

THE USE OF ARTIFICIAL NEURAL NETWORKS TO ASSIST GIS IN DECISION MAKING: A CASE STUDY – PREDICTING THE OCCURRENCE OF *CULICOIDES* CAUSING AFRICAN HORSE SICKNESS

Sanet Eksteen
Department of Geography
Geoinformatics and Meteorology
University of Pretoria
Pretoria, 0002
sanet.eksteen@up.ac.za

Abstract

African Horse Sickness (AHS) is endemic to sub-Saharan Africa. This disease 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. This informs the need for models that can accurately predict potential outbreaks to prevent unnecessary horse deaths.

To this end, various GIS models have been built in an attempt to predict the *occurrences* of the *Culicoides*. Although these models had some success in the prediction of possible outbreaks, the exact relationships between the different variables could not be determined. This led 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 occurrence of *C. Imicola* and *C. Bolitinos*. In addition, using maps to visualise the predictions will make the models more accessible to a wider range of practitioners.

1. Introduction

African Horse Sickness (AHS) is endemic to sub-Saharan Africa. This disease is carried by two midges called *Culicoides imicola* and *Culicoides bolitinos*. Outbreaks of AHS are predicated upon the abundance of these two midges. The disease causes severe illness in horses and has significant economic impact if not dealt with timeously. This informs the need for models that can accurately predict potential outbreaks to prevent unnecessary horse deaths.

In order to address this need, various GIS models have been built in an attempt to predict the abundance of the *Culicoides*. Although these models had some success in the prediction of abundance of the *Culicoides*, the exact relationships between the different variables affecting the occurrence of the *Culicoides* could not be determined. This led to the current study that uses a combination of a GIS and an

artificial neural network (ANN) to develop a model that predict the abundance of *C. Imicola* and *C. Bolitinos*

2. Research in Context

Recent advances in computer technology and associated applications are allowing decision makers to deal with increasing levels of complexity in challenges that many times are either not well understood or deals with so many dimensions and uncertainties that limits the effectiveness of exact methods. By implication, this is also true in the case of GIS, a key decision support tool (Mlisa *et al* 2008).

It is generally accepted that the effectiveness of GIS in decision making is dependent on the appropriate integration of the key components of a GIS namely: people, data, software, hardware and tools and analysis (Craigie, 2008). With the major influence of the application area on the decisions concerning all five GIS components, Van Helden (2005) argued that the application area should be added as a sixth component.

The research for this paper is situated within the same framework. The primary focus of the study is on the GIS components of tools and analysis, software and data. An application with the following characteristics was selected:

- 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 (people);
- integration multiple data sources with various levels of accuracy and quality to focus on the data component (data); and
- significant complexity without exact solutions indicating the need for integrations with non-parametric methods complementing the exact methods typically found in GIS (tools and analysis, software).

A number of applications displaying these characteristics are found in areas such as water quality management and ecology (Jiang *et al*, 2008, Bessa-Gomes *et al*, 2003, Lusk *et al*, 2002). Given the importance of agriculture in the South African economy, an application displaying these characteristics from within the veterinary science was selected. One such problem that is very complex with a strong spatial component is the prediction of abundance of *Culicoides* spp. that causes AHS.

The research focuses primarily on the GIS components. The non-parametric or artificial intelligence part of the research is primarily aimed to demonstrate the viability of the integration of the GIS and AI and is not optimised per se.

3. Objectives

The objective of this study is to develop a GIS incorporating an artificial neural network model that can predict the abundance *C. imicola* and *C. bolitinos* in the Western Cape Province of South Africa.

The secondary objectives of the study are:

- to demonstrate how artificial neural networks can assist or enhance GIS as a decision making tool;
- to determine the variables influencing the occurrence of the *Culicoides*; and
- to determine the correlation between the variables and the occurrence of the *Culicoides*.

4. What is an Artificial Neural Network

An artificial neural network (ANN) is a type of artificial intelligence technique based on how the human brain functions (McCloy, 2006). Lately, 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 inherited two important characteristics of the 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 are a promising alternative to conventional classifiers (Zhang, 2000).

ANNs, like the brain also consists of neurons, called processors (nodes), 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. The most generalised type of ANN consists of three separate layers: an input layer, a hidden layer, and an output layer.

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 (inhibitory effect) or positive (excitatory effect) (Statsoft, 2008). 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.

The training of an ANN has no fixed strategy, most of the training being done through trial and error (Saha, 2003). Training starts with a definition of the number of input

neurons, the number of hidden layers and the number of outputs (Saha, 2003). The purpose of the training of a network is to use these layers to minimise the output error. When training an ANN, the algorithm progresses iteratively through a number of epochs – in ANN terms 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 artificial neural network 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 backpropagation (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.

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).

Once the network has been trained and tested on the test set, the model can be used as a predictive model. When using an ANN for prediction one needs to take care to correctly interpret the output generated.

5. Why use Artificial Neural Networks?

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. (If the exact nature of the relationship is known, it can be modelled directly using a correlation model (Statsoft, 2008)). 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). Further, 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).

From the current understanding of the problem of occurrence of *C. imicola* and *C. bolitinos*, it is a problem well suited to solving with an ANN. Software implementing ANNs are readily available, allowing the research to concentrate on the GIS components and the integration effort instead of focusing on the implementation of the ANN itself.

6. Case Study: Methodology

This study focuses on the occurrence of *C. imicola* and *C. bolitinos* in the Western Cape Province of South Africa, historically an AHS-free zone even though the vector species occur naturally in the area. However, since the first recorded outbreak of AHS in this province, in Stellenbosch in 1999, there have been further outbreaks, 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).

The data obtained and included in the model is based on GIS models previously developed and covers the Western Cape Province of South Africa for the time period December 2005 to December 2006. All the data sets were imported into the GIS and extracted on a monthly basis for the time period covered.

After incorporation into the GIS, the data sets were extracted for use in the ANN. *Culicoides* spp. capture sites were used as extraction points. Raster values for each count (or trap) were extracted for each layer and combined in a spreadsheet containing raster values for all the GIS layers per month. A data extraction model was developed in the GIS to assist with extracting the data. The extracted data was combined in a spreadsheet and used for the training of the ANN. In total 337 traps had been counted for the year 2006. The predicted monthly average count of the *Culicoides* is used as the output variable for the ANN.

Various combinations of different variables were used in the training of the ANN to develop a model with the highest degree of accuracy in predicting the abundance *C. imicola* and *C. bolitinos*. After training the network the ANN is tested to determine the best predictive model. The selected model is used to predict the abundance of the *Culicoides* at unsampled sites. (The Department of Agriculture relies on the traps being set up by farmers and since this was erratic, counts for most months were incomplete.) The results from the ANN predictions are imported back in to the GIS software and a classification map to indicate the abundance of the *Culicoides* in is created.

7. Results

The variables included in the ANN model are similar to previous developed GIS models. Although these models have found rainfall not to be a significant predictor, the ANN model performed better with the inclusion of rainfall as a predictor. The overall accuracy of the ANN model of 83% is roughly similar to the previously developed GIS models. It is important to note that the artificial intelligence component of the model was not specifically optimised, leaving room for further potential increases in accuracy. The variables that display the most significant influence on the occurrence of the *Culicoides* are determined by the ANN. The ANN model 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. These predictions were then mapped against actual recorded AHS outbreaks and recorded abundance of *Culicoides* as benchmark.

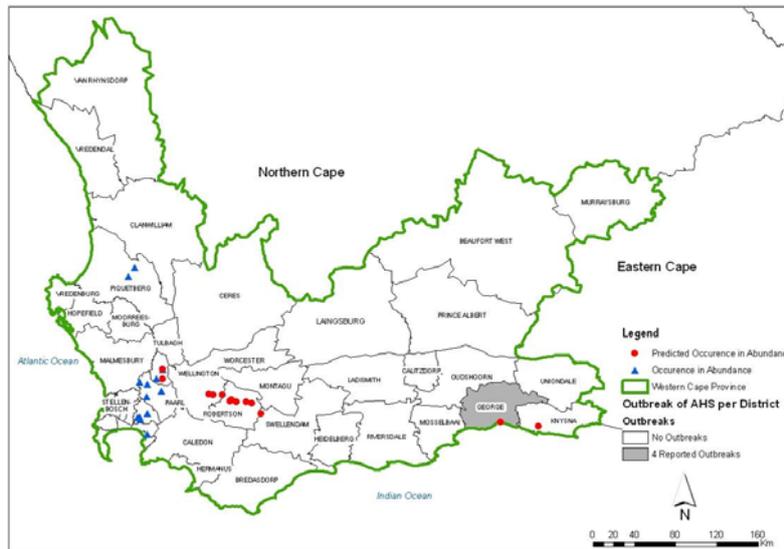


Figure 1: Predicted Abundance of *C. imicola* and *C. bolitinos*: January 2006

The predicted abundance of *C. imicola* and *C. bolitinos* for January 2006 (Fig.1) for the George district coincided with an outbreak of AHS.

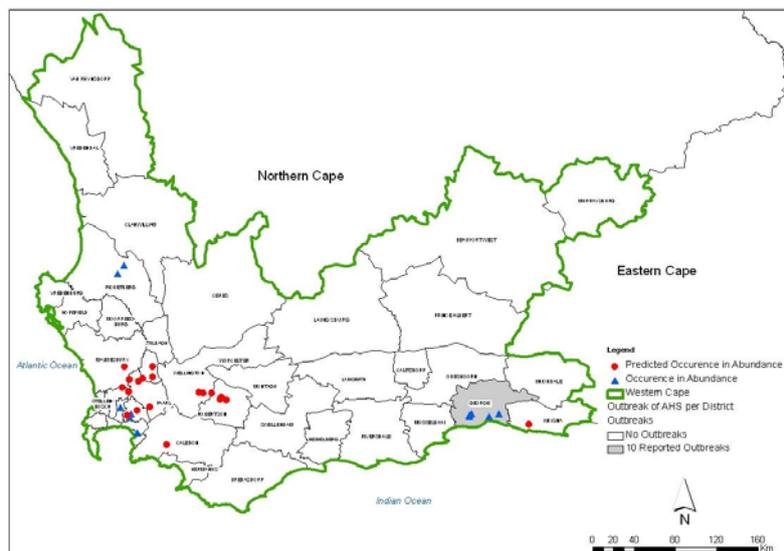


Figure 2: Predicted Abundance of *C. imicola* and *C. bolitinos*: February 2006

The abundance of *Culicoides* predicted by the ANN model for the Stellenbosch district for February (Fig. 2) and March 2006 (Fig. 3) coincided with the actual count. For the

George district for both February and March 2006 (Figs.2 and 3), and for the Robertson district for March 2006 (Fig. 3) the predicted abundance coincided with an outbreak of AHS.

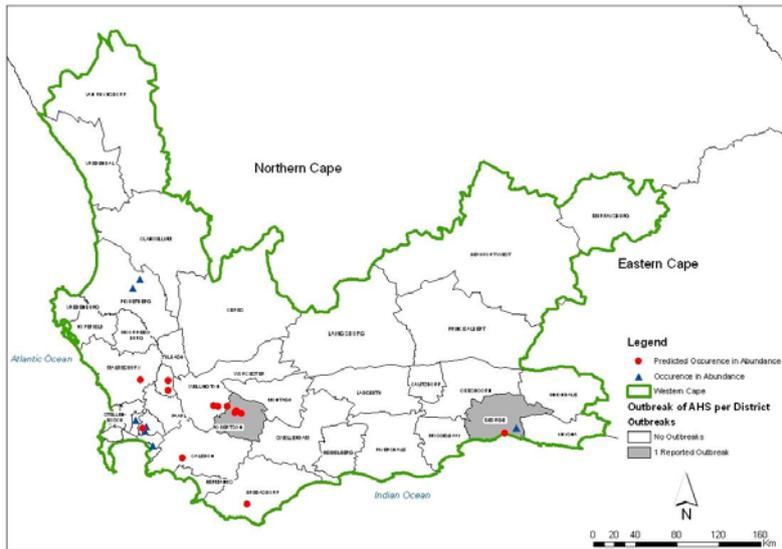


Figure 3: Predicted Abundance of *C. imicola* and *C. bolitinos*: March 2006

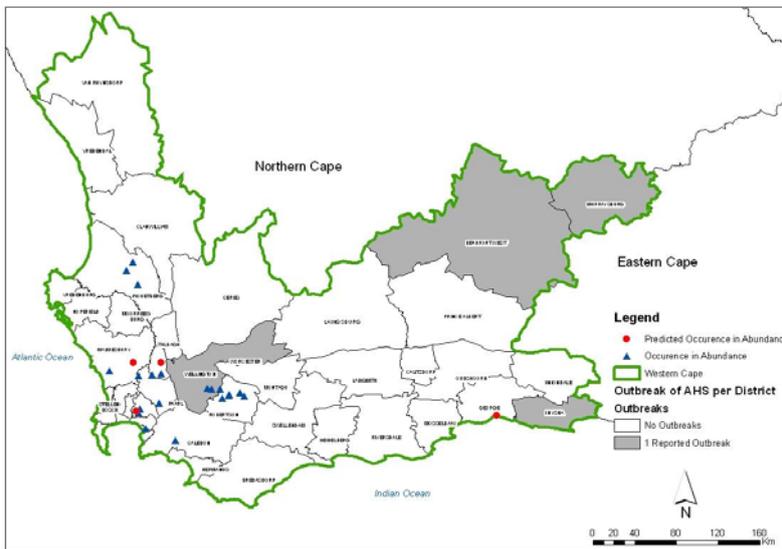
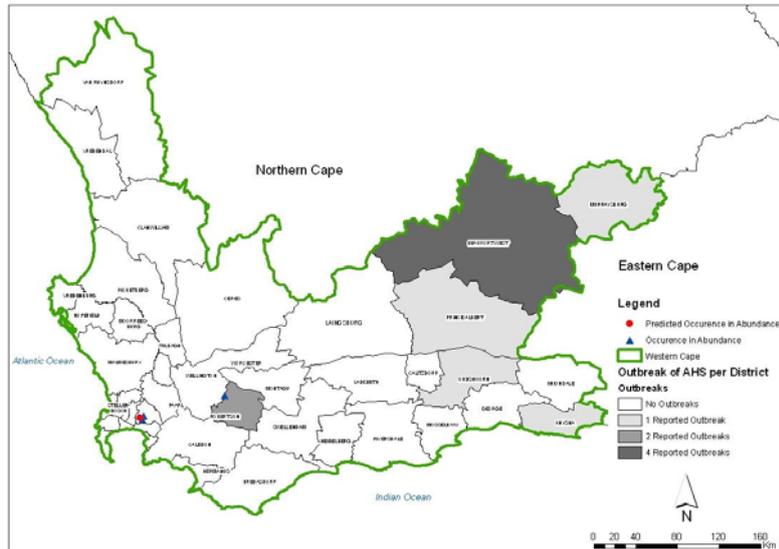


Figure 4: Predicted Abundance of *C. imicola* and *C. bolitinos*: April 2006

There was an outbreak of AHS in the Murraysburg and Beaufort West districts (Fig.4) in April 2006, and in the Murraysburg, Beaufort West and Oudtshoorn districts in May 2006 (Fig.5). Since these districts were underrepresented in the ANN model, no predictions of the abundance of *Culicoides* in these districts were made. 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 5: Predicted Abundance of *C. imicola* and *C. bolitinos*: May 2006

There were no *Culicoides* counts for June, July and October 2006. The ANN model predicted a zero probability of abundance of *Culicoides* for these months, 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*.

8 Conclusions

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.

Even though the AI component was not optimised in the specific case study, the results are encouraging. Exploration of this juncture between exact GIS and non-parametric methods such as pattern recognition provides rich scope for other applications and multi-disciplinary research.

Although AHS does not occur on other continents, similar diseases like Blue tongue disease are carried by the same vectors and will occur under the same circumstances, providing a broader potential application of the developed model.

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. 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.
3. Bolstad, P 2005, *GIS Fundamentals*. Second Edition, Eider Press.
4. Craigie, D 2008, 'Information Integration: A GIS perspective', *Ecological Circuits*, Sept/Oct 2008, pp. 14-19.
5. 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.snr.arizona.edu/~gimblett/cela95.html>.
6. 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 GeoComputation '97 \$ SIRC '97, Retrieved March 2008 <http://www.geocomputation.org/1997/papers/german.pdf>.
7. Hewitson, B C & Crane, RC 1994, *Neural Nets: Applications in Geography*, Kluwer Academic Publishers.
8. 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>.
9. Lord, CC, Venter, GJ, Mellor, PS, Paweska, JT and Woolhouse MEJ 2002, 'Transmission patterns of African horse sickness and equine encephalosis in South African donkeys.', *Epidemiology and Infections*. *Infect* **128**, pp. 265-275.
10. 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.
11. McCloy, KR 2006, *Resource Management Information systems: Remote Sensing, GIS and Modelling*, Second Edition. CRC Press.
12. Mlisa, A, Africa, U & van Niekerk A 2008, 'GIS in the decision-making process' *Position IT*, Sept/Oct 2008, pp. 44-48.
13. Saha, A 2003 'Introduction to Artificial Neural Network Models' Retrieved March 2008, <http://www.geocities.com/adotsaha/NNinExcel.html>.

14. Statsoft, Retrieved June 2008, <http://www.statsoft.com/textbook/stathome.html>
15. Stergiou ,C & Siganos D 1996, 'Neural Networks', Retrieved March 2008 http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
16. Van Helden, P 2005, Introductory GIS – Class Notes, University of Pretoria.
17. 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.
18. 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.