assumed to be irrigated. A two-kilometre buffer zone was created around all irrigated fields as the *Culicoides* spp. can easily spread two kilometres beyond their breeding sites. These buffer zones were converted to a one-kilometre raster layer and imported into the GIS for further analysis.

11. Construction and Analysis of the GIS database

The data sets described above were combined into a GIS and stored on a monthly basis for the period December 2005 to December 2006. Since climate has a delayed effect on the population growth of *Culicoides* spp. of 15-30 days (Meiswinkel *et al.,* 2004; Venter, 2008, personal communication), species counts for a specific month were combined with the NDVI and climate data for the previous month.

3.1 Extraction of Data from the GIS

After incorporation into the GIS, the data were extracted for use in the ANN. For the purpose of extracting the data the *Culicoides* spp. capture sites of the *Culicoides* spp. were used as extraction points. Raster values for each capture site (or trap) were extracted for each layer and combined in an Excel spreadsheet containing raster values for all the GIS layers per month (See Figure 3.2). The '*Extract values per point*' available in ArcGIS 9.2 was used for this purpose.

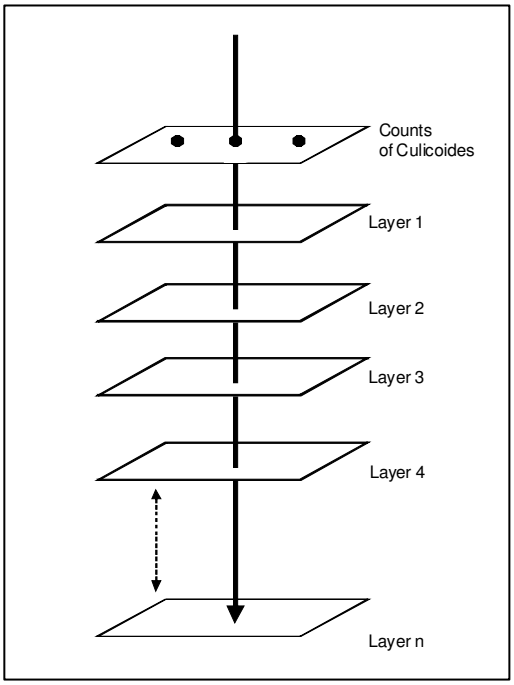


Figure 3.2: Extraction of data per trap

The raster values for climate, altitude, livestock density, NDVIs and LSTs were actual values. These values were either measured in the field or estimated using spatial interpolation. (Spatial interpolation is the calculation of values at unmeasured locations (Bolstad, 2005).) The raster values for the buffer zones calculated for the clay areas, water bodies and cultivated fields were assigned a value 1 or 0. A value of 1 indicates that an extraction point is located within a two-kilometre buffer zone calculated for a relevant feature. A value of 1 indicates therefore that the probability of the abundance of *Culicoides* spp. is high. A value of 0 indicates that the extraction point is located beyond a two-kilometre buffer zone and the probability of the abundance of *Culicoides* spp. is low. Figure 3.3 illustrates the model developed in ArcGIS 9.2 using the *Model Builder* tool to execute extraction of monthly data.

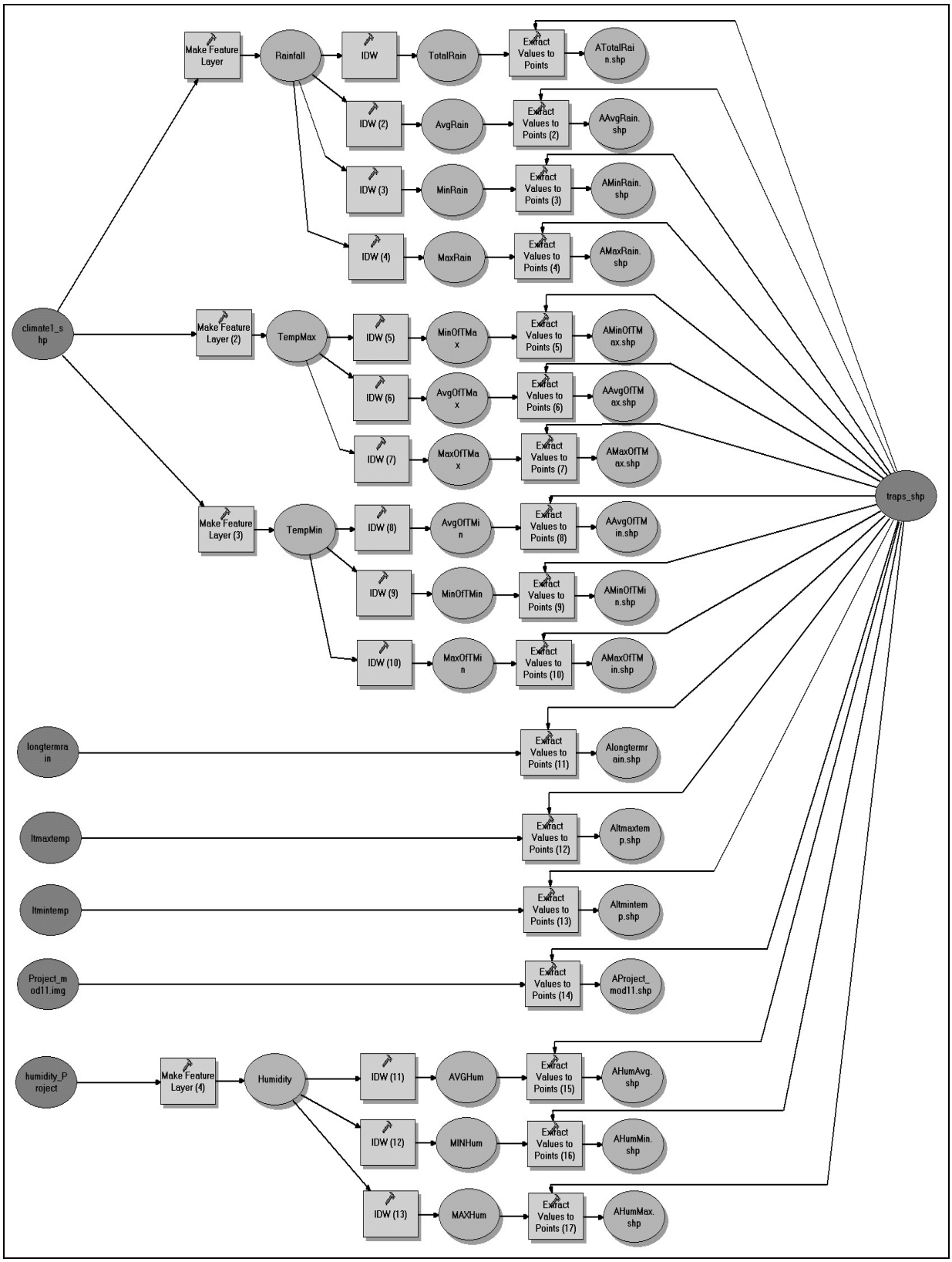


Figure 3.3: Model to extract data per month

A separate model was developed to extract the values for all variables on a monthly basis that stay the same during the time period studied. These layers include altitude, livestock density, clay soils and the buffers around the water bodies and irrigated farms. This model is shown in Figure 3.4

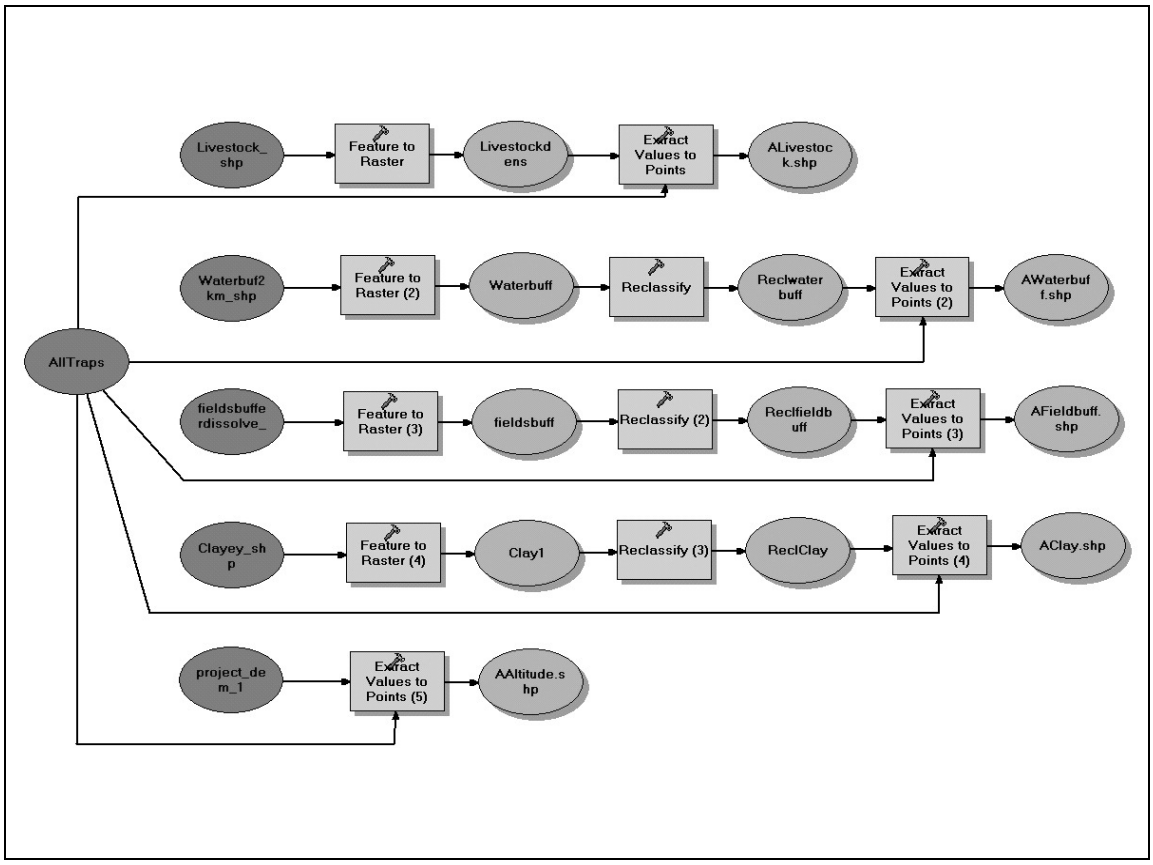


Figure 3.4: Model to extract all other Data Sets (Compiled by S P Eksteen, 2009)

The data extracted from the various GIS layers were combined in an Excel spreadsheet which became the input for the ANN. (See Figure 3.1: Integration of GIS and ANN.) Table 3.1 contains a list of all the variables included in the spreadsheet that is extracted per month per point for the time period December 2005 to December 2006.

Table 3.1: List of variables extracted for use in the ANN

Variable	**Raster Value**
Average Counts of Culicoides	Average Calculated
Total Rainfall	Interpolated Value
Average Rainfall	Interpolated Calculated Value
Maximum Rainfall	Interpolated Calculated Value
Minimum Rainfall	Interpolated Calculated Value
Maximum of The Maximum Temperature	Interpolated Calculated Value
Minimum of The Maximum Temperature	Interpolated Calculated Value
Average of The Maximum Temperature	Interpolated Calculated Value
Maximum of The Minimum Temperature	Interpolated Calculated Value
Minimum of The Maximum Temperature	Interpolated Calculated Value
Average of The Minimum Temperature	Interpolated Calculated Value
Maximum Humidity	Interpolated Value
Minimum Humidity	Interpolated Value
Average Humidity	Interpolated Calculated Value
Long-Term Maximum Temperature Anomalies	Interpolated Calculated Value
Long-Term Minimum Temperature Anomalies	Interpolated Calculated Value
Long-Term Rainfall Anomalies	Interpolated Calculated Value
Clay Areas	1= Inside Buffer, 0 = Outside Buffer
Water Bodies	1= Inside Buffer, 0 = Outside Buffer
Ndvi	Actual Value
Lst	Actual Value
Altitude	Actual Value
Livestock Density	Calculated Value
Cultivated Fields	1= Inside Buffer, 0 = Outside Buffer

12. Constructing and running the ANN

The monthly average count of the *Culicoides* spp. was used as the output variable in the ANN. All the variables extracted from the GIS were used as input variables for the ANN. A total of 99 traps were set up in the Western Cape Province which was to be counted at regular time intervals during January 2006 to December 2006. Therefore a total of 1189 traps were to be counted of which only 337 were done. The averages and total counts of the *Culicoides* spp. as well as the frequency per month are summarised in Table 3.2 (See Annexure 2 for a summary of the counts per trap per year.) The frequency indicates the number of counts that had been done per trap. NNClass that incorporates all the algorithms for ANN training and prediction in an Excel spreadsheet was used as ANN software.

Table 3. 2: Average, total and frequency of counts of *Culicoides* spp. per month for 2006 for all 337 traps.

Month	Total	Average	Frequency
January	166040	2218	46
February	559496	4929	30
March	303753	7364	12
April	341167	2527	68
May	245282	7214	13
June	6650	116	20
July	2993	61	46
August	12813	249	7
September	9398	414	10
October	8635	328	7
November	126498	1095	53
December	34270	650	25

All 337 records in the spreadsheet were investigated to identify the minimum and maximum values for the variables in order to include these values in the training set. The data set was then subdivided into the training set, the verification set and the test set. The

training set consists of 271 records (80%) and the records were chosen so that all the identified minimum and maximum values of the variables were included in the training set. The training set was also geographically representative of the study area and includes records from all climate seasons for the year 2006. The training set included the verification set as the ANN software will randomly select a verification set during the training process. The test set consists of 66 records (20%) and will be used at a later stage to select the best ANN model for the prediction of the occurrence of *Culicoides* spp.

The process of training the ANN was now ready to commence. The training and verification set was imported into the ANN software. The software uses a feed forward network with back propagation as a training algorithm. The training of the network started by including all the variables and with the number of epochs set on 50. The number of epochs were then raised and lowered together with changes in the momentum and learning rate until a minimum percentage misclassification on the training and validation sets were reached. The changes in the parameters also ensured that the ANN finds a global minimum on the error surface. As soon as an acceptable percentage misclassification on the training and validation set was reached the training of the network was stopped. Some of the variables were then omitted and the whole process was repeated.

After the ANN models were trained a number of the models with the least percentage misclassified was chosen and tested using the test set. The best predictive model was chosen and used to predict the occurrence of *Culicoides* spp. at the 852 trap points where no counts where made for the particular month. The results from the ANN were imported back into ArcGIS 9.2 and a classification map of the occurrence of Culicoides in the Western Cape province was created

Chapter 4: Results and Discussion of the Case Study

Chapter 4 highlights the results of the study and identifies the variables and model used to predict the abundance of *Culicoides imicola* and *Culicoides bolitinos*. Subsequently the model is used to predict the abundance of *Culicoides imicola* and *Culicoides bolitinos* in the Western Cape province. The chapter concludes with recommendations for future research.

13. Results

Various combinations of variables were used in the training of the ANN. During each attempt various ANN models were trained as to ensure that the model finds the global minimum and minimise the output error. This is done by retraining the model using the same predictors but changing the number of epochs and hidden layers. A summary of the results of the training of the networks is given in Table 4.1. All variables were included in the model in the first attempt to train the ANN. This resulted in a 13% misclassification of the predicted value when tested against the training set. With the second attempt, no categorical data (irrigated fields, clay areas and water bodies) were included in the training set. This resulted in a slight increase in the percentage misclassified on the predicted values when tested against the training and validation sets. In an attempt to determine the effects of the LST and the NDVI, all other variables were excluded from the model. This resulted in an increase in the percentage misclassified on the predicted values when tested against the training set. The models developed subsequently focused mainly on the inclusion or exclusion of climate data. Most had an acceptable percentage misclassified on the predicted values when tested against both the training and validation sets.

Table 4.1: Model results

Variables included	% Misclassified Trainings Set	% Misclassified Validation Set
1. Model 1: All variables included	13.89	18.18
2. Model 3: All variables included	15.74	18.18
3. Model 6: All variables included	14.35	20
4. Model 8: All variables included	13.43	16.36
5. Model 10: All variables included	15.74	14.55
6. Model 12: Categorical data excluded	18.06	10.91
7. Model 14: Categorical data excluded	12.96	18.18
8. Model 15: Categorical data excluded	11.11	16.36
9. Model 18: Only NDVI and LST	21.3	18.18
10. Model 19: Only NDVI and LST	21.76	16.36
11. Model 25: NDVI and LST, altitude, anomalies	16.67	18.18
12. Model 27: NDVI and LST, altitude, anomalies	19.91	14.55
13. Model 32: NDVI and LST, altitude, anomalies	16.2	20
14. Model 38: NDVI and LST, altitude	21.3	20
15. Model 41: NDVI and LST, altitude	20.83	18.18
16. Model 45: All rain, all temperature, NDVI, LST	18	21
17. Model 46: All rain, all temperature, NDVI, LST	17	21.82
18. Model 47: All temperature NDVI, LST	18.98	20
19. Model 53: All temperature NDVI, LST	18.98	14.55
20. Model 54: All temperature NDVI, LST	14.81	21.82
21. Model 57: All temperature NDVI, LST, rain anomalies	17.59	14.55
22. Model 58: All temperature NDVI, LST, rain anomalies	15.28	18.18
23. Model 59: All temperature NDVI, LST, rain anomalies	13.43	18.18
24. Model 65: Only Climate data (anomalies, humidity excluded)	18	20
25. Model 66: Only Climate data (anomalies, humidity excluded)	15.74	18.18
26. Model 80: Climate and anomalies (no long-term or averages)	18.98	10.91
27. Model 84: Climate and anomalies (no long-term or averages)	17.13	16.36

The models with the lowest percentage misclassification of the predicted values when tested against the training and verification sets were chosen and tested using the test set.

Table 4.2: Results obtained from testing the models

Variables included	% Correctly classified as category 0	% Correctly classified category 1	% Correctly classified
1. Model 8: All variables included	92	49	83
2. Model 15: Categorical data excluded	90	23	77
3. Model 18: Only NDVI and LST	0	0	0
4. Model 19: Only NDVI and LST	0	0	0
5. Model 27: NDVI and LST, altitude, anomalies	0	0	0
6. Model 41: NDVI and LST, altitude	98	23	69
7. Model 46: All rain, all temperature, NDVI, LST	98	1	80
8. Model 53: All temperature NDVI, LST,	100	15	83
9. Model 57: All temperature, NDVI, LST, rain anomalies	88	38	79
10. Model 66: All Climate data only (no anomalies) no humidity	94	23	80
11. Model 84: Climate and anomalies, (no long-term or averages.)	88	15	75

As can be seen in Table 4.2 above, not all the models were accurate predictors. Models 18 and 19, which included only NDVIs and LST, for instance, has an acceptable 21% misclassification of the predicted values when tested against the training set but performed poorly when the predictive capabilities of the models were tested. This poor predictive capability is also seen in the lift chart (see Figure 4.1) developed after using the LST and NDVIs to train the ANN. (Lift charts were explained in Chapter 3.)

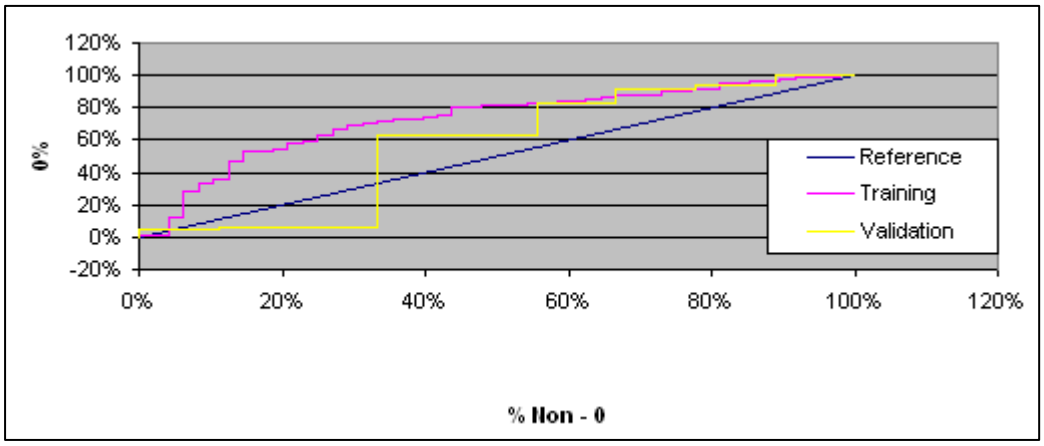


Figure 4.1: Lift Chart for Model 19

Based on the highest percentage correctly classified predictions, Model 8 was selected as the model with the best prediction capabilities. The good prediction capabilities of this model are also illustrated in the lift chart developed after training the ANN (see Figure 4.2)

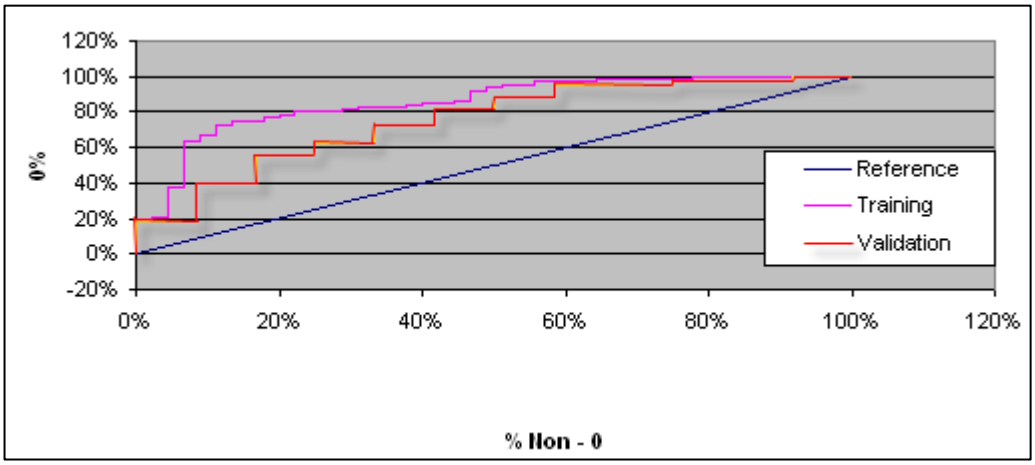


Figure 4.2: Lift Chart for Model 8

The variables included in Model 8 which were found to be significant factors in the predictive accuracy of the abundance of *Culicoides* were:

- NDVI
- LST

- total rainfall
- average rainfall
- maximum rainfall
- minimum rainfall
- long-term rainfall
- maximum of the maximum temperature
- average of the maximum temperature
- minimum of the maximum temperature
- long-term maximum temperature
- maximum of the minimum temperature
- average of the minimum temperature
- minimum of the minimum temperature
- long-term minimum temperature
- maximum humidity
- minimum humidity
- average humidity
- livestock density
- fields
- clay areas
- rain anomalies
- maximum temperature anomalies
- minimum temperature anomalies

Model 8 was subsequently used to predict the abundance of *Culicoides* in the Western Cape Province at trap points where counts were not made for the particular months during the study period.

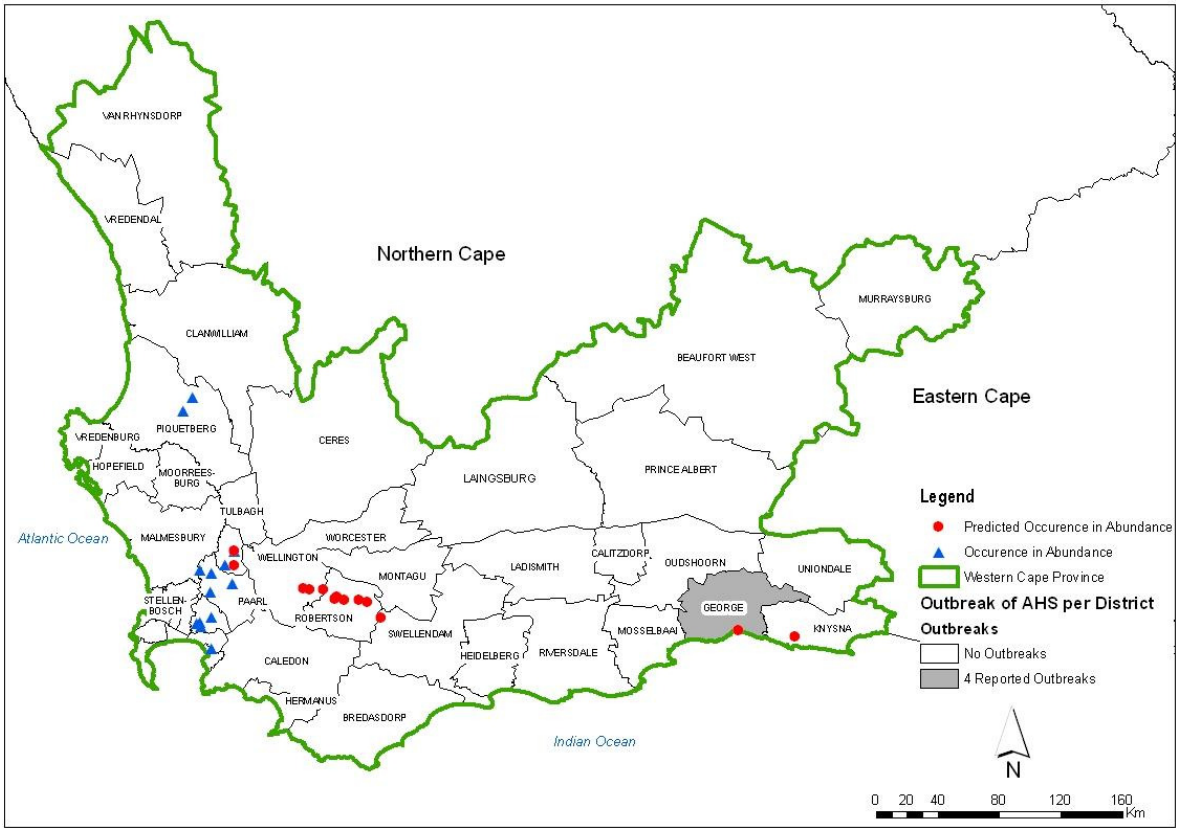


Figure 4.3: Predicted abundance of *C. imicola* and *C. bolitinos*: January 2006

The predicted abundance of *C. imicola* and *C. bolitinos* for January 2006 (see Figure 4.3) for the George district coincided with an outbreak of AHS.

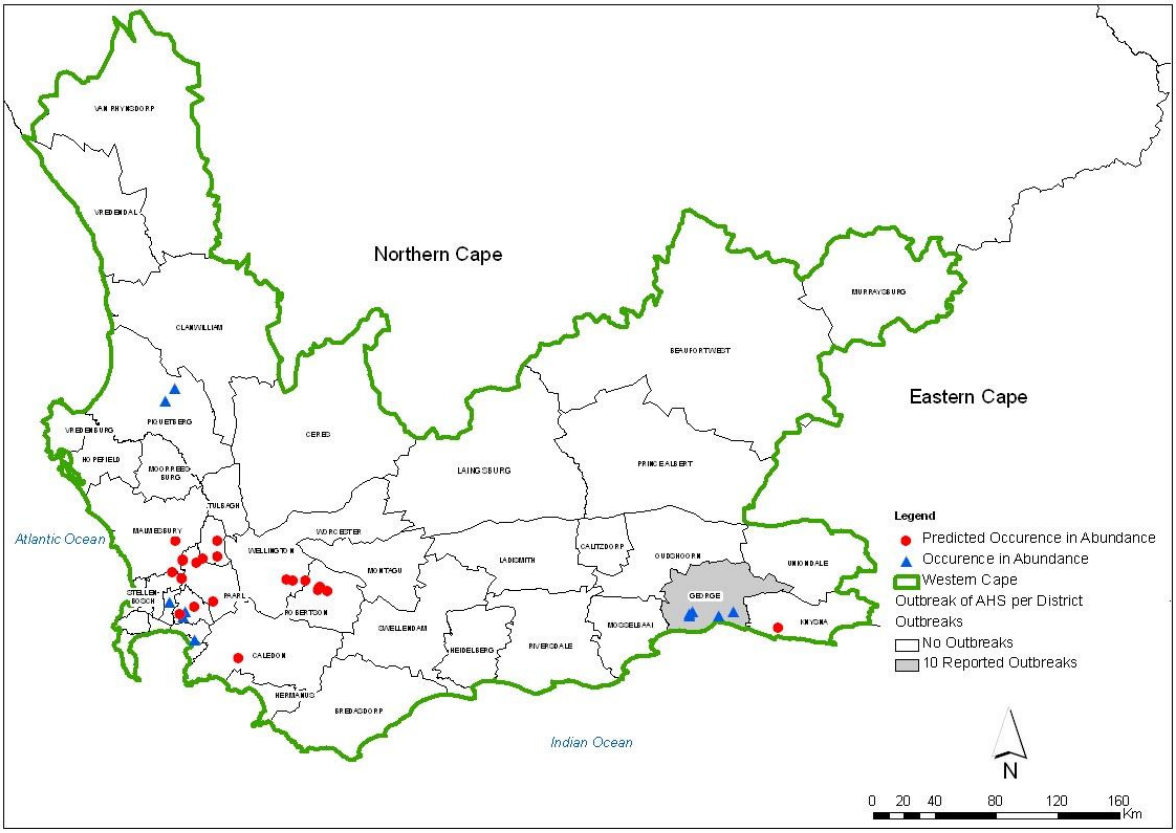


Figure 4.4: Predicted Abundance of *C. imicola* and *C. bolitinos*: February 2006

The abundance of *Culicoides* predicted by the ANN model for the Stellenbosch district for February (see Figure 4.4) and March 2006 (see Figure 4.5) **coincided with the actual count. For the George district for both February and March 2006 (Figure 4.4 and 4.5), and for the Robertson district for March 2006 (Figure 4.5) the predicted abundance coincided with an outbreak of AHS**

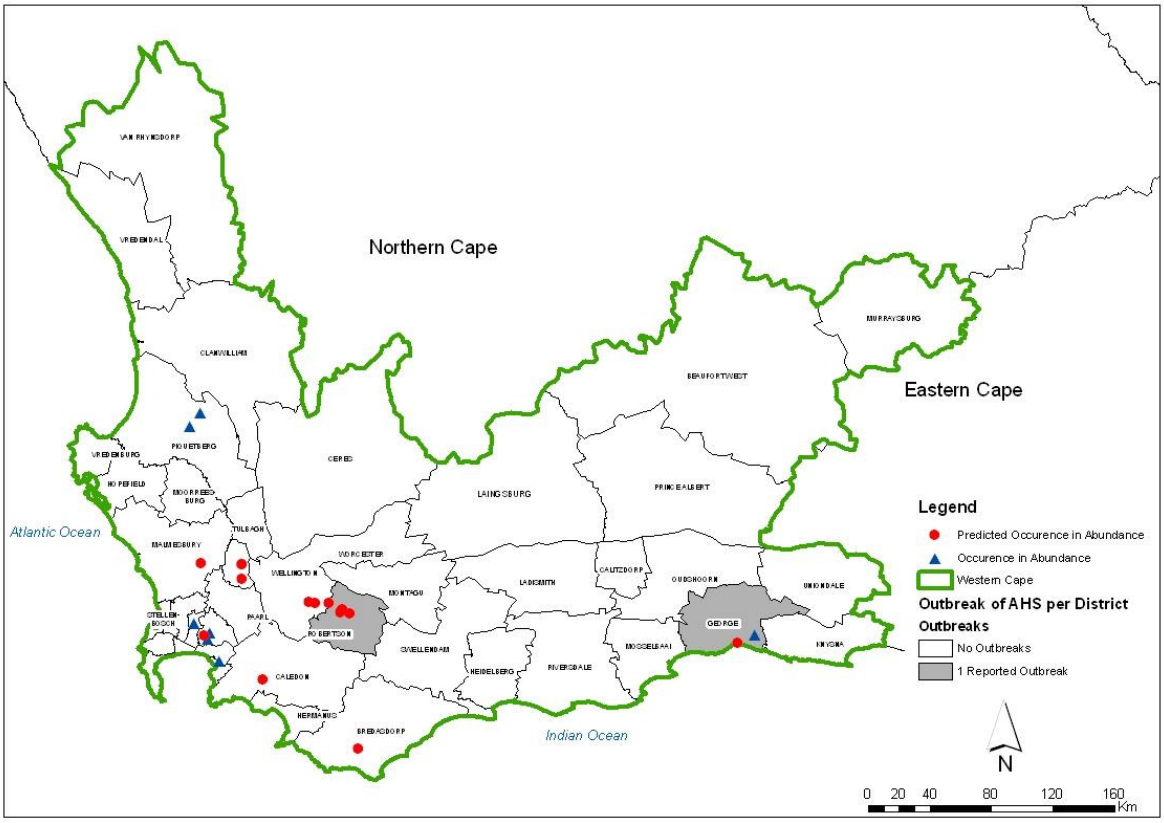


Figure 4.5: Predicted Abundance of *C. imicola* and *C. bolitinos*: March 2006

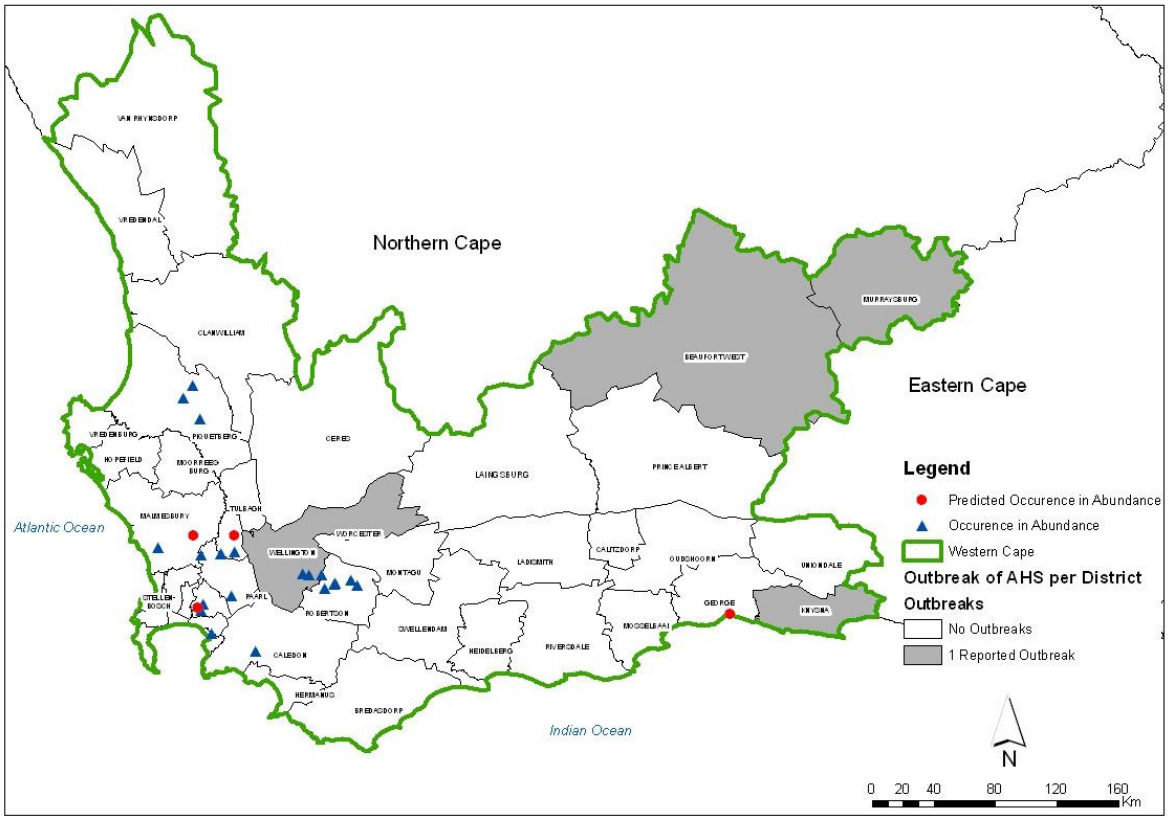


Figure 4.6: Predicted Abundance of *C. imicola* and *C. bolitinos*: April 2006

There was an outbreak of AHS in the Murraysburg and Beaufort West districts (see Figure 4.6) in April 2006, and in the Murraysburg, Beaufort West and Oudtshoorn districts in May 2006 (see Figure 4.7). No predictions of the abundance of *Culicoides* in these districts were made. The reason for this is the unequal distribution of traps the in the study area that has lead to the under representation of these districts in the ANN model. (See also annexure 2 for the location of traps in the study area.) The predicted values for the months April and May 2006 also coincided with actual counts or are in the vicinity of outbreaks of AHS.

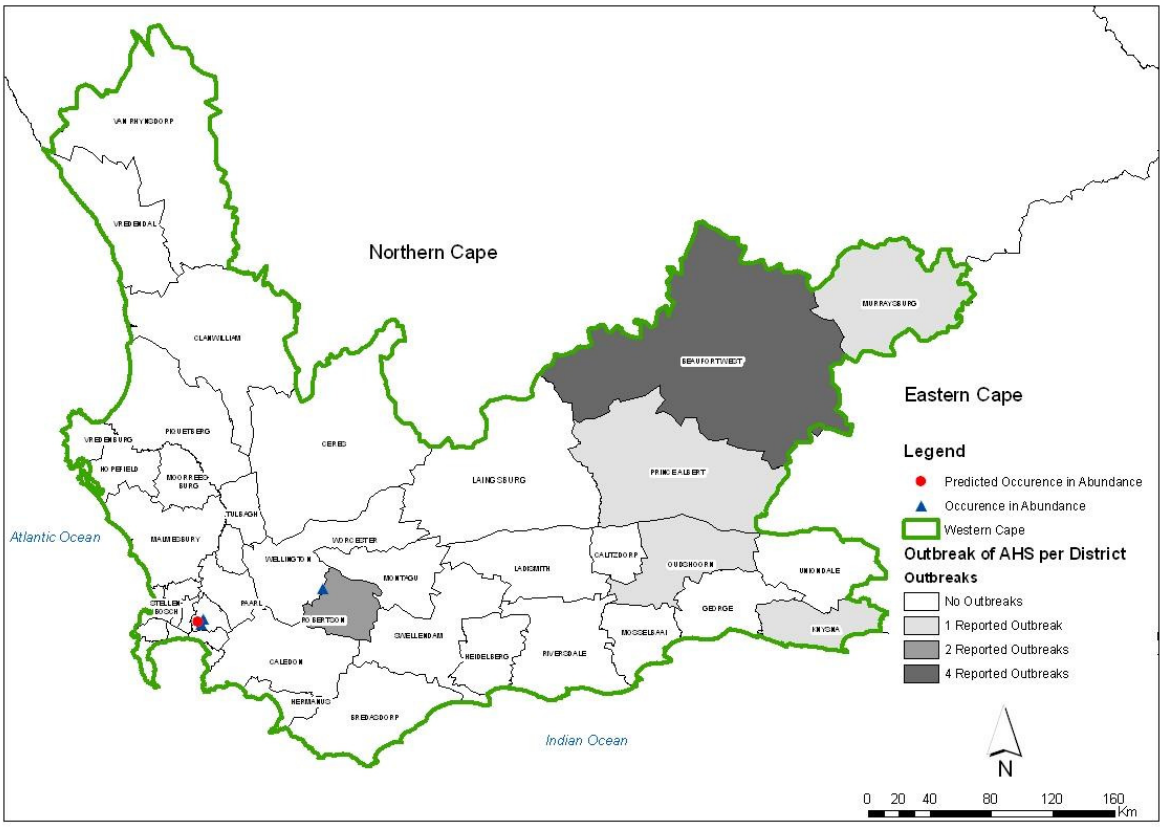


Figure 4.7: Predicted Abundance of *C. imicola* and *C. bolitinos*: May 2006

The ANN model predicted a zero probability of abundance of *Culicoides* for June, July and October, and there were no outbreaks of AHS during this time. During August 2006 there was only one counted abundance of *Culicoides* in the Robertson district, and during September 2006 in the Stellenbosch district and no abundance of the vectors were

predicted. For the experimental results it was deemed not necessary to do predictions for November 2006 since most of traps had been set up by the farmers for the counting of *Culicoides*.

14. Validation of the Model

The model was validated against the actual counts done by the Department of Agriculture while training the ANN (called the verification set) and also by testing the model against the test set. The validation of the model during the training process and the testing against the verification set has indicated a 86% and 82% accuracy respectively. The validation of the model against the test has indicated that 92% of the predictions were correctly classified as no abundance of *Culicoides* whereas 49% has correctly classified an abundance of *Culicoides*. In total 83% of the predictions where correctly classified. Furthermore, the predicted abundance of the *Culicoides* was compared with the counted abundance and the actual outbreaks of AHS. The outbreak of AHS occurs under certain climate conditions and correlates strongly with the abundance of *Culicoides* spp. (Meiswinkel *et al.,* 2004). The outbreaks of AHS are published by the Department of Agriculture per district and the exact locations of the outbreaks are not known. Since the diseases spreads rapidly the whole district as well as the adjacent districts will be warned if at risk. If an abundance of Culicoides are known due to actual counts or predictions the district will be warned as such. Therefore if an abundance of *Culicoides* spp. can be predicted in a district where there was an outbreak of AHS it is an indication of the good prediction capabilities of the model.

Although the actual and predicted counts of the *Culicoides* are done at a specific point it should not be interpreted as such as the midges are not restricted to that specific point and can travel long distances especially on prevailing winds.(Meiswinkel *et al.,* 2004) Therefore if the predicted abundance of Culicoides are in the vicinity of actual counts indicating an abundance it is a further indication of the good prediction capabilities of the model.

15. Discussion

The ANN model proved to be highly accurate in predicting the abundance of *C. imicola* and *C. bolitinos,* with a prediction capability of 83%, which roughly corresponds to that of the GIS model developed by Wittman *et al.* (2001). The variables included in the GIS models of Baylis *et al.* (1999) and of Wittman *et al.* (2001) to predict the probable abundance of *Culicoides* spp. were mostly included in the current ANN model. For the Baylis *et al.* (1999) GIS model, rainfall was considered not to be a significant factor, while for the Wittman *et al.* (2001) model the specific site for which abundance was predicted determined whether or not rainfall was significant. The ANN model described here performed better as a predictor with rainfall included. This is supported by the research of Meiswinkel *et al.* (2004), which showed a clear link between abundance of *Culicoides* spp. and above-average rainfall. The ANN model also performed better when climate anomalies, not included in the GIS models mentioned above, were included.

Improved accuracy of prediction of the ANN model over these two GIS models was achieved with the inclusion of livestock density data and field boundary data (Meiswinkel *et al.*, 2004; Baylis *et al.*, 1999). Field boundaries, however, still do not indicate the

farming practices, insecticides used, and management of animal dung, all of which can lead to an abundance of *Culicoides* (Baylis *et al.*, 1999). In contrast to the findings of Baylis *et al.* (1999), simplifying the ANN model by using only NDVI and LST was unsuccessful with the model having poor prediction capabilities. Although the resultant ANN model includes many variables, data for these are readily obtainable and available in electronic format, with the exception of *Culicoides* counts. Counts involve costly field work, a factor which may hamper further development of the model (Venter, 2008, personal communication). However, once the model is fully developed to include extreme minimum and maximum climate anomalies, further field work setting up traps to count *Culicoides* should not be necessary.

The ANN model described here does not take into account the effect of wind conditions on the abundance and distribution of *Culicoides* spp. This is a shortcoming also of the GIS models developed by Baylis *et al.* (1999) and Wittman *et al.* (2001). Prevailing winds influence the number of *Culicoides* caught in the traps: in strong winds the *Culicoides* may become stationary and fewer will be trapped. On the other hand, *Culicoides* can travel for long distances on prevailing winds and cause outbreaks of diseases in unsuspected areas. (Meiswinkel *et al.*, 2004) Another limitation of the ANN model, and also of the GIS model of Baylis *et al.* (1999), is the uneven distribution of *Culicoides* traps. There was a high density of traps in the Stellenbosch, Paarl, Malmesbury, George, Wellington districts, but no traps in the Prince Albert, Beaufort West, Murrayburg, VanRynsdorp, Vredendal and Clanwilliam districts. So, the ANN model is well-trained to predict *Culicoides* in abundance for districts where there are many traps but poorly trained for areas with fewer or no traps. If the actual counts of the Culicoides are evenly distributed throughout the study area it may be possible to predict the abundance of the Culicoides for the whole

study area and not just at specific points. The ANN model is not trained to do predictions in certain areas because there are no actual counts that can be used to train the ANN. One drawback of the combined use of a GIS and an ANN to predict the abundance of *Culicoides* spp. is that there is no direct interface between the GIS and the ANN: a high level of software knowledge and computer training is still required.

The use of ANN models to predict abundance of *Culicoides* spp. is important in the modelling of outbreaks of diseases carried by these insects. Such predictions can lessen the impact of the outbreak of vector-borne diseases by vaccinating animals at risk in time (Lord *et al.*, 2002). The ANN model was used successfully to predict the abundance of *C. imicola* and *C. bolitinos* for the year 2006 at sites where no counts were made. The model can also be used to predict the abundance of *Culicoides* for subsequent years provided that there are minor anomalies in the monthly temperature or rainfall.

16. Recommended Future Research

This project leads to many further research opportunities. These include firstly, the adaptation of the model to make not only monthly but also seasonal predictions of the abundance of *Culicoides* spp. Monthly predictions should be more accurate as these models include more detailed information but would involve the acquisition of many variables per month. The model can also be expanded to include other *Culicoides* species, whose possible role in the transmission of AHS cannot be ruled out (Lord *et al.*, 2002). Secondly, the model can also be expanded to predict outbreaks of AHS. Although the outbreaks of AHS correlate with the abundance of Culicoides certain climate variables also play an important role. The relationships among the outbreak of AHS, abundance of

Culicoides and the climate variables are also not known (Mellor *et al.*, 2004). Thirdly, the ANN model can also be expanded to include the prediction of outbreaks of other vector-borne diseases, such as bluetongue, epizootic haemorrhagic disease and equine encephalitis, which also are associated with abundance of *Culicoides* spp. The principle of combining a GIS and ANN could also be tested and applied to predict other diseases of which the exact relationships among the variables are not known. This study combines a GIS and ANN but a GIS can also be combined with other artificial intelligence systems e.g. decision trees or Markov models.

Chapter 5: Conclusion

This dissertation illustrated how a GIS and ANN can be combined for improved decision making in situations that combine spatial issues with significant complexity. A case study was conducted to illustrate the improved decision making capabilities obtained by utilising both a GIS and an ANN. In particular, the case study focused on predicting the abundance of *Culicoides* spp. in the Western Cape Province using a combination of a GIS and an ANN. This project was chosen because of its complex nature and the high number of predictors influencing the abundance of *Culicoides* spp. The presence of *Culicoides* spp. in abundance can lead to the outbreak of various vector-borne diseases, including African horse sickness. Such outbreaks can be avoided by identifying sites where abundance of *Culicoides* is likely to occur and vaccinating animals at risk. Accurate prediction of sites of probable abundance of *Culicoides* spp. has therefore great veterinary and economic benefits. Although GIS models have been developed which are able to predict abundance of *Culicoides* spp., there is still uncertainty as to which variables can be used as predictors and what the exact nature of the relationship among these variables is. In order to clarify this problem a combined GIS-ANN approach was employed.

Chapter 2 defined the field of study: namely GIS and ANN. The processes for the construction and implementation of these technologies were explained. GIS is still an evolving science while ANNs as a sub-field of computational or artificial intelligence is relatively well-developed. Chapter 3 captured the methodology used to illustrate the complementary use of GIS and ANN. It focused on the use of a GIS and ANN to predict the abundance of *C. imicola* and *C. bolitinos* in the Western Cape Province. The data sets

included in the development of the model were based on the GIS models developed respectively by Baylis *et al.* (1999) and by Wittman *et al.* (2000). Two models were developed and tested, and a best predictive model selected. Chapter 4 outlined the models that were developed and tested, as well as the model finally used for the prediction. The model with the highest accuracy and best prediction capability was chosen. The ANN model was subsequently used to predict the abundance of *Culicoides* at sites where counts were not done for certain months. These predicted values were then imported into the GIS and a classification map indicating the counted occurrence of *Culicoides* and predicted abundance was displayed. The predicted results were also compared with the actual counts of *Culicoides* spp. and outbreaks of AHS where available.

The research aims of the thesis were reached in the following ways:

1. Highlight the increasing role and integration of artificial intelligence within mainstream GIS applications.

   Through the theoretical development and specifically the validation with a specific case study, the potential and role of the integration of artificial intelligence techniques in mainstream GIS applications was illustrated. It is important to note that this study is largely indicative of the role and potential of this integrated approach and did not attempt to address the full spectrum of integration possible between GIS and artificial intelligence.

2. Evaluation of the process of integrating a GIS and ANN

A cursory study of published literature indicated that the integration of GIS and ANNs from a systems perspective is lacking. In most projects using GIS and ANNs the data was exported from the GIS to the ANN and the results from the ANN imported into the GIS. This essentially illustrates that the combination rather than the close integration of the two systems is the dominant use scenario. Significant opportunities exist to integrate rather than just combine the GIS and ANN.

Since not all GIS software include a direct interface between the GIS and ANN, a higher level of computer literacy is required for the complimentary use of the two systems than when tightly integrated. This stems from the need to deal effectively with data exchange as well as knowledge of two systems instead of one. The visualisation of the predictive result in the GIS makes the results understandable to a wider range of practitioners.

3. Demonstrate the integration of GIS and ANN by means of a case study

A GIS/ANN model was developed to predict the abundance of the two *Culicoides* spp. in the Western Cape province of South Africa. The case study has a strong spatial component illustrating the value of a GIS in decision making and allows for the broader use of the results through visualisation, making it more accessible. It features the integration of multiple data sources and has significant complexity without exact solutions, indicating the need for integration of GIS and ANN's

strengths. The developed model performed well with a prediction accuracy of 83% comparing favourable with GIS-only studies. These predicted values were then imported into the GIS and a classification map indicating the counted occurrence of *Culicoides* and predicted abundance was displayed. The predicted results were also compared with the actual counts of *Culicoides* spp. and outbreaks of AHS where available as further validation of the model.

4 Determine the variables influencing the occurrence of the Culicoides spp

Various ANN models were trained to determine the predictors influencing the abundance of *Culicoides* spp in the Western Cape province. The model with the best prediction capabilities was identified. The variables included in the model were NDVIs, LST, total rainfall, average rainfall, maximum rainfall, minimum rainfall, long-term rainfall, maximum of the maximum temperatures, average of the maximum temperatures, minimum of the maximum temperatures, long-term maximum temperatures, maximum of the minimum temperatures, average of the minimum temperatures, minimum of the minimum temperatures, long-term minimum temperatures, maximum humidity, minimum humidity, average humidity, livestock density, field boundaries, clay areas, rain anomalies, maximum temperature anomalies and minimum temperature anomalies.

5. Predict the abundance of Culicoides at trap points where counts were not made for the particular months during the study period.

The GIS/ANN model developed for the case study was subsequently used to predict the abundance of *Culicoides* spp. at trap points where no counts were made for particular months for the year 2006. The prediction of abundance were in the vicinity of actual counts of abundance of Culicoides or coincides with outbreaks of AHS (indicating abundance of *Culicoides* spp.). This is an indication of the good prediction capabilities of the model.

17. Assessing the Scientific Meaning of the Study

The scientific value of this study can be assessed by examining the research aims that were reached (presented above) and the contribution towards and the contribution from the various disciples used to conduct this study namely geoinformatics (geographical information systems and science), artificial intelligence and veterinary science.

Firstly, the study outlined the suitability of the complementary use of GIS and ANN for better decision making capabilities. The study indicates that GIS has good decision making capabilities and performs particularly well in cases where exact relationships among the variables impacting the issue are known. In cases where the relationships among the variables are unknown or the data are noisy, an ANN can be used in synenergy with the GIS to enhance the decision making capabilities of the system. The complementary use of GIS and ANN involves both the disciplines of geoinformatics and artificial intelligence.

Secondly, the study contributes towards the discipline of veterinary science. The model described in this study lays the foundation for the development of ANN models to predict the likelihood of outbreaks of other diseases where the relationship among the variables is

not known. Once the model is fully developed, the need for further costly field work will be reduced. The application potential of the model can be expanded since other diseases like Blue Tongue are carried by the same vectors and occur under the same circumstances. The study further contributes towards veterinary science since a cursory study indicated that there is no published attempt to predict the abundance of *Culicoides* spp. using a GIS/ANN model.

18. Final summation

The use of GIS and ANN to predict the occurrence of *Culicoides* in abundance has demonstrated successfully how techniques such as ANNs can assist GIS in decision making, especially where the datasets incorporate uncertainty or if the relationships between the variables are not known. The results of the study are encouraging and provide a rich set of scenarios for further research. Exploration of this juncture between exact GIS and non-parametric methods such as ANN provides significant scope for other applications and multi-disciplinary research.

References:

1. Abouzakhar, NS, Gani A, Manson G, Abuitbel M & King D 2003,' Bayesian Learning Networks Approach to Cybercrime Detection', Post Graduate Networking Conference.
2. All, J, Wulfe, A & Iovanna, A 2008, 'Using Geoinformatics to Examine Residential Radon Vulnerability' *Southeaster Geographer*, vol. 48, Issue 1, pp. 97-109.
3. Barredo, J 2007, 'Major flood disaster in Europe: 1950-2005.' *Natural Hazards*, vol. 42, Issue 1 pp. 125-148.
4. Baylis, M, Meisswinkel, R & Venter GJA 1999, 'Preliminary attempt to use climate data and satellite imagery to model the abundance and distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in southern Africa.', *Tydskrif van die Suid Afrikaanse Veterinere Vereningng*, vol. 70(2), pp. 80-89.
5. Baxt WG 1990, 'Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion.' *Neural Computing*, vol. 2, pp. 480-489.
6. Bédard Y, Gosselin P, Rivest S, Proulx MJ, Nadeau M, Lebel G, Gagnon MF 2003, 'Integrating GIS components with knowledge discovery technology for environmental health decision support.', *International Journal Of Medical Informatics*, vol. 70 (1), pp. 79-94.
7. Bessa-Gomes, C & Petrucci Fonseca, F 2003 'Using artificial neural networks to assess wolf distribution patterns in Portugal.', *Animal Conservation*, vol. 6, Issue Number 3, pp. 221-229.
8. Bolstad, P 2005, *GIS Fundamentals.* Second Edition, Eider Press.
9. Boone, I, Thys, E, Marcotty, T, de Borchgrave, J, Ducheyne, E & Dorny, P 2007 'Distribution and risk factors of bovine cysticercosis in Belgian dairy and mixed herds.' Preventive Beterinary Medicine vol. 82 page 1-11.
10. Bourlard, H & Morgan, N 1993, 'Continuous speech recognition by connectionist statistical methods.' *IEEE Transactions. Neural Networks*, vol. 4, pp. 893-909.
11. Breetzke, GD 2006, 'Geographical information systems (GIS) and policing in South Africa: a review.' *Policing*, vol. 29, Issue 4, p. 723.
12. Cavero, D, Tölle, KH, Henze, C, Buxadé, C & Krieter, J 2008, 'Mastitis detection in dairy cows by application of neural networks.', *Livestock Science*, vol. 114, Issues 2-3, pp. 280-286.
13. Craigie, D 2008, 'Information Integration: A GIS perspective', *Ecological Circuits*, Sept/Oct 2008, pp. 14-19.
14. Davis, B 2001, *GIS: A Visual Approach*, 2nd edition, Thomson Delmar Learning.
15. Deadman, PJ & Gimblett, HR 1997, 'Applying neural networks to vegetation management plan development.', *AI Applications: Natural Resources, Agriculture & Environmental Science*, vol. 11 Issue 3, P. 107.
16. Deadman, P & Gimblett, R 1997, 'Merging Technologies: Linking Artificial Neural Networks to Geographic Information Systems for Landscape Research and Education.' Retrieved April 2008, http://www.srnr.arizona.edu/~gimblett/cela95.html.

17. Demers, MN 2000, *Fundamentals of Geographic Information Systems, Second Edition*, John Wiley and Sons.
18. Durr, P & Martin, W 2005. *The GISVET'04 Special Edition.* Available Online: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TBK-4H16NT6-1&_user=59388&_coverDate=10%2F12%2F2005&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000005298&_version=1&_urlVersion=0&_userid=59388&md5=4668d52af1190641621395a962aba80f
19. Garzón, MB, Blazek, R, Neteler, M, Dios, R, Sanchez, D, Helios, S, Furlanello, C 2006, 'Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L. in the Iberian Peninsula.', *Ecological Modelling*, vol. 197 Issue 3/4, pp. 383-393.
20. Genchi, C, Rinaldi, L, Cascone, C, Martarino, M & Cringoli, G, 2005, 'Is heartworm disease really spreading in Europe?' *Veterinary Parasitology* vol. 133, issues 3-4, pages 137-148.
21. German, G, Gahegan, M & West, G August 1997, 'Predictive Assessment of Neural Network Classifiers For Applications in GIS.', Paper Presented at Second Annual Conference of GeoComputaion '97 $ SIRC '97, Retrieved March 2008 http://www.geocomputation.org/1997/papers/german.pdf.
22. Graham, TE & Goswami I 2001, 'Baltimore's Urban Environment Using GIS and Neural Networks.' Presented at ESRI User Conference 2001, Retrieved March 2008 http://gis.esri.com/library/userconf/proc01/professional/papers/pap699/p699.htm.
23. Guyon, I 1991, 'Applications of neural networks to character recognition.', *Int. J. Pattern Recognit, Artif. Intell*, vol. 5 pp. 353-382.
24. Ha, SR, Park, SY, Park, DH 2003, 'Estimation of urban runoff and water quality using remote sensing and artificial intelligence.', *Water Science and Technology*, vol. 47 (7-8), pp. 319-325.
25. Heald, CW, Kim, T, Sischo, WM, Cooper, JB & Wolfgang, DR 2000, 'A computerized mastitis decision aid using farm based record: An artificial neural network approach.', *Journal of dairy science*, vol. 83, pp. 711-720 .
26. Hewitson, B C & Crane, RC 1994, *Neural Nets: Applications in Geography*, Kluwer Academic Publishers.
27. Jiang, Y & Nan, Z 2006, 'Integration of Artificial Neural Network with GIS in Uncertain Model of River Water Quality.' Retrieved March 2008, http://www.geocomputation.org/1997/papers/german.pdf.
28. Kalipeni, E & Zulu, L 2008, 'Using GIS to Model and Forecast HIV/AIDS Rates in Africa, 1986-2010.' *The Professional Geographer*, vol. 60, Issue 1, pp. 33-53.
29. Lacher RC, Coats PK, Sharma SC & Fant LF 1995, 'A neural network for classifying the financial health of a firm.', *Eur. J, Oper. Res.*, vol. 85, pp. 53-65.

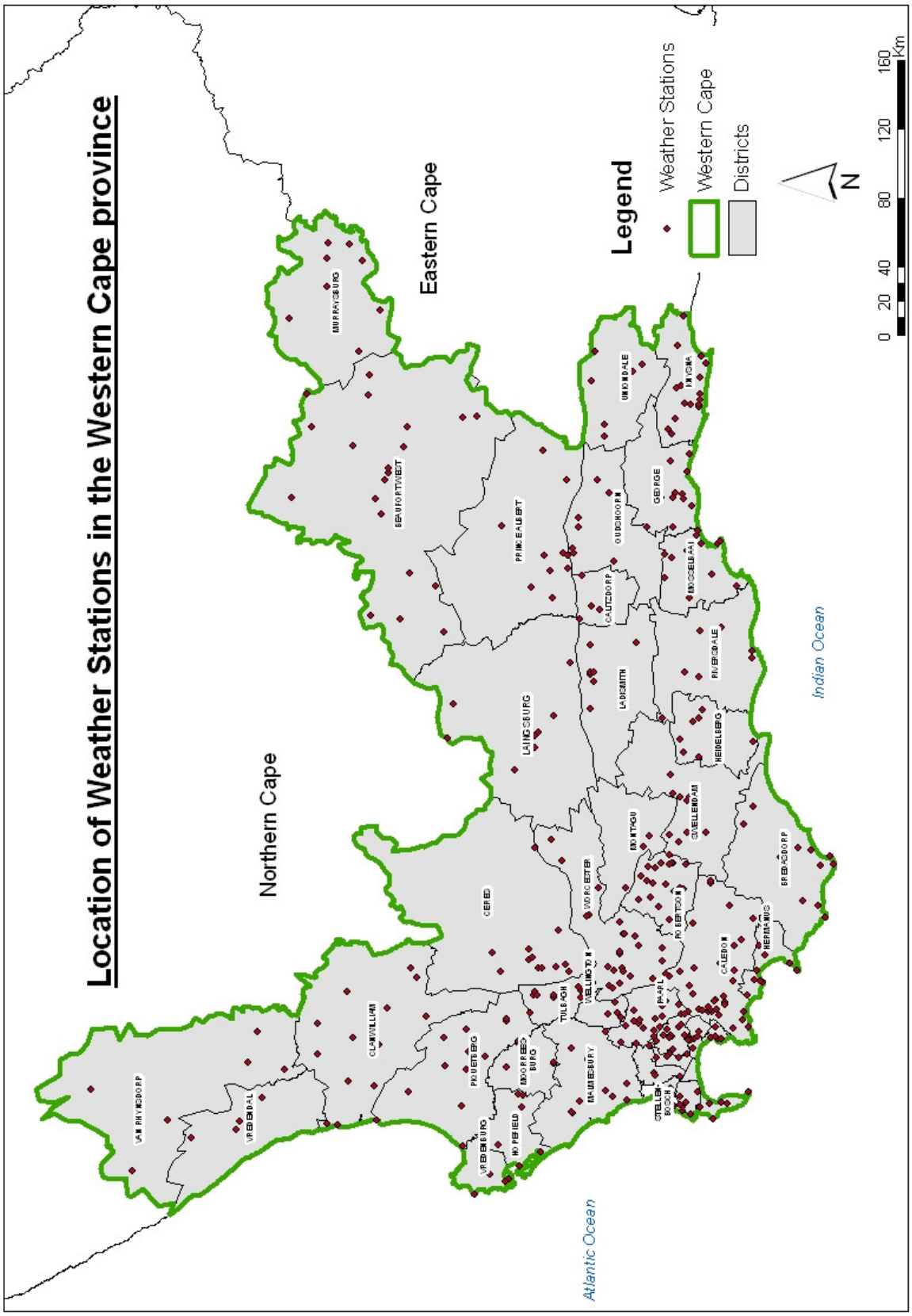
30. Lee, S, Ryu, J, Min K, & Won, J 2003, 'Landslide Susceptibility Analysis Using GIS and Artificial Neural Network.', *Earth Science Processes and Landforms*, vol. 28, pp. 1361 – 1376.
31. Lord, CC, Venter, GJ, Mellor, PS, Paweska, JT & Woolhouse, MEJ 2002, 'Transmission patterns of African horse sickness and equine encephalosis in South African donkeys.' *Epidemiology and Infections. Infect* **128**, pp. 265-275.
32. Lusk, JJ, Guthery, FS, George, RR, Peterson, MJ & DeMaso, SJ 2002, 'Relative abundance of bobwhites in relation to weather and land use.', *Journal of Wildlife Management*, vol. 66(4), pp. 1040-1052.
33. Mas, JF, Puig, H, Palacio, JL & Sosa-Lopez, A 2003, 'Modelling deforestation using GIS and artificial neural networks', *Environmental Modelling & Software*, vol 19, Issue 5, pp. 461-471.
34. Mccloy, KR 2006, *Resource Management Information systems: Remote Sensing, GIS and Modelling*, Second Edition. CRC Press
35. McDonald, AT & Foster JA 2000, 'Assessing pollution risks to water supply intakes using geographical information systems (GIS).', *Environmental Modelling & Software*, vol. 15, Issue 3, pp. 225.
36. Meiswinkel R, Venter GJ & Nevill EM 2004, 'Vectors: *Culicoides* spp.', *Infectious diseases of livestock*, vol. One, pp 93-136.
37. Mellor PS & Hamblin C 2004, 'African Horse Sickness', *Veterinary Research*, vol. 35 pp. 445-466.
38. Muttiah, R, Srinivasan, R & Engel B 'Development and Application of Neural Network Interface for GRASS GIS' Retrieved July 2009, http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/muttiah_ranjan/muttiah.html
39. Mlisa, A, Africa,U & van Niekerk A 2008, 'GIS in the decision-making process' *Position IT*, Sept/Oct 2008, pp,44-48.
40. Ni JR & Xue A 2003, 'Application of Artificial Neural Network to the Rapid Feedback of Potential Ecological Risk in Flood Diversion Zone.' *Engineering Applications of Artificial Intelligence*, vol. 16, Issue 2, pp.105-119.
41. Ochola, WO & Kerkides, P 2004, 'An integrated indicator-based spatial decision support system for land quality assessment in Kenya.', *Computers & Electronics in Agriculture*, vol. 45, issue 1-3, p. 24.
42. Pastell, ME & Kujala, M 2007, 'A Probabilistic Neural Network Model for Lameness Detection', *Journal of Dairy Science*, vol. 90, pp. 2283-2292
43. Perdiguero-Alonso, D, Montero, F, Kostadinova, A, Raga, JA & Barrett, J 2008, 'Random forests, a noel approach for discrimination of fish populations using parasites as biological tags.', *International Journal of Parasitology*, vol. 38, issue 12, pp. 1425-1434.
44. Pijanowski, BC, Brown, DG, Shellito, BA & Manik GA 2002, 'Using neural networks and GIS to forecast land use changes: a Land Transformation Model' , *Computers, Environment and Urban Systems*, vol. 26, Issue 6, pp. 553-575.
45. Pijanowski, BC, Shellito, BA, Bauer, ME, & Sawaya, K.E 2001, 'Using GIS, Artificial Neural Networks and Remote Sensing to Model Urban Change in

the Minneapolis-St. Paul and Detroit Metropolitan Areas.' ASPRS Proceedings 2001, Retrieved April 2008 http://64.233.183.104/search?q=cache:Pi68rqeOkqAJ:hussein.atoui.googlepages.com/Pijanowskietal2001USINGGISARTIFICIAL.pdf+Using+GIS,+Artificial+Neural+Networks+and+Remote+Sensing+to+Model+Urban+Change+in+the+Minneapolis-St.+Paul+and+Detroit+Metropolitan+Areas&hl=en&ct=clnk&cd=1&gl=za.

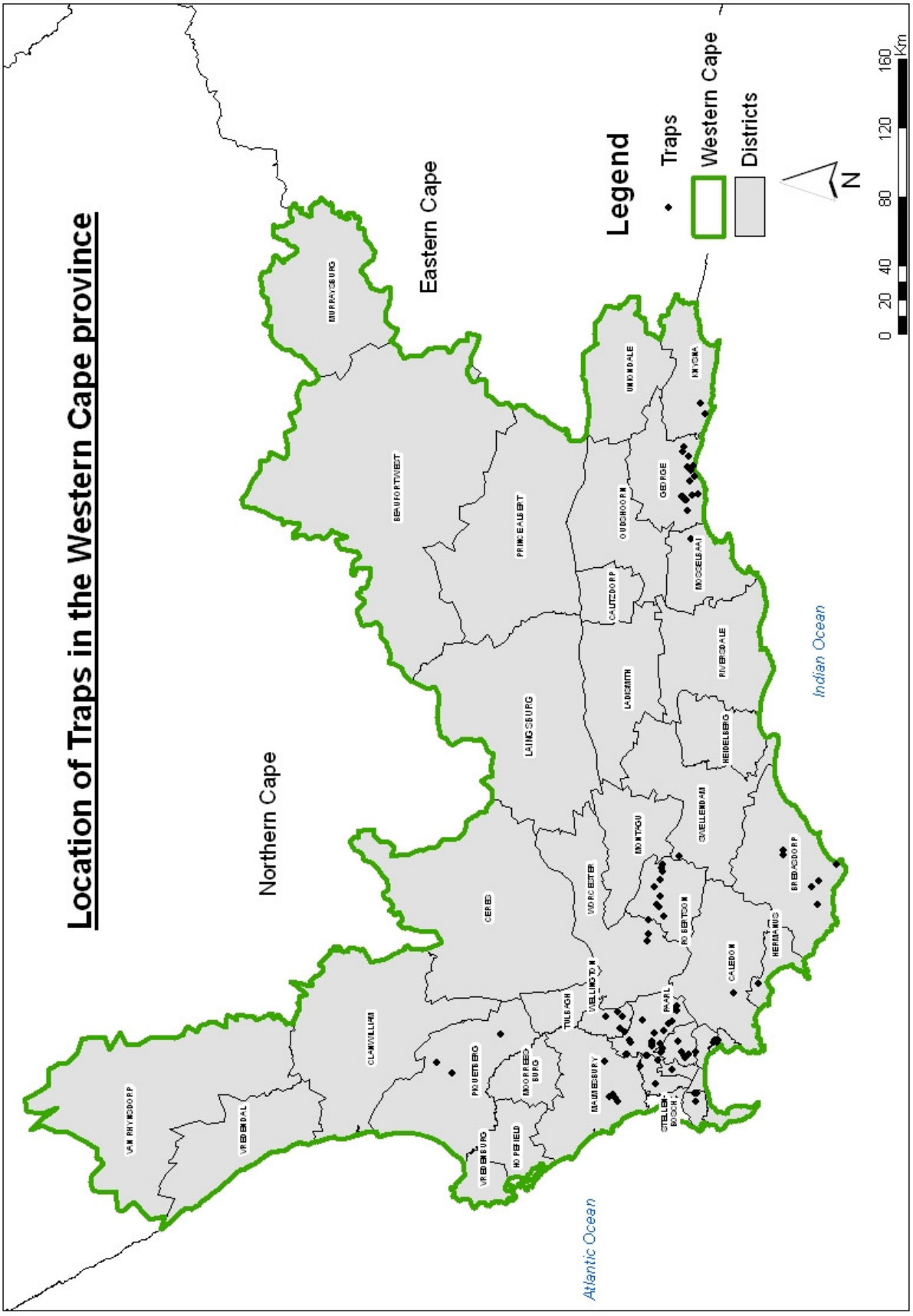
46. Pope, R, Anderson, B & Rickel, BW 1998, 'Using fuzzy systems, object – orientated programming and GIS to evaluate wildlife habitat.', *AI Applications: Natural Resources, Agriculture & Environmental Science*, vol. 12, issue 1-3, p. 31.
47. Pradan, B, Mansor, S, Lee, S & Buchroithner, MF 2008, Application of a data mining model for landslide hazard mapping', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVII, Part B8.
48. Ramirez, A 2004 'Geographic Information Systems and its Role in Biological Risk Management.' Iowa State University, http://www.cfsph.iastate.edu/BRM/resources/General/GeographicInformationSystesRoleBRM_Sept2004.pdf
49. Rich, E, 1983, *Artificial Intelligence*., McGraw Hill.
50. Rigol-Shanchez, JP, Chica-Olmo, M & Abarca-Hernandez, F 2002, 'Artificial neural networks as a tool for mineral potential mapping with GIS', *International Journal of Remote Sensing*, vol. 24, no.5, pp 1151-1156.
51. Rogers, DJ & Williams, BG 1993, 'Monitoring trypanosomiasis in space and time.', *Parasitology*, vol. 106, pp. S77-92
52. Roush WB, Cravener TL, Kirby YK & Wideman RF (Jr) 1997, 'Probabilistic Neural Network Prediction of Ascites in Broilers Based on Minimally Invasive Physiological Factors', *Poultry Science*, vol. 76, pp. 1513-1516
53. Roush, WB, Wideman, RF (Jr), Cahaner ,A, Deeb, N & Cravener, TL 2001, 'Minimal number of chicken daily growth velocities for artificial neural network detection of pulmonary hypertension syndrome (PHS)', *Poultry Science*, vol. 80, issue 3, pp. 254-259.
54. Russel, SJ 1995. *Artificial Intelligence: A modern Approach.* Prentice Hall.
55. Saha, A 2003 'Introduction to Artificial Neural Network Models' Retrieved March 2008, http://www.geocities.com/adotsaha/NNinExcel.html.
56. Sarip, AG 2005, 'Integrating Artificial Neural Networks and GIS for Single-Property Valuation.' Proceedings of the PRRES Conference 2005, Retrieve April 2008 http://www.prres.net/papers/Sarip_Integrating_Artificial_Neural_Networks_And.Pdf.
57. Statsoft, Retrieved June 2008, http://www.statsoft.com/textbook/stathome.html
58. Stergiou ,C & Siganos D 1996, 'Neural Networks', Retrieved March 2008 http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

59. Thurston, J January 2002, 'GIS & Artificial Networks: Does Your GIS Think?', Retrieved March 2008, http://www.integralgis.com/pdf/Neural%20Networks.pdf.
60. Vafeidis, AT, Koukoulas, S, Gatsis, S & Gkoltsiou, K 2007, 'Forecasting lan-use changes with the use of neural networks and GIS', *Geoscience and Remote Sensing Symposium*, 23- 28 July 2007, pp. 5068-5071.
61. Van Helden, P 2005, Introductory GIS – Class Notes, University of Pretoria.
62. Venter GJ 2008, Senior Researcher, Onderstepoort Veterinary Institute , Agricultural Research Council.
63. Walker, R & Craighead, L 1997, 'Analysing Wildlife Movement Corridors in Montana Using GIS', Paper Presented at ESRI User Conference 1997, Retrieved September 2008.
64. Wang, F 1994, 'The Use of Artificial Neural Networks in a Geographical Information System for Agricultural Land-Suitability Assessment', *Environment and Planning*, vol. A 26(2), pp. 265-284
65. Ward, DG, Wenbin, W, Yaping, C, Bellingham, LJ, Marin, A, Johnson, P, Lyons, B, Feist, SW & Stentiford, GD 2006, 'Plasma Proteome Analysis Reveals the Geographical Origin and Liver Tumor Status of Dab (Limanda limanda) from UK Marine Waters.', *Environmental Science & Technology*, vol. 40, issue 12, pp. 4031-4036.
66. Winston, PH, 1992, *Artificial Intelligence*. Addison-Wesley Pub.
67. Wittmann, EJ, Mellor PS & Baylis, M 2001, 'Using climate data to map the potential distribution of *culicoides imicola* (Diptera Ceratopogonidae) in Europe.' *Revue scienfique et technique, Office International des Epizooties* vol. 20, pp. 731-740.
68. Wu, C, Ayers, PD & Anderson, AB 2004, 'GIS and Neural Network Methods for Potential Road Location for Military Land Management'. Paper number 041148, ASAE Annual Meeting.
69. Wu, Q, Xu, H & Pang,W 2007, 'GIS and ANN coupling model: and innovative approach to evaluate vulnerability of karst water inrush in coalmines of north China', *Environ Geol*, vol. 54, pp. 937-943.
70. Yang, Y & Rosenbaum, M 2001 'Artificial Neural Networks linked to GIS for Determining Sedimentology in Harbors.' *Journal of Petroleum Science and Engineering*, vol. 29, pp. 213-220.
71. Yoo, C & Kim, J 2007, ' Tunneling performance prediction using an integrated GIS and neural network', *Computers and Geotechnics*, vol. 34 Issue 1, pp. 19 -30.
72. Zhang, P 2000, 'Neural Networks for Classification: A Survey', *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, vol. 30, pp. 451-462.

Annexure 1:


Location of Weather Stations in the Western Cape province
Northern Cape
Eastern Cape
Indian Ocean
Atlantic Ocean
Legend
Weather Stations
Western Cape
Districts
N
0 20 40 80 120 160 Km

Annexure 2:


Location of Traps in the Western Cape province
Northern Cape
Eastern Cape
Legend
Traps
Western Cape
Districts
N
0 20 40 80 120 160 Km
Indian Ocean
Atlantic Ocean

Annexure 3: List of Traps summarised per year (2006)

Name of Trap	Total	Average	Frequency
Botrivier, Klein Paradys	2387.52746	596.881865	4
Cape Town, Hyjo	13	3.25	4
Cape Town, Milnerton Race Course	2	0.166665	4
Cape Town, Philippi training centre	0	0	3
Cape Town, Philippi, Golden Grove Farm	0	0	1
Cape Town, Philippi, Kings Kraal	0	0	1
George, Anne & Caprice	14	14	1
George, Barnyard	147	36.75	4
George, EL-BE-AR ranch	409	409	1
George, Fancourt	1813.22432	233.27804	4
George, Farmlands	3061.93981	3061.93981	1
George, Forest Hill Farm	743	122.8	5
George, George Riding Club	863	147.2	5
George, Hoekwil, Bajaanskloof	3429.23596	1143.078653	3
George, Hoekwil, Kiddbuddie	2174.35208	56.70491	10
George, Hoekwil, Kingfisher Lodge	48	48	1
George, Morning Star	1533.72561	383.4314025	4
George, Oakhurst	27	27	1
George, Outeniqua miniture horses	30013.99889	3334.88877	1
George, Perdepoort	1966.20475	393.24095	5
George, Rocky Mountain Trails	118	59	1
George, Tarentaalbos	343	343	1
George, The Ark	2663.02909	665.7572725	4
George, Victoria Bay, Carmel Equestrian C	580	290	2
George, Wilderness Adventures	1918.26441	304.1274017	6
George, Wilderness, Hirschberg	379	379	1
Gordens Bay, Broadlands	164871.2814	3673.698175	11
Gordens Bay, Cindy's Livery	513	513	1
Gordens Bay, Firlands Equestrian Centre T	905	452.5	1
Gordens Bay, Hillside	1168.5	233.7	5
Gordens Bay, Stellentia	494	177.25	2
Hermanus, High Seasons	181	45.25	4
Knysna, Horse rescue	3446.55552	136.48611	4
Piketberg, (Porterville) Rietvlei Noord	1726.39896	575.46632	3
Piketberg, Metonshoek stud	53468.35912	2593.211016	8
Piketberg, Smitsvlei	1457.15705	364.2892625	4
Piketberg, Wilgerbosdrift	114813.7725	2542.253059	12
Robertson, Alchemy	579	193	3
Robertson, Alfalfa Dairy calves	3090.77077	975.534055	2
Robertson, Dageraad	36451.26379	1355.830848	4
Robertson, Galloway	569	569	1
Robertson, Highlands stables	1911.5082	955.7541	2
Robertson, Highlands Stallions	147	147	1
Robertson, Highlands yearlings	3953.36044	1976.68022	2
Robertson, Litchfield	1112	224	3
Robertson, Maine Chance	45439.82876	2765.014523	8
Robertson, Normandy	321	160.5	2
Robertson, Riverton	2377.46018	792.4867267	3

Robertson, Rondeheuwel	3509.09792	600.1829867	3
Robertson, Showgrounds	12257.18024	4085.726747	3
Robertson, Zandvliet	379	189.5	2
Stellenbosch, Animal Zone	227	227	1
Stellenbosch, Arc en Ciel	2117.08463	705.6948767	3
Stellenbosch, Avontuur	70	3.413888333	6
Stellenbosch, Beaumont Stud	2654.01049	459.0017467	3
Stellenbosch, Bona Vista	1245	284.875	4
Stellenbosch, Brakenfell, Connemara	35479.67933	3026.580925	4
Stellenbosch, Briza kennels and cattery	5554.47764	819.9417467	6
Stellenbosch, Del Vera	337	36.25	3
Stellenbosch, El Dorado	3528.32948	882.08237	4
Stellenbosch, Elsenburg	1730	88.375	8
Stellenbosch, Fairyhouse	112	28	4
Stellenbosch, Highflyer	352	88	4
Stellenbosch, Inca Vale	3017.5794	1005.8598	3
Stellenbosch, Juhanta	12234.73953	3058.684883	4
Stellenbosch, Klein Optenhorst	46091.29942	15363.76647	3
Stellenbosch, La Pitite Rochelle	503.22649	167.7421633	3
Stellenbosch, L'Auberge Rosendal	31839.65344	7959.91336	4
Stellenbosch, Linquenda	513718.0292	20627.42355	9
Stellenbosch, L'ormarins	147	36.75	4
Stellenbosch, Moseoatuania	1613.74967	537.9165567	3
Stellenbosch, Mr Burger	112	112	1
Stellenbosch, Natte Vallei	99	24.75	4
Stellenbosch, Pine Ranch	1803	450.75	4
Stellenbosch, Reinel stud	1126	375.3333333	3
Stellenbosch, Rhodes Food Group	332	83	4
Stellenbosch, Riverworld	654	163.5	4
Stellenbosch, Robertsvallei	1139	37.25555556	9
Stellenbosch, Sorento	90	45	2
Stellenbosch, Spieka	0	0	1
Stellenbosch, Spier	1325	1325	1
Stellenbosch, Steadfast	345	86.25	4
Stellenbosch, Steinmetz	614.64811	204.8827033	3
Stellenbosch, Tatooine Lane	379	94.75	4
Stellenbosch, Trough End	600429.3108	21825.32092	8
Stellenbosch, Tygerberg Zoo	660	310.5	2
Stellenbosch, Varsfontein	1010.93686	252.734215	4
Stellenbosch, Vredenheim	6625.80702	331.465351	10
Stellenbosch, Welgemeend	7	7	1
stellenbosch, Wellington, Afton Grange	4331.98208	4331.98208	1
Stellenbosch, Wellington, Arc-en-Ciel	2917.45574	2917.45574	1
Stellenbosch, Wellington, Oaklands	29	29	1
Stellenbosch, Woodhill racing stud	19178.29825	2649.402175	4
Struisbaai, Elandsdrift	297.09548	297.09548	1
Struisbaai, Kraskalk, Dohne sheep	10	10	1
Struisbaai, Melkhout	103	103	1
Struisbaai, Morning Glory Beach Rides	0	0	1
Struisbaai, Vlooikraal	963.33333	963.33333	1
Struisbaai, Wiesdrift	51	51	1