MIKE Data Analysis Strategy

R.W. Burn, F.M. Underwood and N.D. Hunter
December 2003

1. Introduction

MIKE is an ambitious monitoring system with multiple objectives and consequently complex data requirements. Given the scope (both geographical and over time) of MIKE, it is vital that there is a clear strategy for both the management of the data and its analysis in order to meet the overall objectives. At the present stage of development, the main database structures are already in place and the data management strategy is well under way. Here we address the issue of a strategy for the analysis of MIKE data.

'Data analysis' is understood in its broadest sense, ranging from simple summaries such as tables and charts for routine reports to statistical inference based on models for analysing trends and establishing relationships between variables. Much of the data analysis for MIKE can be achieved with quite simple statistical methods. The principle to adopt is to strive to keep the analysis as simple as possible and to avoid unwarranted complexity. However, there are occasions when the overall objectives of MIKE will demand more advanced approaches.

2. Objectives

The starting point for a strategy for the analysis of MIKE data must be the **MIKE/ETIS objectives** (COP 10.10, rev. in COP 12) …

(i) measuring and recording levels and trends, and changes in levels and trends, of illegal hunting and trade in ivory in elephant range states, and in trade entrepots;

(ii) assessing whether and to what extent observed trends are related to changes in the listing of elephant populations in the CITES Appendices and/or the resumption of legal international trade in ivory;

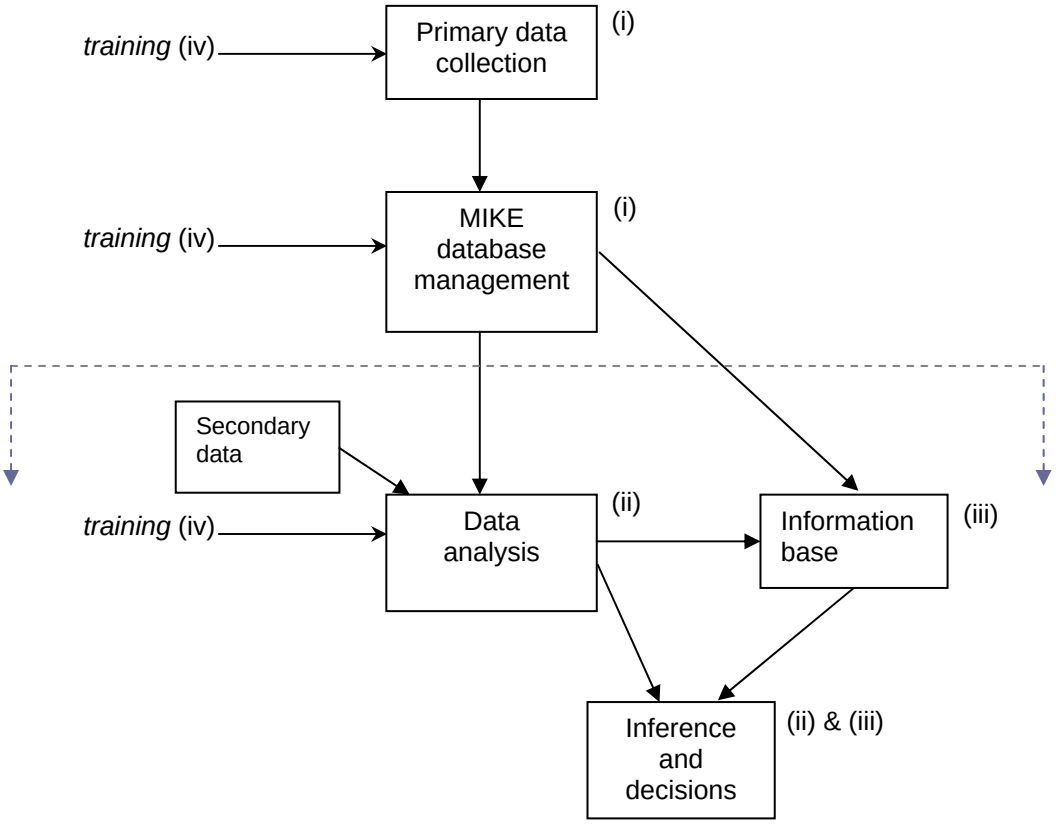
(iii) establishing an information base to support the making of decisions on appropriate management, protection and enforcement needs;

(iv) building capacity in range states.

In broad terms, objective (i) relates to data collection and objective (ii) to data analysis. The foundation of the 'information base' required by objective (iii) is the MIKE database which is currently being implemented. The objectives imply the need for a degree of data analysis capability at each level in the MIKE structure. The analysis required to meet objective (ii) is the global analysis of trends and patterns of association and will largely take place at the top level, i.e. the Central Coordinating Unit (CCU). Some data analysis is also called for in making use of information at national and site levels for the management and decision making mentioned in objective (iii).

Objective (iv), as well as being a product of the MIKE process in its own right, aims to provide the training necessary to accomplish the other objectives. The participation of MIKE staff in data analysis entails a need for training in the ideas of statistical inference and data analysis techniques. This training is seen as an integral part of the data analysis strategy. It is important to understand that this training for data *analysis* is additional to the existing training in data *management*, although there is some overlap (see Section 9).

The flowchart below is a simple model of the MIKE information system and how it relates to the objectives (labelled (i), (ii), etc in the diagram). The data analysis strategy concerns the part of the flowchart below the broken line. However it is important that considerations of data collection and database management are not totally divorced from the data analysis strategy. Decisions made about the data collected and how it is stored have a direct influence on how effectively data can be analysed and, conversely, analysis objectives necessarily impact on the collection and management of data.



3. The MIKE information hierarchy

The MIKE information system functions at various levels in a geographical hierarchy. These are:

regions (Asia and Africa);
sub-regions within regions (e.g. East Africa, Central Africa, etc.);
countries within sub-regions;

sites within countries.

The entire system is overseen by the CCU which in turn reports to the CITES Secretariat and thus to the CoP's.

Although there is a flow of information *up* the hierarchy, each level has its own requirements with regard to support for management and decision-making. The system must simultaneously meet these needs at all levels. Data collection takes place primarily at the lowest level in the hierarchy (sites), but some kind of data analysis is required at every level. For the purposes of comparability and consistency of reporting, it is important that, at least for the more routine analyses, the strategy aims to develop standardised analysis procedures across all members of each level of the hierarchy.

The movement of information from sites through countries and sub-regions up to CCU level will demand considerable vertical integration and collaboration. The local processing and analysis of data at each level is important to ensure a proper sense of ownership of the data and to empower stakeholders to make their own use of the information they are collecting. At the same time, it is crucially important for the global objectives of MIKE that there is an easy flow of data through the hierarchy.

For effective statistical analysis at sub-regional, regional and CCU levels, **it is crucial that raw data, and not just summarised or aggregated data, are always transmitted from each level in the hierarchy to the next**. On this point, it is possible that there will be issues to resolve at national level in certain cases. Sensitivities regarding ownership of data are not uncommon and often quite understandable. However, the importance of having access to raw data for analysis, at least at CCU level, cannot be over-emphasised.

4. MIKE data components and sources

For convenience, we recall MIKE's main data collection activities and summarise them below:

elephant population surveys

- dung surveys by line transects for forest elephants, and data from dung decay surveys;
- possibly data from other types of survey (mark-recapture, direct observation, etc.)
- aerial surveys for savannah elephants;

law enforcement monitoring (LEM) from ground patrol activities

- incidence of poaching and other illegal activities,
- carcass counts*,
- law enforcement effort data from patrols;

site-level covariate information

- data on human activities in the proximity of elephant ranges,
- ecosystem type and habitat variables;

secondary data sources (desk research)

* Data on carcass counts will also arise from population surveys, local knowledge and other sources.

- law enforcement budgets, personnel, etc.,
- background socio-economic information,
- ivory and other elephant product seizure data from ETIS,
- domestic ivory markets;
- maps of sites and areas where elephants might be found;

other data (intelligence information etc.)

The primary data are stored and managed in the MIKE database system. Sources of secondary data include the ETIS database maintained by TRAFFIC and web-based sources on background information such the World Bank. For the purposes of ETIS data analysis, TRAFFIC maintains a country-level database on background variables. This should be shared with MIKE to avoid duplication of effort. Further country-level data will presumably be available from national statistics offices and other government departments.

5. Analysis goals

To translate the broad aims implied by the MIKE objectives into operational reality, decisions concerning precisely what is to be estimated and which questions are to be investigated are required. In order to get the analysis process going, such decisions are required at an early stage. A provisional list of questions at site level has been decided upon. It is likely that further questions and goals will emerge as the process evolves and these initial decisions should not be allowed to ossify into a fixed routine of data analysis.

Once the questions to be answered have been identified the next step is to decide on the variables required to address them. The MIKE process will generate a lot of data, in terms of both the number of variables and also sequentially over time. Selecting variables to meet defined analysis aims has to be achieved through an iterative process of exploratory data analysis and statistical modelling, as outlined in Section 6.3 below. There will always be a need for thinking of new (informative) analyses and responding to new questions and information requirements.

6. Top level data analysis

MIKE objective (ii), assessing the impact of CITES policy on trends in illegal killing of elephants, is the most challenging requirement of all and requires a careful analysis of *all* factors thought likely to influence trends. A strategic decision taken early on in the MIKE process is that abundance estimates of elephant populations will normally be required as essential information in the analysis of trends in illegal killing: changes in illegal killing can only be fully understood in relation to changes in elephant population numbers. Thus the global analysis of MIKE data must not only describe trends in illegal killing but must also quantify the impact of these trends on elephant populations.

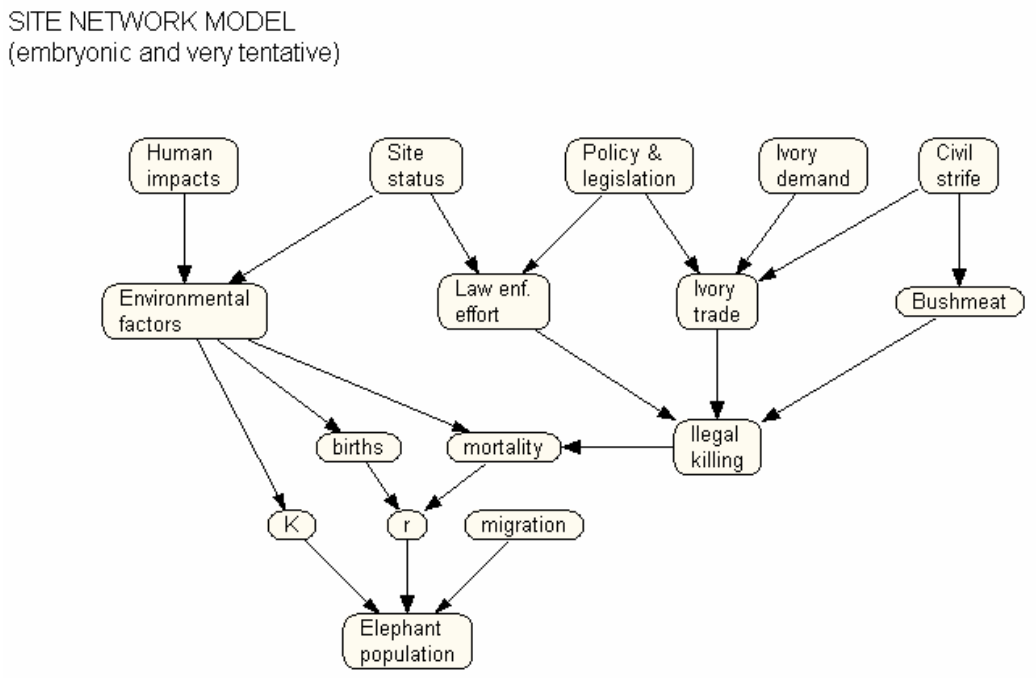
The first task of the global analysis is the aggregation of sub-regional and national data to produce descriptive statistics representing overall trends and summaries. Although dramatic changes over short time periods will soon become evident, it is anticipated that more gradual trends may require a considerable period of time before they can definitively be established. The second, more difficult task is to assess the

effects of factors that may influence the incidence of illegal killing and the analysis of global/regional trends. A feature of the MIKE data is that there are potentially quite complex interrelating patterns of causality between the variables. A modelling framework that can assess these causal relations, at least informally, is Bayesian network models (described below). The findings of this kind of analysis are often not sufficiently precise to establish the statistical significance of the effects of factors on outcomes, but they can be used to indicate which relationships are worthy of a more rigorous statistical analysis. The procedure would therefore be to construct Bayesian network models for elephant population numbers and measures of illegal killing, followed up by statistical modelling of key outcome variables and influencing factors.

6.1 Bayesian network models

Since Bayesian networks are perhaps less familiar than the more usual statistical methods, an outline of the basic ideas, with references, is given in Appendix 1. Briefly, the idea is to model patterns of causal relations with a causal diagram together with probabilistic relations between the nodes (variables). The probabilities linking the nodes would preferably be estimated from data, or in the absence of suitable data, they can be subjective probabilities elicited from expert judgement. Software packages (see Section 8) including powerful computational algorithms are available that permit useful inferences from the resulting structure.

A tentative model of the kind of Bayesian network model that would be useful for assessing the impacts of causal effects on elephant numbers is shown below. This model is not at all definitive and is shown here just to give a flavour of the kind of modelling that can be achieved with Bayesian networks.



In practice, the development of a network model proceeds in stages. First the qualitative assignment of causal links (as displayed in the above example) evolves from careful consideration by groups of scientists with expert knowledge of the field (Jensen, 2001). The next stage is to review all available data with a view to

constructing table of conditional probabilities that quantify the links between the nodes. The example network model above would require data from the full range of sources: detailed MIKE data from the sites as well as data from secondary sources and ETIS. In cases where no adequate data are available, a process of elicitation (O'Hagan, 1998) from panels of experts can be used to construct subjective probabilities. A strength of Bayesian network models is that unobservable (latent) variables can be included in the model, for example the 'K' node (carrying capacity) in the above model. For these nodes, there would be no directly observable data and the elicitation of subjective probabilities would be appropriate.

The fictitious example above is a *static* model - it takes no account of repeated observations over time - and is therefore not appropriate for investigating trends. However, taking the model as a 'time slice' (Jensen, 2001), it can be made into a *dynamic* Bayesian network. This would allow the modelling of time-dependent and even lagged effects.

To meet the objectives of the global analysis of MIKE data, different network models would be constructed: for example, one for analysing causes of changes in elephant populations and another for analysing factors affecting illegal killing.

6.2 Statistical modelling

We mention here some particular statistical methods which will be required for the more rigorous analysis of relationships. Precisely which relationships are analysed will result from decisions on the key questions that MIKE needs to address, mentioned in Section 5. For the analysis of trends, when sufficient data eventually become available, statistical methods for smoothing time series, such as loess or splines, are appropriate. A class of regression-like models that accomplishes this smoothing and at the same time models the effects of covariates is generalised additive models (GAMs) (Hastie & Tibshirani, 1990).

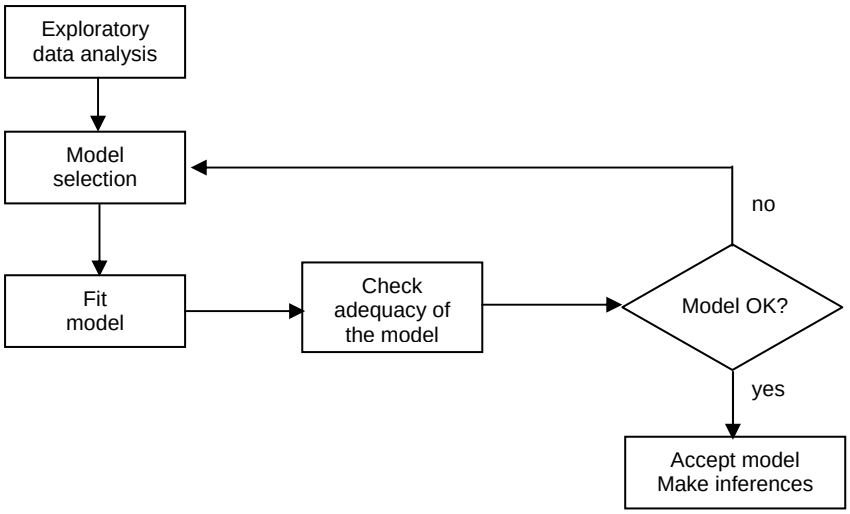
The hierarchical nature of the data structures indicate the need for multi-level modelling (Goldstein, 1995), and the analysis of factors influencing trends in time would require statistical models for longitudinal data (Diggle *et al*, 1994). These methods, together with smoothing techniques were used in the recent analysis of ETIS data presented to the CITES 12th Conference of the Parties (Milliken *et al*, 2002).

For the purposes of extrapolating elephant abundance estimates from survey sites to larger regions, the use of spatio-temporal modelling by GAMs should be explored, provided adequate covariate information is available. This approach has been successfully applied to a number of animal abundance estimation problems (see Augustin *et al* for a fisheries example). This is an example of model-based inference (Borchers *et al*, 2002; Thompson, 1992).

6.3 An evolving data analysis strategy

Assessing the relationships between variables is accomplished by fitting statistical models to data (regression models and various generalisations of them). Statistical modelling proceeds by an iterative process of data exploration, fitting a tentative model, examining the adequacy of the model, fitting a modified model, and so on until a reasonable balance between accuracy and simplicity can be found. This is a collaborative process between statisticians and scientists.

The data analysis strategy must have enough flexibility to allow this exploratory and interactive approach to data modelling.



Exploratory data analysis embodies a variety of techniques for

(1) the initial screening of data with a view to identifying outliers and other quality checks;

(2) preliminary investigation of relationships between variables;

(3) in situations where there are many variables, attempting to reduce the dimensionality of the problem by identifying redundancies or finding derived variables.

Methods for achieving (1) and (2) are mainly graphical and descriptive. Techniques that have been found useful for (3) include multivariate methods such as principal components analysis and variable clustering methods (Krzanowski, 1988).

The *model selection* process, as depicted in the flow-chart above, is iterative and should be seen as a part of the inference process (Buckland et al, 1997). In the first instance, considerable thought needs to be given to initial candidate models. It is likely that the model selection process will reveal several equally viable models for the data, in which case it would be sensible to attempt model averaging or other multi-model approaches (Burnham & Anderson, 2002).

Although it *is* possible (and efficient) to pre-program certain routine parts of the analysis, a fully automated 'push-button' analysis system for the entire MIKE process would be both undesirable and unobtainable, at least in the short to medium term. What is required at this stage is a strategy which is sufficiently flexible to develop with the system. In time, more will be learnt about the processes being monitored, problems will be solved and new techniques will evolve.

Although we represent data analysis near the end of the 'data process' (data collection, data management and data analysis), the results from data analyses will feed back into future data collection, which also incorporates design of relevant studies, and data management decisions. In this way the whole data process evolves through time.

6.4 Analysis of LEM data

Two key issues yet to be resolved in the analysis of LEM data are:

(i) how to measure law enforcement effort, and
(ii) how to adjust for effort in measuring incidence of illegal killing.

These questions must be settled before any coherent analysis of LEM can be achieved.

What is needed is a simple measure that can be applied consistently across sites, one measure for forest sites and one for savannah sites. Some preliminary analysis of existing LEM data should be undertaken to help derive this measure. Clearly, the results of law enforcement patrols (carcass counts, indicators of illegal killing, etc.) will depend on the amount of effort expended by the patrol. By analogy with methods of estimating fish stocks, 'catch-effort' methods have been suggested (Jachmann, 1998). As yet there do not appear to be robust measures of law enforcement effort that are generally applicable, to make it work well. It is likely that much can be learned from data LEM collected in the MIKE process itself. When sufficient data become available, it should be possible to use statistical modelling methods to investigate relationships between patrol results and candidate variables for measuring effort.

A number of issues concerning the way in which MIKE data are collected will inevitably impose limitations to the inferences and extrapolations that are statistically valid. The problems are:

> The route for a patrol may not be chosen according to any statistical sampling plan, random or otherwise. Patrol routes that are determined purposively, possibly even chosen as a result of intelligence reports, would lead to biased estimates of incidence of illegal killing. In some cases, patrols may repeatedly follow the same path, or one of a small set of paths. This would also lead to bias. If patrol routes are not chosen according to a statistical design, inference to a larger domain is generally not justified (unless some form of model-based inference is possible – see below).
>
> Although great care has been taken in the design of the data collection forms, it remains true that the members of a patrol are primarily engaged in law enforcement and not scientific observation, with the result that reliability and quality of the data may be compromised.

These considerations impose constraints on the degree to which inferences from LEM data can be extrapolated to any larger domains. However, these limitations are to some extent mitigated by at least two factors:

> Eventually, it may be possible to use spatial modelling techniques and then use model-based inference to derive estimates over larger ranges. This would require that relevant covariate information (spatial and other) are systematically collected not only over the range covered by existing patrols, but over the wider domain over which the estimates are required to be valid. However, although theoretically possible, this option is a bit of a long shot!
>
> It should still be possible to use LEM data from patrols as covariates to assist in the analysis of elephant population survey data. In particular, rates of illegal killing in the general vicinity of the survey site will probably turn out to be useful.

7. National and site level analysis

The main requirement at national and site levels is the analysis of population survey and LEM data. As well as providing the data needed for the global analysis at

regional and CCU levels, the analysis must also meet the local needs of monitoring and management. Much of the analysis will consist of simple data summaries in tabular and graphical forms for routine reporting purposes. However, if local staff are to get the most out of their data, some statistical analysis would be useful. An important aspect of data processing which must take place at site level is data quality control, which has to be done before any kind of data summary or analysis. These requirements have implications for capacity building, an issue taken up in Section 9. Analysis done at CCU or regional level will often be of direct interest at national and site levels. This would apply especially to methods such as spatial modelling and some of the more sophisticated trends analysis. National site officers should therefore have easy access to these results.

7.1 LEM data

Once the measures of law enforcement effort and procedures for catch-effort analysis mentioned in Section 6.4 are decided, it will be possible to produce analyses of adjusted measures of carcass counts, rates of illegal killing, mortality statistics, etc. After a period of development, most of these routine analysis procedures should be programmed. This would not only lighten the burden of repetitive analysis but also ensure a uniform reporting format which would enable easier comparisons between sites and over time. However, there is a real possibility that these routine procedures will periodically change. As more data become available, for instance, the CPUE relationships will need to be updated. The system must have enough flexibility to respond to these changes.

Some *ad hoc* statistical analysis (such as regression or analysis of variance) would enhance the ability of site and national staff to learn from their data. For instance, from time to time it would be useful to be able to compare survey sites or law enforcement regimes, or to explore the effects of other covariates.

7.2 Population survey data

For savannah sites, the main requirement is the analysis of aerial survey data to obtain unbiased population estimates and standard errors. Abundance estimates of forest elephants are mainly obtained from line transect dung surveys. The analysis of data from line transect surveys requires the methods of distance sampling (Buckland *et al*, 2001) and the use of the DISTANCE software. Dung count surveys should be accompanied by dung decay surveys and the mean decay time estimated using, for example, the methods of Laing *et al*. In Asian range states, in addition to abundance estimates, monitoring the sex ratio in elephant populations, possibly derived from a DNA analysis of dung, will also be important, and it is possible that other survey methods (Section 4) will be used, especially for small or sparse populations.

An issue relating to the design of population surveys in general is the question of statistical power (Green, 1994). A realistic power analysis needs to be developed at an early stage to ensure the efficient use of resources in both forest and savannah sites. This analysis, which would be done at global level, should provide guidelines for use at site and national levels for survey design.

8. Software

In addition to the ArcView GIS software, the usual configuration of Microsoft Office software, especially Excel and Word, and the custom MIKE database management system, a general purpose statistical package should be installed on all sites. For this

we propose the Genstat system (NAG, Oxford, 2003), subject to an initial evaluation of its suitability. Version 7 of Genstat will be available around October 2003. An agreement has been signed between the MIKE Director and the suppliers of Genstat (NAG/VSN) for the supply of the software, including some special extensions to its functionality for the MIKE process (the details of which have yet to be decided).

For analysis of Bayesian network models, the Netica program (Norsys Software Corp., 1998) is required. This would be required only at CCU, or perhaps regional level. Netica is available quite cheaply (around USD 400) for non-commercial purposes, and only one licence is required (for the CCU).

Program DISTANCE (Buckland *et al*, 2001) is required on all sites doing line transect surveys. It is freely available from the web-site of St. Andrews University, UK. Other software requirements will be met by custom programs written for specific tasks such as the analysis dung decay survey data and for aspects of LEM data analysis.

9. Training

Capacity building in range states is one of the MIKE objectives and is necessary both for accomplishing the overall MIKE tasks and for the broader needs of the range states themselves. Experience has shown that one-off training events with no follow-up have limited value. An on-going programme of training for some years to come should be seen as one of MIKE's activities. The long term objective is to achieve as much self-sufficiency in analysis techniques as is possible.

As mentioned in Section 2, training for data analysis should be seen as complementary to the training in the use of the MIKE database system, currently under way.

The overall aims of the training are:

(a) to equip staff at all levels with the skills required for managing and analysing data to meet MIKE's objectives;

(b) to build capacity in range states so that national and site staff can make use of their data to the greatest effect for their own purposes.

9.1 A training strategy

The strategy that we propose is to initiate the training programme with a series of sub-regional workshops. The effectiveness of this initial training can subsequently be made more sustainable in two ways:

(1) Strive to identify at least one key participant from each sub-region who would be able to take on the role of secondary trainer themselves. (In sub-regions which are not predominantly Anglophone, the choice of key participant must clearly be made according to local language requirements.) This would help the further dissemination of ideas on data management and analysis throughout sites.

(2) Work towards making creative use of web-based methods for further dissemination and exchange of ideas. There is great potential here for the future, although perhaps limited in the short term, at least in certain countries, by difficult access to the internet. The web could certainly be used for the traditional "distance learning" approach, but a much more exciting prospect is

to use the web to set up communications between sites for the exchange of ideas and discussion of common problems, with or without the intervention of "trainers". Experience suggests that e-mail access may be easier in some parts of Africa and Asia than internet access, and it would be quite easy to set up mail-lists for the purposes of this kind of exchange. Countries where even that is difficult will not, at least in the short term, benefit from this approach.

9.2 The initial training workshops

It is proposed that training workshops are organised in the sub-regions, according to language needs, as follows:

(a) West & Central Africa (Anglophone);
(b) West & Central Africa (Francophone);
(c) East Africa;
(d) Southern Africa;
(e) South Asia;
(f) South-East Asia.

There is likely to be considerable variation in training needs at national and site levels and the precise content of the training will depend on assessments from national officers and other senior staff. The following description is indicative of what should be achieved.

Duration: 3 to 4 weeks for each workshop.

Objectives:

(a) to develop data management skills appropriate for analysis;
(b) learn principles of effective data summary;
(c) learn statistical methods for basic analysis of data;
(d) acquire training skills.

Content:

(1) Structure of MIKE data; review of MIKE database system; exporting data to Excel; data quality checks.
(2) Data management with Excel: list format; filtering and sorting; simple lookups.
(3) Data summary and descriptive statistics with Excel: basic calculations, means and standard deviations, rates; pivot tables and charts.
(4) Introduction to Genstat.
(5) Basic statistical inference: confidence intervals; comparing means; simple linear regression.
(6) Time series: smoothing and trends.
(7) Analysis of population survey data.
(8) Extracting LEM data from the MIKE database and producing tabular and graphical summaries.
(9) Training the trainers: discussion and planning of training at national and site levels.

Specialists in training in applied statistics and data management should be contracted to run these initial workshops (Section 10) and the MIKE Data Coordinator should

also contribute. However, provided the training of trainers is effective, reliance on these outside resources should be reduced and eventually phased out altogether.

9.3 The long term perspective

Training needs must be periodically reviewed. The MIKE process is itself evolving and will continue to do so. New techniques become available and software is always changing and (hopefully) improving. There is also a natural turnover in staff at all levels. Even without these changes taking place, as mentioned above, one-off training events are never sufficient to ensure a sustainable level of skill. To be effective, capacity building itself needs to be seen as a sustainable activity. Future training will be greatly enhanced both by feedback from work experience and by evaluation of training by the participants themselves. These exchanges will eventually be greatly facilitated by the use of the internet as outlined above.

The effectiveness of the training programme should be carefully monitored by the CCU and training workshops as outlined above should be organised on a regular basis, annually or at least every two years, or even on an *ad hoc* basis if the need is indicated by the monitoring.

10. Human resources: who does what?

The aim of the training programme outlined above is that analysis of data at national and site levels will ultimately be the responsibility of local staff. However, some inputs from regional officers, with support from external consultants acting as advisers may be necessary in the short term. Support across regions or sub-regions could also be envisaged.

At regional and CCU levels, where the more advanced statistical methods are called for, collaboration between the MIKE Data Coordinator and statisticians will be necessary. Statisticians with expertise in the methods described above should be contracted by MIKE to assist in the development of the data analysis strategy. Given the key role of capacity building in the MIKE process, training specialists should also be engaged, especially in the early stages. The Data Coordinator should also make substantial contributions to the training.

Acknowledgements

We are grateful to Réné Beyers, Holly Dublin, David Borchers, Ken Burnham, Ian Dale, Tom Milliken, Iain Douglas-Hamilton and Roger Stern for useful discussions and communications. TAG members during the April 2003 and December 2004 TAG meetings also contributed valuable ideas.

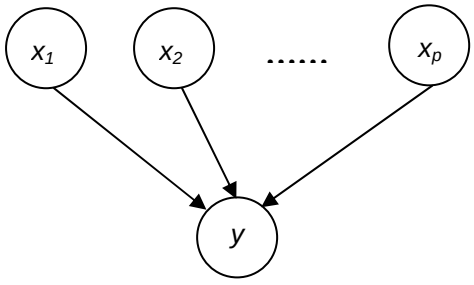
Appendix 1

Bayesian Belief Networks

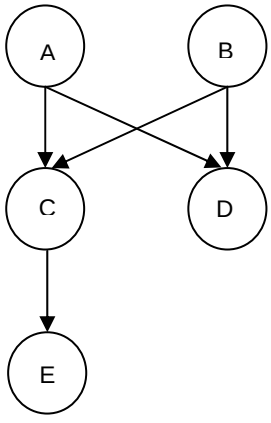
Early applications of Bayesian belief networks (BBNs) were in medical diagnosis and genetics, but recently there has been an explosion in their use, including for environmental impact assessment, tracing faults in computer systems and software, robotics and many other areas (Jensen, 2001; Cowell *et al*, 1999). A growing area of interest is the management of natural resources under uncertainty. For example, a BBN model was developed for assessing the impacts of land use changes on bull trout populations in the USA (Lee, 2000). Another recent application of BBNs is modelling uncertainties in fish stock assessment and the impact of seal culling on fish stocks (Hammond & O'Brien, 2001). Marcot et al. (2001) have used BBNs for evaluating population viability under different land management alternatives, while Wisdom *et al* (2002) used BBNs in conservation planning for the greater sage-grouse. BBNs have been used for investigating factors associated with adaptive co-management of artisanal fisheries (Halls & Burn, 2002).

Why Network Models?

Traditional statistical modelling defines and builds models for a response (outcome) in terms of sets of explanatory variables (attributes). Each explanatory variable in a model is seen as *directly* impacting on the response variable. With explanatory variables x_1, x_2, …, x_p, and response y, the situation can be represented by the following diagram:



In reality, however, it can happen that the relationships between variables are not as simple as this model allows. The effect of one x-variable on the response y may be mediated through another x-variable, or through two or even more x-variables. It could also happen that some of the x-variables affect some of the others. Indeed, with datasets containing many variables, it is easy to envisage quite complex patterns of association. The roles of 'response' and 'explanatory' become blurred, with variables taking on each role in turn.



In the simple example shown here, variables *E* and *D* could be regarded as 'responses', and *A* and *B* as 'explanatory'. But *C* seems to play both roles. It looks like a response with *A* and *B* acting as explanatory variables, and it is an 'explanatory' variable for *E*. The variables are modelled as random variables and the links are probabilistic. A link from *A* to *C* would be interpreted as meaning that the value of *A* affects the value of *C* by means of influencing the probability distribution of *C*.

Historically, these models evolved largely in the artificial intelligence (AI) community, and form the basis of *expert systems*. Generally they are not tools for statistical inference but rather they are mechanisms for encoding probabilistic causal relationships and making predictions from them. Because of their AI background, it is not surprising that the

current terminology of network models is quite different from statistical jargon, and is perhaps less familiar. Sometimes there is an exact correspondence between an AI term and a statistical one, the two terms being different names for the same concept.

Bayesian Networks

The general class of models that we will use consist of a number of *nodes* (random variables) connected by *directed* links. A node which has a directed link leading from it to another node is called a *parent* node and the second one is a *child* node. Cycles are not permitted: that is, it is not possible to start from any node and, following the directed links, end up on the same node.

A model with these properties, after specifying the probabilities which govern the links, is called a *Bayesian belief network* (BBN), or just a Bayesian network. Most of the currently available software for building and analysing BBNs requires that the nodes are discrete, taking only a finite set of possible values, and we assume this to be the case in what follows. Continuous variables can be accommodated by grouping their values into class intervals. An introductory account of BBNs is given by Jensen (2001) while a more rigorous and complete treatment is Cowell *et al* (1999).

To explain the basic ideas, consider the simple example above. For simplicity, assume that all of the nodes are binary variables, taking values T or F (true or false). The probabilistic mechanism which governs the relationship between, say, *E* and its parent *C* is the *conditional probability distribution* of *E* given *C*. This can be expressed as a table:

	E|*C*		
C	*F*	*T*	*Sum*
F	p_{00}	p_{01}	*1*
T	p_{10}	p_{11}	*1*

The table of conditional probabilities for node *C*, which has parents *A* and *B* would have the following form:

		C|*A*,*B*		
A	*B*	*F*	*T*	*Sum*
F	*F*	p_{000}	p_{001}	*1*
F	*T*	p_{010}	p_{011}	*1*
T	*F*	p_{100}	p_{101}	*1*
T	*T*	p_{110}	p_{111}	*1*

A node with no parents (*A* or *B* in the example) would have just a *prior* probability table:

A		
F	*T*	*Sum*
p_0	p_1	*1*

The complete specification of a BBN consists of

(a) the set of nodes,
(b) the directed causal links between the nodes,
(c) the tables of conditional probabilities for each node.

Estimating the Conditional Probabilities

In practice, there are several possible ways of obtaining estimates for the conditional (and prior) probabilities. If sufficient data are available then cross-tabulating each node with its parents should produce the estimates. There are alternatives to deriving the probabilities from data, however. It is possible to use *subjective* probabilities or *degrees of belief*, usually encoded from expert opinions. In many of the early applications of BBNs in medical diagnosis this was generally the approach that was used. There has been some recent research into developing systematic ways of *eliciting* prior beliefs from experts and building probability distributions from them (O'Hagan, 1998).

Evidence and Updating

In the simple example above, if the states of the nodes (i.e. the values of the variables) *A* and *B* were known, then it would be possible to use the rules of probability to calculate the probabilities of the various combinations of values of the other nodes in the network. This kind of reasoning in a BBN can be called 'prior to posterior', in the sense that the reasoning follows the directions of the causal links in the network. Suppose now that the state of node *E* were known. What could be said about the other nodes? The *updating algorithm* of Lauritzen and Spiegelhalter (1998) allows us to calculate the posterior probabilities of all other nodes in the network (and this works for *any* BBN), given the known value at *E*, or indeed, given any combination of known nodes. In the jargon of expert systems, 'knowing' the value of a node is called 'entering evidence'. This is 'posterior to prior' reasoning and allows us to infer something about the states of nodes by reasoning *against* the direction of the causal links. The updating algorithm is a very powerful tool in BBNs and enables us to make useful predictions and examine 'what if' scenarios with ease. Various software packages are available which facilitate the construction of BBNs and implement the updating algorithm. For this project, we propose the program Netica (Norsys, 1998).

References

Augustin N.H., Borchers D.L., Clarke E.D., Buckland S.T.., Walsh M. (1998). Spatiotemporal modelling for the annual egg production method of stock assessment using generalized additive models. *Can. J. Fish. Aquat. Sci.* 55, 2608-2621.

Borchers D.L., Buckland S.T., Zucchini, W. (2002). *Estimating Animal Abundance.* Springer.

Buckland S.T., Burnham K.P., Augustin N.H. (1997). Model selection: an integral part of inference, *Biometrics*, 53, 603-618.

Buckland S.T., Anderson D.R., Burnham K.P., Laake J.L., Borchers D.L., Thomas, L. (2001) *Introduction to Distance Sampling.* Oxford. Further material and DISTANCE software available from http://www.ruwpa.st-and.ac.uk/distance/.

Burnham K.P., Anderson D.R. (2002). *Model Selection and Multimodel Inference (2nd edition).* Springer, New York.

Cowell R.G., Dawid A.P., Lauritzen S.L., Spiegelhalter D.J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Diggle P.J., Kung-Lee Yiang, Zeger S.L. (1994) *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Goldstein H. (1995) *Multilevel Statistical Models (second edition).* Arnold, London.

Green, R.H. (1994). Aspects of power analysis in environmental monitoring; in Fletcher D.J. & Manly B.F.J. (Ed.) - *Statistics in Ecology and Environmental Monitoring*, Univ. of Otago Press, Dunedin, N.Z.

Halls A.S., Burn R.W. (2002). *Interdisciplinary Multivariate Analysis for Adaptive Co-management.* DFID Technical Report for Project R7834.

Hammond T.R., C.M. O'Brien. (2001). An application of the Bayesian approach to stock assessment model uncertainty. *ICES J. Marine Science* **58**, 648-656.

Hastie T., Tibshirani R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Jachmann H. (1998). *Monitoring illegal wildlife use and law enforcement in African savanna rangeland*. Wildlife Resources Monitoring Unit, Environmental Council of Zambia.

Jensen F.V. (2001). *Bayesian Networks and Decision Graphs.* Springer, New York.

Krzanowski W.J. (1988). *Multivariate Analysis.* Oxford University Press.

Laing S.E., Buckland S.T., Burn R.W., Lambie D., Amphlett, A. (2003). Dung and nest surveys: estimating decay rates. *J. Appl. Ecology* **40**, 1102-1111.

Lauritzen S.L., Spiegelhalter D.J. (1998). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Stat. Soc.B*, **50**, 157-224.

Lee D.C. (2000). Assessing land-use impacts on bull trout using Bayesian belief networks, in Ferson, F., Burgman M. *Quantitative Methods in Conservation Biology*, Springer, New York.

Marcot, B. G., R. S. Holthausen, M. G. Raphael, M. Rowland, M. Wisdom (2001). Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management* **153**, 29-42.

Milliken T., Burn R.W., Sangalakula L. (2002). An analysis of trends of elephant product seizure data in ETIS. *CITES CoP12, Doc. 34.1*.

NAG (2003). *Genstat*. Numerical Algorithms Group, Oxford. www.nag.co.uk/stats/tt_soft.asp or www.vsn-intl.com/genstat/

Norsys Software Corp. (1998). *Netica*. www.norsys.com/netica

O'Hagan A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician* **47** Part 1, 21-35.

Thompson, S.K. (2002). *Sampling (2nd edition)*. Wiley.

Wisdom, M.J., Wales, B.C., Rowland, M.M., Raphael, M.G., Holthausen, R.S., Rich, T.D., Saab, V.A. (2002). Performance of Greater Sage-Grouse models for conservation assessment in the Interior Columbia Basin, USA. *Conservation Biology* **16**, 1232-1242.