# Intelligent Data Analysis for Conservation: Experiments with Rhino Horn Fingerprint Identification

**Rajan Amin**
Institute of Zoology, Zoological Society of London, UK
Rajan.Amin@zsl.org

**Max Bramer**
Faculty of Technology, University of Portsmouth, Portsmouth, UK
max.bramer@port.ac.uk
www.btinternet.com/~max.bramer

**Richard Emslie**
IUCN SSC African Rhino Specialist Group, KwaZulu-Natal, South Africa
remslie@kznwildlife.com

**Abstract** Conservation is an area in which a great deal of data has been collected over many years. Intelligent Data Analysis offers the possibility of analysing this data in an automatic fashion to map characteristics, identify trends and offer guidance for conservation action. This paper is concerned with the use of techniques of Intelligent Data Analysis for an important task in animal conservation: the identification of the species and origin of illegally traded or confiscated African rhino horn. It builds on an earlier analysis by the African Rhino Specialist Group. It is demonstrated that it is possible to distinguish between both species and country of origin with a high degree of accuracy and that the results are also likely to be suitable for use in court.

## 1. Introduction

The 2000 *Red List of Threatened Animals* published by the International Union for Conservation of Nature and Natural Resources (IUCN) identifies one in every four mammals and one in every eight birds as being at risk of extinction. The number of critically endangered species has increased significantly since the 1996 IUCN 'Red List' was published. IUCN and other national conservation agencies and NGOs are therefore seeking to increase the level of physical and legal protection of threatened species and habitats in order to conserve biodiversity.

Intelligent Data Analysis techniques offer the potential to assist the conservation process by improving the analysis of the mass of data that exists, much of which is multivariate and noisy. Apart from improving ecological understanding of how systems work, these techniques could provide much needed and improved practical tools e.g. to classify plants and animals, identify subspecies and hybrids, assess suitability and carrying capacities of areas for potential reintroduction, or to identify the origin of illegally traded animal and plant 'products'.

This paper describes the novel use of Intelligent Data Analysis techniques to identify the origin and species of illegally traded African rhino horn.

## 2. Rhino Horn Identification

There are five species of rhino alive today, two in Africa (White and Black rhino) and three in Asia (Indian, Javan and Sumatran rhino). Rhinos were among the first species to be included on the Convention on International Trade in Endangered Species of Wild Fauna and Flora's (CITES) 'Appendix I' list. This prohibited international trade in rhino products by CITES parties.

In Africa there was a 96% reduction in numbers of the critically endangered Black rhino from 1970-92. Numbers of the critically endangered Northern White rhino also declined from about 2,230 in 1960 to only 15 by 1984. Reintroduction initiatives have met with mixed success. The major cause of population decline is the demand for rhino horn, which has been used as an ingredient in traditional Asian medicine for 2,000 years and also for making the ceremonial curved daggers worn in some Middle-Eastern countries [1].

Although the international trade in rhino horn has been banned under the CITES convention since 1976, the demand remains high. Horn poaching generally involves killing the rhino and hacking off their horn(s). Demand may also inadvertently have been stimulated in Africa by the publication of exaggerated values of rhino horn in the press.

International conservation agencies are working to monitor and where possible eliminate illegal trade in rhino horn. To assist in the identification and prosecution of illegal rhino poaching/horn dealing cases and to improve knowledge of trade routes it would be highly desirable to be able to identify the species and origin of illegally traded horn; and if possible to do so in a manner which would be likely to prove convincing to a jury. This need resulted in the coordinating body for African rhino conservation, IUCN Species Survival Commission's African Rhino Specialist Group (AfRSG) initiating its continental horn fingerprinting project.

## 3. Rhino Horn Fingerprinting

Although it may appear as if a rhino horn is made of bone, it is in fact more akin to compacted hair and fingernails. Its chemical composition reflects what the animals have eaten throughout their lives. The chemical properties of their food are absorbed into the horn through digestive processes. Furthermore, the chemistry of this food varies in response to an area’s underlying geology, geomorphological history, and climate.

Rhino populations now largely occur in discrete separated populations, many of which are fenced. For this reason it has been hypothesised and subsequently confirmed that the chemical composition of the horns of rhino in the same park have strong similarities to each other and that these differentiate them from rhino living (or killed) in other parks. The white rhino is a grazer (eating tropical grasses) while the black rhino is a browser (eating succulent plants, trees and herbs). As these different plant types use different photosynthetic (chemical) pathways, there are chemical differences in the food the two species eat [2,3].

A report on the first phase of the AfRSG's horn fingerprinting project to its sponsors, the World Wide Fund for Nature (WWF), [2] states that a chemical analysis of rhino horn offers the potential for determining both the probable geographical source of the horn and the species of rhino that produced it.

A combination of variable values corresponding to a given rhino horn is known as its *fingerprint*. The development of a tool for identifying the source of an African rhino horn from its fingerprint has been rated by the AfRSG as a *Continentally Important* project.

In the early 1990s several studies determined that element and isotope concentrations and their ratios found in rhino horns varied between species and park origin [3, 4, 5]. The potential for horn fingerprinting indicated by these studies resulted in the AfRSG assembling an extensive rhino horn chemistry database and initiating a long-term project to further develop statistical models for the identification of the species and source of rhino horn. In the first phase of this project, the AfRSG successfully collected 361 rhino horn samples from 36 parks and 5 countries. The horns were analysed using three different chemical techniques and a database was compiled of their chemical characteristics. Discriminant Function Analysis (DFA) was explored as an approach for the development of the horn fingerprint identification tool.

# 4. Collection and Chemical Analysis of Rhino Horn Samples

The initial phase of the AfRSG study is described in detail in [2].

Horn samples were obtained from South Africa, Namibia, Kenya, Swaziland and Zimbabwe for the two species of African rhino (Black and White). Good coverage was obtained for the major rhino populations in South Africa, Namibia and Swaziland. Samples were obtained from 27 Black rhino populations and 22 White rhino populations. In 1997 the sampled populations conserved around two thirds of Africa's rhinos. While there are gaps in the coverage (especially for Zimbabwe and Kenya) the AfRSG project has made significant progress in establishing a continental horn database.

The horn samples were about 5cm$^3$ in size. They were cut up into smaller samples and analysed in three different laboratories, each using a different technique: carbon and nitrogen analysis using mass spectrometry, common and trace element analysis using inductively-coupled-plasma optical-emission-spectroscopy (ICP-OES) and heavier isotope analysis using laser ablation inductively-coupled-plasma mass spectrometry (LA-ICP-MS). Most but not all of the horns in the database were analysed in all three laboratories. The carbon and nitrogen analyses were undertaken at the University of Cape Town. The other analyses were carried out by Anglo American Research Laboratories in Johannesburg.

In the first analysis, four variables were measured: %C and %N, the percentage of Carbon and Nitrogen respectively, together with $\delta^{13}$C and $\delta^{15}$N which measure the 'delta ratios' of two isotopes, $^{13}$C to $^{12}$C and $^{15}$N to $^{14}$N respectively.

The second analysis (ICP-OES) quantified the abundance of 4 common element variables (Aluminium, Iron, Calcium and Magnesium) and 16 trace elements. In addition to these, a total of 66 ratios of the 12 most common elements measured by ICP-OES were calculated, e.g. Fe/Al (Iron divided by Aluminium).

The third chemical analysis measured the relative abundance of 132 isotopes of 58 elements using LA-ICP-MS, e.g. Cadmium (Cd) 110 111 112 113 114. Summing isotope values for elements with more than one isotope gave an additional 12 variables, e.g. SumCd (Cadmium). In addition, some of the more common isotopes were used to calculate nine potentially useful isotope ratios, e.g. Sr88Rb85 (Strontium88/Rubidium85).

Data were then examined for approximate normality, and where necessary the data were subjected to a Log+1 transformation to approximately realize a normal distribution.

Some of the isotopes are not required by rhino for normal metabolic functions and others were very rare with the result that a number of elements and isotopes occurred in such low quantities in some horn samples that they were beyond the detection capabilities of the machines, and could not be measured.

Principal component analysis (PCA) applied to the ICP-OES common and trace elements and ratios reduced the data to 12 principal components (OES1 - OES12). Two PCA runs were performed on the LA-ICP-MS data. The first made use of all 132 isotopes and produced 18 principal components (FLA1 - FLA18). The second used only the most common 41 isotopes, together with seven of the element sum variables such as SumCd, and derived 6 new principal components (MLA1 - MLA6).

As ratios of selected isotopes and elements have discriminatory potential, a set of 15 ratios of the 6 MLA components was also calculated. Only 7 of these were shown to have any discriminatory ability and the other 8 were discarded.

# 5. Initial Data Analysis by the AfRSG

The principal technique of data analysis used was classical linear *Discriminant Function Analysis* (DFA). Essentially this method predicts the group (i.e. the species, country etc.) to which a case belongs by deriving a series of mathematical functions that provide the greatest possible discrimination amongst groups. The *Statistica 5* software package was used to do this.

It was found that the best results were obtained by analysing the data in a hierarchical fashion, i.e. first establish the species, then the country, then the area and finally the park.

The AfRSG analyses confirmed the earlier discovery by Lee-Thorp and co-workers [3] that variable $\delta^{13}C$ was particularly good for discriminating amongst species. Analysis also showed that variable $\delta^{15}N$ was useful for distinguishing amongst species. The study also found that these two variables were related to climatic indicators in different areas. However, it was found that species identification was not as straightforward as initially expected. Using only $\delta^{13}C$, some black rhino samples from the Kunene area of Namibia were misclassified.

There were 356 usable samples available for constructing a model at the top level of the hierarchy to discriminate between species. However, lower down in the hierarchy many of the sample sizes were very small (e.g. all the rhino of a particular species in a particular country in a given area).

It is reported that many of the DFA models generated successfully classified all the samples used to build the model (100% 'post-hoc' classification success). However, the dangers of model overfitting were understood and the possibility of validating the derived models by techniques such as k-fold cross-validation or 'jack-knifing' was considered.

In the case of k-fold cross-validation, the original data is first divided into k approximately equal parts (generally 5 or 10). Next, k separate models are generated. Each of the k parts in turn is used to validate a model, and the other k-1 parts are used to generate it. The results of these k experiments are then combined to give an estimate of the accuracy of the modelling process on genuinely unseen data. Jack-knifing is the extreme case of k-fold cross-validation, where k is equal to the total number of examples available in the data. Unfortunately facilities for both forms of model validation were not available in the *Statistica 5* package.

Reference [2] concludes that 'Jack-knife validation is clearly the method that should be used to validate models in future'. In the absence of such validation they state 'while these results are very encouraging … readers still need to be cautious and treat these results as preliminary'.

Overall the results of these initial analyses were very promising. However the problems associated with small sample sizes, high data dimensionality and the need for careful validation of models were recognised. It is possible that these issues would be better addressed using techniques other than DFA.

A further important issue is the desirability of producing predictive models that can easily be explained in court and are likely to prove convincing to a jury. It is possible that this issue too may be better addressed by techniques other than DFA.

The remainder of this paper is concerned with bringing two widely-used techniques of Intelligent Data Analysis: Artificial Neural Nets and Automatic Induction of Classification Trees to bear on the problem of rhino horn identification and represents the next phase of the AfRSG's rhino horn fingerprinting project.

## 6. Developing Classification Models Using Neural Nets

Neural networks have been applied to many problems of learning classification models. They generally perform better than traditional methods when the problem is inherently non-linear (which is highly likely to be the situation with this data).

The most important property of a classification model is its ability to *generalise*. Simpler neural network models are capable of generalising better because they have fewer parameters to estimate. It is therefore important to follow a model development cycle aimed at reducing the number of variables as far as possible.

Although neural network parameters are estimated by minimising error on a training sample, the aim is to produce a prediction model that will perform well on out-of-sample data. It is therefore important to derive an estimate of performance out of sample. A number of techniques have been developed to derive these estimates. It is customary to divide the data available into three approximately equal parts: a *training* set, a *validation set* and a *test set*. The training set is used in conjunction with the validation set to train the model using a range of alternative configurations and initialisations of the network. The test set is used to estimate the performance of the final model on unseen data.

Sufficiently large sample sizes are needed to enable the available data to be divided into three smaller datasets. There is certainly a cause for concern that sample sizes were inadequate to do this in the case of the rhino data, especially at the lower levels of the species-country-park hierarchy, where there is only a restricted quantity of data available to construct and evaluate the neural network model. For this reason a jack-knifing approach was used. If there are N samples in a particular dataset, N models are constructed, in each case using just one sample for the test set, with the remaining N-1 samples used to build the model.

A considerable gain in the proportion of the data available for training was achieved using a technique called *Bayesian regularisation*, which avoids the need for a separate validation set during the training of Multi-layered feed-forward networks [6]. This was the principal form of Neural Network used in this study. For each partition of the data, 10 initialisations of the network were trained. An alternative to this network that can be used where the dataset is very small and the classes are unbalanced is the Probabilistic Neural Network (PNN) [7]. This was also used in this study.

The Matlab software package was used to process the data, with some additional use of the *Predict* package from Neuralware Inc.

## 7. Developing Classification Models Using Automatic Rule Induction Techniques

The Top-Down Induction of Decision Trees (TDIDT) algorithm is a widely used method for generating classification rules in the form of a decision tree. Assuming all variables are numerical, the antecedent (left-hand side) of each rule is a conjunction of terms such as $x<a$ or $x\geq a$, where $x$ is a variable and $a$ is a constant value. The consequent (right-hand side) of each rule is a classification such as *class=Black*. This is known as *disjunctive normal form*.

Use of the TDIDT algorithm goes back to the 1960s [8]. It is the basis of two very well known classification algorithms ID3 [9] and C4.5 [10] and many variants on these. At each stage of the tree generation process a choice of variables needs to be made. The most commonly used way of doing this is probably on the basis of maximising a measure known as *Information Gain* and that will be the choice assumed in this paper. Further information is given in [11].

The AfRSG report comments that 'To be successfully used in court the results (evidence) must be explainable simply and graphically'. Rule based models offer the potential to achieve this aim by providing an explicit representation of the underlying model. Each rule is a conjunction of simple tests on the values of variables, which should be readily understandable by the layman. The complete set of rules fit together into a tree structure, which is a form of the familiar flowchart.

The method is much less vulnerable than Neural Nets to the 'curse of dimensionality', i.e. problems of computational complexity associated with a large number of variables. The Information Gain criterion effectively 'filters out' unimportant variables. The method requires a training set and a test set, but no validation set, so it is generally possible to obtain good results with smaller datasets than for Neural Nets.

The potential value of using a classification tree approach was recognised by the AfRSG team, which commented that 'model outputs and structure lend themselves to being displayed graphically and are intuitively understandable' and 'analysis allows for a hierarchical analytic approach which is logically appealing and more flexible than traditional analysis'. Unfortunately the classification tree analysis module available in the most recent version of the *Statistica* package was not available in the older version (*Statistica 5*) used by the AfRSG for its analysis.

For the purposes of the study described below the implementation of the TDIDT (with Information Gain) algorithm in the ***Inducer*** Rule Induction Workbench [12] was used. This implementation supports both k-fold cross-validation and jack-knifing.

# 8. Experiments with Rhino Horn Identification

Some experiments using Intelligent Data Analysis techniques to identify the species and origin of a rhino horn from its chemical composition are described below. These are part of an ongoing programme of analysis which it is hoped in time will lead to the development of a fully reliable horn fingerprinting tool to establish the species and origin of an illegally traded rhino horn beyond reasonable doubt. As for the experiments described in Section 5, the data was analysed in a hierarchical fashion, i.e. species, then country, then area, then park.

## 8.1 Rhino Horn Fingerprint Data

The experiments made use of 52 of the variables identified during the original AfRSG study: the four variables from the carbon and nitrogen analysis, together with the 36 variables derived from Principal Component Analysis (OES1-OES12, MLA1-MLA6, FLA1-FLA18), four of the summed isotope values (such as SumCd) and eight of the isotope ratios (such as Sr88Rb85). These were the same variables used in the data analysis stage of the initial AfRSG study.

Five of the 361 samples gathered by the AfRSG were considered by them to be unusable, leaving 356. Thus the largest dataset used had 356 records (samples) each comprising the values of 52 variables plus a classification (species). This was subdivided into a number of smaller datasets for other classification tasks, all with 52 variables, containing only rhino of a particular species, rhino from a single country or from a specific park, area etc.

Forty of the 356 samples contained at least one and frequently several missing data values. The Neural Net and Discriminant Function Analysis algorithms were unable to cope with missing values so the number of samples was reduced to 316 for most of the experiments. The exception was the first experiment described below which used only the four Carbon/Nitrogen variables, for which there are no missing values in any of the 356 samples. (The TDIDT algorithm implemented in ***Inducer*** is able to cope with missing values by estimating their values, so can make use of all 356 samples.)

It is clear from [2] that the data suffers from many other problems as well as missing data values. These include the possibility of noise introduced by measurement errors or inconsistency in measurement techniques.

## 8.2 Species Discrimination

A Multi-Layered Perceptron (MLP) was produced using Matlab on all 356 of the usable samples to discriminate between Black and White rhino. In this and all subsequent experiments with the neural network methods it was first necessary to reduce the number of variables substantially from the 52 in the datasets described in Section 8.1 above.

Preliminary data analysis showed that the four Carbon and Nitrogen variables %C, $\delta^{13}$C, %N and $\delta^{15}$N were likely to be significant in discriminating between the Black and White rhino species. This was confirmed by using a genetic algorithm technique, which selected these 4 variables out of the 52 variables in the dataset and they were accordingly the variables used in generating the MLP. The MLP was trained with the four variables listed above using the Bayesian regularized Levenberg-Marquardt optimisation technique. It was tested by jack-knifing.

The *confusion matrix* below shows the number of times that each species was correctly or incorrectly classified.

|  | Predicted Species |  |
|---|---|---|
| Actual Species | White | Black |
| White | 178 | 0 |
| Black | 2 | 176 |

**Table 1. Confusion Matrix for Species Identification**

The overall predictive accuracy was 99.44%. The output of the neural network is in fact a probability of an input sample belonging to a White or Black rhino species. This probability can be used to calculate error bars. Both the misclassified samples had slightly higher error-bars than those for the rest of the data samples. This could be used to flag these samples as irresolvable.

A Probabilistic Neural Network (PNN) was also generated, using the same four variables, with 355 hidden Gaussian basis functions and a width parameter (chosen by an iterated search between 0.01 and 1) of 0.03. The model was tested by jack-knifing.

The TDIDT algorithm was used to generate a further model in the form of a decision tree. In this case all 52 variables were used during model (i.e. rule) generation. The ***Inducer*** package was able to cope easily with the additional variables. The algorithm was run using a number of the available options and settings and tested using jack-knifing. It was found that in this case the same (best) results were obtained using a maximum tree depth of just one as when the tree was allowed to grow without restriction.

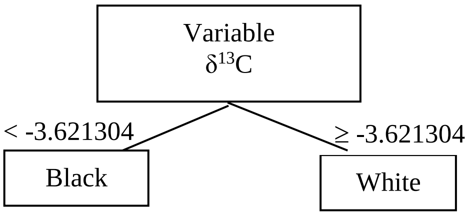
The models generated by PNN and TDIDT were both marginally inferior to that generated by MLP, giving one additional misclassification. This was also the result originally obtained using Discriminant Function Analysis. Thus all the methods used produced models with over 99% predictive accuracy.

The other methods do not output any explicit representation of the rules needed to discriminate between White and Black rhino on the basis of the values of the variables stored in the dataset, but TDIDT does so in a convenient form as follows.

In keeping with the earlier findings just two rules are needed to discriminate between the two species with over 99% accuracy:

1: IF $\delta^{13}C < -13.621304$ THEN Class = Black
2: IF $\delta^{13}C \geq -13.621304$ THEN Class = White

The above rules can be represented in graphical form as a simple decision tree:



It was noticed that all the methods used, including DFA, misclassified the same Black rhino samples and that these both originated from the Kunene area in North West Namibia. The earlier work by the AfRSG indicated that possible confusion between the species would be in very arid areas [2]. Analysis of additional Kunene black rhino horn samples would assist in further refining species identification.

## 8.3 Country Discrimination

Table 2 shows the distribution of the 316 samples that had no missing data values amongst the five countries of origin, broken down by species.

| | Namibia | South Africa | Kenya | Swaziland | Zimbabwe | TOTAL |
|---|---|---|---|---|---|---|
| Black | 38 | 97 | 10 | 0 | 18 | 163 |
| White | 5 | 137 | 0 | 10 | 1 | 153 |

**Table 2. Number of Horn Samples From Each Country**

Earlier AfRSG analysis showed that decomposing the problem of classification into parks, areas etc. by species gave better results than treating both species together. This was borne out by further experiments with Intelligent Data Analysis techniques.

*8.3.1 White Rhino*

For this experiment a dataset of the White rhino samples (all 52 variables) was used in order to predict the country of origin. The one sample for Zimbabwe was omitted, giving 152 records. A Probabilistic Neural Network was generated using all 52 variables as input. Iterated search revealed that a spread of $\sigma = 0.3$ to 0.35 was appropriate for the Gaussian basis functions. The model was cross-validated by jack-knifing. The overall predictive accuracy was 97.36%. This compares well with the use of Discriminant Function Analysis with the same 52 variables, which gave a classification accuracy of 94.7%. The following confusion matrix was obtained.

| | Predicted Country | | |
|---|---|---|---|
| Actual Country | Namibia | South Africa | Swaziland |
| Namibia | 4 | 1 | 0 |
| South Africa | 0 | 137 | 0 |
| Swaziland | 0 | 3 | 7 |

**Table 3. Confusion Matrix for Country Identification: White Rhino**

Although TDIDT also gave 'only' 94% accuracy, the method demonstrated that only 7 rules, involving tests on the values of just six variables were needed to gain that level of accuracy. They are listed below.

1: IF %C < 39.9 AND Sr88Rb85 < 0.754 AND OES8 < 0.989
THEN Class = Swaziland [1]
2: IF %C < 39.9 AND Sr88Rb85 < 0.754 AND OES8 ≥ 0.989
THEN Class = South Africa [16]
3: IF %C < 39.9 AND Sr88Rb85 ≥ 0.754 AND $\delta^{15}$N <0.898
THEN Class = Swaziland [9]
4: IF %C < 39.9 AND Sr88Rb85 ≥ 0.754 AND $\delta^{15}$N ≥ 0.898
THEN Class = South Africa [3]
5: IF %C ≥ 39.9 AND OES12 < 0.686 THEN Class = South Africa [113]
6: IF %C ≥ 39.9 AND OES12 ≥ 0.686 AND $\delta^{13}$C < -10.1
THEN Class = South Africa [5]
7: IF %C ≥ 39.9 AND OES12 ≥ 0.686 AND $\delta^{13}$C ≥ -10.1
THEN Class = Namibia [5]

The figures in parentheses indicate the rule coverage, i.e. the number of samples on which each of the rules is based. As before it would be straightforward to draw a graphical representation of these rules in flowchart form.

*8.3.2 Namibia v Swaziland*

A Probabilistic Neural Network was developed to discriminate between White rhino from Namibia and those from Swaziland, for both of which there are very few samples. The PNN was trained on the samples from Namibia and Swaziland and a

Density Mixture Model was built using the South African samples. The objective here is to use the Density Mixture Model to classify a sample as South African or otherwise (i.e. from Namibia or Swaziland). The PNN can then be used to classify the non-South African samples. The PNN model was cross-validated by jack-knifing. The confusion matrix obtained is given below.

|  | Predicted Country |  |
|---|---|---|
| Actual Country | Namibia | Swaziland |
| Namibia | 4 | 1 |
| Swaziland | 0 | 10 |

**Table 4. Confusion Matrix for Country Identification: Namibia v. Swaziland**

Only one discrimination error was made. However, its error bar was almost 0.5 while the rest of the data sample error bars were less than 0.01. This sample can be labelled as unclassified giving 100% accuracy for those classified, which is better than the DFA result of 94.10%. A Kohonen Self Organising Map was also trained to model the unconditional probability density of the South African White rhino samples only (density mixture model). The model was able to discriminate between the South African and the other samples.

TDIDT was also used to discriminate between the five samples from Namibia and the 10 from Swaziland. Only 2 rules are needed to do this:

1: IF %C < 41.2 THEN Class = Swaziland
2: IF %C ≥ 41.2 THEN Class = Namibia

### *8.3.3 Generating Additional Data for Black Rhino*

For this experiment a dataset of Black rhino samples was used to predict the country of origin. The ten samples from Kenya were omitted. As for White Rhino, the distribution of samples between the three remaining countries (classes) was highly unbalanced (38, 97 and 18), with by far the largest number coming from South Africa. In this case the problem of unbalanced classes was addressed by estimating the probability density of each class and generating 138 new data points by sampling from the distributions. This gave 97 samples from each of the three classes, Namibia, South Africa and Zimbabwe.

A Multi-layered feed-forward network was designed to classify the samples. The model was trained by Bayesian regularized Levenberg-Marquardt optimisation, and tested by 10-fold cross validation. The confusion matrix obtained is given as Table 5. The overall predictive accuracy is 96.56%, which is an improvement on the result from Discriminant Function Analysis (95.9%).

|  | Predicted Country |  |  |
|---|---|---|---|
| Actual Country | Namibia | South Africa | Zimbabwe |
| Namibia | 95 | 1 | 1 |
| South Africa | 3 | 89 | 5 |
| Zimbabwe | 0 | 0 | 97 |

**Table 5. Confusion Matrix for Country Identification: Black Rhino**

The same problem was tackled using TDIDT in a much simpler fashion. The 38 samples from Namibia were duplicated and the 18 samples from Zimbabwe were replicated four times, giving a distribution of (76, 97, 90) for the three countries. In this case the result obtained using jack-knifing was a predictive accuracy of 95.82%. All samples from Namibia and Zimbabwe were correctly classified, with 8 of the South African samples wrongly classified as being from Namibia and 3 of them wrongly classified as being from Zimbabwe.

## 8.4 Park Determination

South Africa is divided into nine provinces and it was decided to determine how reliably it was possible to distinguish between horn samples from six black rhino and six white rhino populations in the province of KwaZulu-Natal. Seeking to distinguish amongst different parks within a region is a more rigorous test of data analysis techniques than distinguishing amongst parks that are widely separated. There are more than 2 samples for each species for each of the parks (74 samples for White rhino and 58 for Black rhino). The confusion matrices were as follows.

| Actual | Predicted Park |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Park | Park 1 | Park 2 | Park 3 | Park 4 | Park 5 | Park 6 |
| Park 1 | 29 | 5 | 2 | 0 | 1 | 0 |
| Park 2 | 4 | 13 | 0 | 0 | 1 | 0 |
| Park 3 | 3 | 1 | 2 | 0 | 0 | 0 |
| Park 4 | 0 | 1 | 0 | 2 | 0 | 2 |
| Park 5 | 0 | 2 | 0 | 1 | 2 | 0 |
| Park 6 | 0 | 0 | 0 | 2 | 1 | 0 |

**Table 6. Confusion Matrix for Park Identification: White Rhino**

| Actual | Predicted Park |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Park | Park 1 | Park 2 | Park 3 | Park 4 | Park 5 | Park 6 |
| Park 1 | 28 | 0 | 1 | 1 | 0 | 0 |
| Park 2 | 1 | 5 | 0 | 0 | 0 | 0 |
| Park 3 | 2 | 1 | 3 | 0 | 0 | 0 |
| Park 4 | 2 | 0 | 0 | 4 | 0 | 0 |
| Park 5 | 1 | 1 | 1 | 0 | 2 | 0 |
| Park 6 | 3 | 0 | 2 | 0 | 0 | 0 |

**Table 7. Confusion Matrix for Park Identification: Black Rhino**

In this case Intelligent Data Analysis methods gave a predictive accuracy of 64.9% for White rhino and 72.4% for Black rhino. The former result is a considerable improvement on the DFA result of just 40.5% predictive accuracy. The very unbalanced distribution of classes and the small number of samples for some classes (parks) are likely to prove problematic for any method of analysis. However the results are very encouraging and predictive accuracy should increase substantially in future as sample sizes per park increase.

### 8.5 Potential Use of Novelty Filters

One requirement for turning horn fingerprinting into a practical routine forensic test is that it is necessary to be able to detect whether some samples are likely to have come from areas not yet covered by the continental rhino horn chemistry database.

One possible approach would be to train a Kohonen Self Organising Map as a novelty filter. This is briefly discussed in Section 8.3.2.

## 9. Discussion and Conclusions

The experiments described above are part of the AfRSG's ongoing programme aimed at turning horn fingerprinting into a practical and reliable forensic tool. The concern that the original models may have been overfitted was confirmed by the later analysis. It is also clear that the two methods of Intelligent Data Analysis used, neural nets and automatic rule induction, improve upon Discriminant Function Analysis as a means of analysing the rhino horn data and are less prone to problems of model overfitting.

The AfRSG's interest in using classification trees would appear to be justified. They give an explicit representation of the decision process in the form of rules as well as having a natural graphical representation as flowcharts, thus helping to meet the AfRSG report's requirement that 'results … must be explainable simply and graphically'. Neural networks can be formulated to give probability outputs, which is very useful. They can also be used as novelty detectors to identify whether or not samples have come from areas not yet included in the continental horn fingerprinting database.

The Intelligent Data Analysis work described here has taken a step closer to producing a field operational system.

The small number of samples currently available for classification at park level will create difficulties for any method; but these are likely to be smaller for rule induction than for the other methods used. The obvious (although highly understandable) deficiencies in the data have not prevented a high level of predictive accuracy being obtained in many cases. However, for practical use, the predictive accuracy of the technique at park level needs to be increased. An experiment is planned where additional samples from a few parks will be analysed to determine how many samples ideally need to be collected per park to raise predictive accuracy to an acceptable level.

Intelligent Data Analysis is often assumed to be concerned with the processing of large volumes of data. By contrast the task of rhino horn classification is concerned with small datasets with noisy data values, missing values, unbalanced class distributions and for some classes (e.g. individual parks) only very few examples. These all present significant technical challenges.

The development of a rhino horn fingerprinting tool is a high priority. This is likely to combine a range of analytic techniques with facilities for displaying information in graphical form both for use by conservationists and for possible presentation to juries.

Conservation is an area in which a great deal of data has been collected over many years. Intelligent Data Analysis offers the possibility of analysing this data in an automatic fashion to map characteristics, identify trends and offer guidance for conservation action.

## Acknowledgements

We would like to acknowledge the AfRSG Chairman, Dr Martin Brooks, for his role in initiating the AfRSG horn fingerprinting project, and his initial work in raising the necessary funding and obtaining many of the samples. The work of the initial pioneers in this field [3,4,5] provided the base on which subsequent work has been built. During the project planning phase, advice from Julia Lee-Thorp (University of Cape Town), Roger Hart (University of the Witwatersrand) and Anna Jolles (University of Princeton) was invaluable.

The AfRSG thanks all those in the field who provided it with samples. Julia Lee-Thorp also kindly made additional samples available to the AfRSG that were left over from her initial horn fingerprinting work. Julia also is thanked for facilitating the carbon and nitrogen analyses at the University of Cape Town. Dr Richard Holdsworth and Nannete Schutte of Anglo American Research Laboratories (AARL) are thanked for expediting the analysis of the samples at AARL and for their useful comments.

WWF is gratefully acknowledged for funding the collection of horn samples, and their original chemical and statistical analysis and their partial support of the AfRSG Scientific Officer. The Italian funded SADC Regional Programme for Rhino Conservation is thanked for partially funding the next phase of data analysis.

# References

[1] Emslie, R.H. and Brooks, P.M. (1999). African Rhino. Status Survey and Conservation Action Plan IUCN/SSC African Rhino Specialist Group. IUCN, Gland, Switzerland and Cambridge, UK.

[2] Emslie, R.H., Brooks, P.M., Lee-Thorp, J.A., Jolles, A., Smith, W. and Vermaas, N. (2001). Development of a Continental African Rhino Horn Fingerprinting Database and Statistical Models to Determine the Probable Species and Source of Rhino Horn. AfRSG Rhino Horn Fingerprinting for Security Project 9F0084.1. Unpublished Report to WWF

[3] Lee-Thorp, J.A., van der Merwe N.J. and Armstrong R.A. (1992). Final Project Report ZA309: Source Area Determination of Rhino Horn by Isotropic Analysis. Unpublished WWF Report

[4] Hall-Martin, A.J., van der Merwe N.J., Lee-Thorp J.A., Armstrong R.A., Mehl, C.H., Struben, S. and Tykot, R. (1993). Determination of Species and Geographic Origin of Rhinoceros by Isotropic Analysis and its Possible Implication to Trade Controls. Proceedings of International Rhino Conference, San Diego, California, p.123-124

[5] Hart, R.J., Tredoux, M. and Damarupurshad, A. (1994). The Characterisation of Rhino Horn and Elephant Ivory Using the Technique of Neuron Activation Analysis. Final report on a project undertaken on behalf of the Department of Environmental Affairs, South Africa

[6] Mackay, D.J.C. (1992). Bayesian Interpolation. Neural Computation, Vol. 4, No. 3, pp. 415-447

[7] Masters, T. (1993). Practical Neural Network Recipes in C++, Academic Press

[8] Hunt, E.B., Marin J. and Stone, P.J. (1966). Experiments in Induction. Academic Press

[9] Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning, 1: 81-106

[10] Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann

[11] Bramer, M.A. (2000). Automatic Induction of Classification Rules from Examples Using N-Prism. In: Research and Development in Intelligent Systems XVI. Springer-Verlag, pp. 99-121

[12] Bramer, M.A. (2000). Inducer: a Rule Induction Workbench for Data Mining. In Proceedings of the 16th IFIP World Computer Congress Conference on Intelligent Information Processing (eds. Z.Shi, B.Faltings and M.Musen). Publishing House of Electronics Industry (Beijing), pp. 499-506