Report Cover Page

ACERA Project
0807
Title
Alternative methodologies for establishing pest and disease freedom
Author(s) / Address(es)
Greg Hood, Bureau of Rural Sciences, Department of Agriculture Fisheries and Forestry, Canberra, ACT Tony Martin, Department of Agriculture and Food, Bunbury Western Australia Simon Barry, CSIRO Mathematical and Information Sciences, Canberra, ACT
Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)
Manuscript
Summary
This report presents a technical development. This development is important because it simplifies some recent advances in scenario tree techniques that are being used increasingly, especially in animal disease management, to substantiate claims of disease freedom. Scenario trees are a new and popular method by which surveillance systems can be evaluated. One of their limitations is that for relatively complicated systems, such as surveillance systems for animal diseases that use multiple sources of information, the scenario trees can become large and complex. This report demonstrates how to simplify large scenario trees by 'pruning' redundant branches. It also shows that in many situations, scenario trees can be presented as Bayesian networks, thereby taking advantage of a considerable body of knowledge, software and expertise for building and testing systems that support claims of disease freedom.

ACERA Use only	Received By:	Date:
	ACERA / AMSI SAC Approval:	Date:
	DAFF Endorsement: ( ) Yes ( ) No	Date:

Alternative methodologies for establishing pest and disease freedom: ACERA 0807

Greg Hood; Bureau of Rural Sciences
Department of Agriculture, Fisheries and Forestry

Manuscript

Friday, 12 June 2009

Acknowledgements

This report is a product of the Australian Centre of Excellence for Risk Analysis (ACERA). In preparing this report, the authors acknowledge the financial and other support provided by the Department of Agriculture, Fisheries and Forestry (DAFF), the University of Melbourne, Australian Mathematical Sciences Institute (AMSI) and Australian Research Centre for Urban Ecology (ARCUE).

Disclaimer

This report has been prepared by consultants for the Australian Centre of Excellence for Risk Analysis (ACERA) and the views expressed do not necessarily reflect those of ACERA. ACERA cannot guarantee the accuracy of the report, and does not accept liability for any loss or damage incurred as a result of relying on its accuracy.

Table of contents

List of Tables

List of Figures

1. Executive Summary

Stochastic scenario trees are a new and popular method by which surveillance systems can be analyzed to demonstrate freedom from pests and disease. For multiple component systems—such as a combination of a serological survey and systematically collected observations—it can be difficult to represent the complete system in a tree because many branches are required to represent complex conditional relationships.

Here we show that many of the branches of some scenario trees have identical outcomes and are therefore redundant. We demonstrate how to prune branches and derive compact representations of scenario trees using matrix algebra and Bayesian belief networks. The Bayesian network representation is particularly useful for calculation and exposition. It therefore provides a firm basis for arguing disease freedom in international fora

2. Introduction

Under the Sanitary and Phytosanitary Agreement of the World Trade Organization (WTO), measures taken to protect animal, plant or human health must be scientifically justified and supported by evidence [1]. Countries which seek to impose sanitary measures can demonstrate freedom from particular pests, weeds or diseases (the word *disease* is used hereafter in a generic sense to refer to a pest, weed or disease) using a surveillance system, which may comprise a range of ongoing observations and tests—for example, observations by qualified personnel and structured serological surveys. Likewise, countries which wish to export commodities can use a surveillance system to demonstrate freedom from disease in the sources of those commodities.

In many cases, however, even if signs of a disease have not been seen for many years, we cannot be confident that the disease is truly absent because introduction of the disease could be recent and most surveillance systems are not 100% sensitive. To maximize confidence in freedom from a disease, countries can use surveillance systems with multiple (preferably independent) components to increase the conditional probability of detection (that is, the probability of detection assuming the disease is present). In the animal health literature [2], demonstrating freedom from disease by calculating the probability of a disease being present across multiple components of a surveillance system has been a hot topic since the WTO began operations in 1995. Recently, Martin *et al.* [3, 4] presented a general methodology to analyse and support claims of freedom in complex surveillance systems using stochastic scenario tree models. The method assumes that if a disease is present in a country it must be present at some minimal prevalence when the surveillance is conducted. This is called the design prevalence, P^*. For a single component of a surveillance system—such as a serological survey—the method partitions the reference population into groups within which all units have the same probability of being detected as diseased (assuming that the population is infected at the design prevalence). This division of the population allows calculation of the sensitivity of surveillance in both representative sampling schemes and targeted schemes (in which the sampling units are not independent). For multiple component systems, such as surveillance by field and abattoir sampling, the sensitivity of the combined components is estimated by serial application of Bayes theorem.

Analysis of multiple-component systems using scenario trees is flexible, provides a quantitative estimate of the probability of freedom, can account for lack of independence between components, and allows incorporation of historical data. Scenario tree diagrams are also promoted by the World Organisation for Animal Health [5, 6] since they provide a clear and logical presentation of pathways, clarify ideas and assist in communicating the results of a risk analysis. The stochastic scenario tree methodology, therefore, has many advantages over traditional alternatives such as statistical analysis of structured surveys (which are expensive and difficult to implement) and qualitative assessments (which are subjective and not easily repeated). Scenario trees are also computationally simple and relatively easy to review compared to recent approaches using simulation models to evaluate the effectiveness of surveillance systems [7, 8].

But the method has some drawbacks. For complex surveillance systems, one or more (potentially large) scenario trees must be constructed using spreadsheets or other

computer software. Unless specialist tools are used, implementation of large trees in software is prone to calculation errors and can be difficult to audit. This diminishes their value for risk communication—defined under international standards as the open, interactive, iterative and transparent exchange of information [7]. As we shall see in the following case studies, the scenario tree format can also obscure conditional relationships amongst risk factors, especially in complex models.

To improve risk communication, we present alternative methods for analyzing multiple-component surveillance systems. The methods are analogues of the scenario tree method—providing the same results in all cases—but can be simpler to implement using current software, and also provide alternative views of the system that are relatively easy to interpret and audit. The first method collapses the redundancy inherent in the multiple limbs of scenario trees and uses matrix algebra to simplify calculation of the sensitivity of each component of a surveillance system. The chief advantage of this method is that it produces a compact representation of the complete surveillance system and eases calculation of component sensitivities—though some practitioners would argue that the scenario tree representation is simpler to interpret.

The second method reconstructs a scenario tree as a Bayesian belief network (BBN). BBNs are widely used for machine learning [8] and are increasingly used for modelling ecological systems [9, 10], but have received little attention in the surveillance literature. We will show that a BBN provides a compact diagram of the structure of a surveillance program, simplifies calculations, and extends the range of software that can be used for analysis. The sometimes tedious calculations required for analysis of surveillance systems can be performed by simply updating beliefs in components of the network using readily available software.

In what follows, we introduce some key terminology and concepts from Martin *et al.* [3, 4] and provide brief introductions to each of the new methods. We then use each method to analyse both a simple contrived example and the real example of Danish surveillance for classical swine fever described in detail by Martin *et al.* [3]. To facilitate applications of the scenario tree methodology and its analogues outside existing veterinary applications, we provide tools for implementing the matrix method in spreadsheets, example code in the R statistical language [11], and examples of belief networks in various formats in the online appendices.

3. Key terminology and concepts

In general, we follow the terminology introduced by Martin *et al.* [4], who consider that each surveillance system component, *SSC*, targets individuals (units) within a reference population (sampling frame). Tests or observations of individuals are defined as having a characteristic sensitivity *Se* (the probability that an infected individual yields a positive test) and perfect specificity (that is, the probability that a negative test result comes from a truly negative individual is one). Perfect specificity implies that all steps are taken to ensure that any positive results are thoroughly investigated to confirm or deny the (possibly false-positive) result—an assumption that simplifies the analysis but is not strictly necessary.

A key quantity in the scenario tree method is the component sensitivity, *CSe*, the probability that the *SSC* would give a positive outcome given the individuals tested and that the reference population is infected at the design prevalence, *P**. The purpose of the scenario tree is to partition the sampling frame into groups with a homogeneous history (where, for surveillance systems, history is defined as all factors that can affect the outcome of a test or observation). This allows calculation of the probability that a unit would yield a positive test at *P** for each of the terminal nodes. All dependence between units—such as membership of the same region—is incorporated into the tree so that, given *N* terminal nodes, *CSe* can be calculated as

$$CSe = 1 - \prod_{i=1}^{N} \Pr(T+)_i \,, \tag{1}$$

where $\Pr(T+)_i$ is the probability of a positive test in the *i*th terminal node. For a system with *J* independent components, the sensitivity of the system can be calculated as

$$SSe = 1 - \prod_{j=1}^{J} (1 - CSe_j) \,. \tag{2}$$

When the system components are not independent, *SSe* is calculated by sequential application of Bayes theorem, as described by Martin *et al* [4]. And given *SSe*, Bayes theorem likewise yields a point estimate of the posterior probability of freedom given a negative surveillance outcome and some prior probability of disease being present in the country. Martin *et al* [4] extend the approach to account for sequential sampling schemes and uncertainty in the parameters—issues not considered here since the only differences in the methodologies lie in the calculation of sensitivities.

4. Matrix method

Because the purpose of a scenario tree is to partition all units in the sampling space into groups with a similar history, each node in a scenario tree represents a historical state with the Markov property that the conditional probability distribution of future states depends only upon the present state. The properties of a scenario tree can therefore be computed by representing the tree as a Markov chain [12]. To do so, we construct a transition matrix, each row of which forms a probability distribution of future outcomes for each state (it is row stochastic). The transition probabilities in this matrix correspond to the branch probabilities of an equivalent scenario tree representation.

Consider the simplest surveillance application we can contrive, in which samples from an infected population are subject to a single test with sensitivity, *Se*. A scenario tree representing this system has three nodes representing the three possible states of a sample: (1) the undifferentiated state before testing is completed, (2) a positive and (3) a negative outcome (Figure 1).

[Figure 1 hereabouts]

The corresponding transition matrix in canonical form is:

$$\mathbf{P} = \begin{array}{c} \\ \text{from} \end{array} \begin{array}{c} \begin{array}{ccc} & \text{to} & \\ \end{array} \\ \begin{bmatrix} 0 & Se & 1-Se \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}, \tag{3}$$

which shows that the Markov chain has two absorbing states corresponding to positive and negative outcomes of the test. The chain reaches absorption after just one step after which the samples have been partitioned into those with positive and those with negative outcomes. Using the distribution vector $\mathbf{v}_0$ = [100 0 0] (representing the initial state of the chain with all samples in the undifferentiated state), the distribution after one step can be calculated using the matrix expression

$$\mathbf{v}_1 = \mathbf{v}_0\mathbf{P}\,. \tag{4}$$

Given, for example, 100 samples with *Se* = 0.95, we obtain the expected distribution $\mathbf{v}_1$ = [0 95 5]. And because absorption is reached with one step, we also have $\mathbf{v}_\infty$ = [0 95 5].

Note that, since the state of the system after partitioning into the absorbing positive and negative outcomes is of no interest, we can ignore subsequent outcomes using the simplified transition matrix **P** = [0, 0.95, 0.05]. In the node diagram, this would be represented by omitting the two "self" loops.

In this trivial example, the formulation of a scenario tree as a Markov chain has no obvious utility. Benefits arise in more complicated systems in which a reduced transition matrix can be formulated which retains the behavior of the full system. This simplification is achieved using a branch pruning algorithm similar to that used for binary

decision trees [13]. We illustrate this approach fully in the case studies below. In brief, the method entails (1) pruning of outcomes having no interest (negative test results), and (2) pooling states which have different prior histories but identical distributions of future outcomes. These operations yield a simplified matrix amenable to manipulation and inference using spreadsheets or any software that allows matrix multiplication.

For representative sampling schemes—where the proportions of units processed falling into each of the categories specified in the tree are the same as their proportions in the reference population—we can calculate the probability that any randomly selected unit in the population will give a positive outcome, *CSeU*, using the matrix product $\mathbf{v}_0\mathbf{P}^k$, where k is number of steps to absorption in the Markov chain described by matrix $\mathbf{P}$. For a single sample, the initial state has zeroes everywhere except for its first element and we therefore need only compute the matrix power $\mathbf{P}^k$. Because the outcomes of the chain (and the scenario tree) are mutually exclusive, the probability of a positive result is the sum of those elements of $\mathbf{P}^k$ representing positive outcomes. To see this, notice that for N samples drawn from a reference population, each cycle of the Markov chain represents a partitioning of the N samples according to their disease risk or test response category. The number of such partitionings will equal the number of branching levels in the equivalent scenario tree representation. Given a k-level scenario tree, a complete partitioning of N samples is conducted by completing k cycles of the corresponding Markov chain—these operations being equivalent to equations 2 and 3 of Martin *et al.* [4]. The overall component sensitivity, *CSe*, is then calculated using equation 4 of Martin *et al.* [4]. That is:

$$CSe = 1 - (1 - CSeU)^n. \tag{5}$$

For targeted sampling schemes, in which the units processed are not independent, the matrix representation of scenario trees is not so useful for calculation of *CSe* because a new transition matrix must be prepared for each unique set of risk and test outcomes—a task best left to computer code.

5. Bayesian belief networks

The second methodology can simplify calculation of *CSe* for both representative and targeted sampling schemes. The conditional independence of nodes in scenario trees means that they can be represented using Bayesian belief networks (BBN). A BBN is a probabilistic graphical model in which nodes represent variables and arcs encode conditional independence. For surveillance systems, a BBN can be constructed *de novo*, but can also be formed as a translation of a scenario tree representation following the same simplification steps used for representing a scenario tree as a transition matrix. One benefit from such a translation is that commercial and open-source software for analysing BBNs is readily available, and, once constructed, computation of *CSeU* for representative sampling schemes merely requires compilation of the network to obtain the probability of a positive result. The capacity for belief updating in BBN software permits straightforward analysis of sampling schemes in which the units processed are not independent. For each sample, the known information (e.g. risk categories from which the sample was drawn and the result of the test—which will always be negative in our case) is updated in the BBN. The posterior probabilities of infection at the unit and higher (e.g. herd) levels are then calculated by belief updating (probabilistic inference). Using the ensemble of posterior probabilities derived from all processed units, calculations like equation 5 then provide estimates of *CSe*.

We now illustrate the matrix and BBN methods using two case studies: the first is a contrived example of surveillance for Foot and Mouth Disease (FMD) in an island nation; the second is the more complex case study of Martin *et al.* [(4)] which computes the sensitivity of Danish surveillance for classical swine fever (CSF) using serological tests collected at abattoirs.

6. Case Study 1. FMD surveillance

Consider a country stratified into two cattle rearing zones. One zone (*Low*, comprising 70% of the country's area) is far from neighboring countries and the risk of FMD infection is low; the other (*High*) is close to neighboring countries where FMD is endemic and the risk of infection in that zone is judged by expert opinion to be high (say, *RR_High* = 5). We assume all livestock are owned by smallholders, so that there are no large herds and cattle are distributed evenly across the zones. Experts believe that the prevalence of infection, if established, would be at least $P^*_A = 0.01$ and have embarked on a program to sample cattle in each zone using an enzyme-linked immunosorbent assay with a sensitivity of *ELISASe* = 0.95. All positive tests are thoroughly investigated and so we assume the specificity of the test is 100%. A summary of fundamental parameters and derived quantities appears as Table I.

[Table I hereabouts]

[Figure 2 hereabouts]

6.1 Matrix formulation

The full scenario tree for this study (Figure 2a.) shows that the surveillance program has three classification levels for each surveillance outcome—zone status, animal status and test result. To construct a simplified transition matrix, we first observe that negative outcomes are of no interest and that uninfected cattle do not produce positive outcomes. Branches that lead to wholly negative outcomes can therefore be pruned. Second, since zone status affects the prevalence of infection but not the result of the ELISA, we can further simplify the model by pooling animals from both zones—after calculating the relative proportions of infected animals in each zone. Application of these two simplifying steps results in the node diagram shown as Figure 2b, in which nodes are numbered according to the row numbers of the transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & Pr\_High & 1-Pr\_High & 0 & 0 \\ 0 & 0 & 0 & EPIA\_Low & 0 \\ 0 & 0 & 0 & EPIA\_High & 0 \\ 0 & 0 & 0 & 0 & ELISASe \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \qquad (6)$$

The elements of **P** are transition probabilities (Table I) and their values are reflected in the width of the arrows joining nodes in Figure 2b. *Pr_High* (the proportion of animals in the *High* zone) and *ELISASe* are given *a priori*; the effective probabilities of infection (*EPIA_Low* and *EPIA_High*) are calculated for the low and high risk zones using the method discussed in Martin *et al.* [4]. The sequence of calculations is:

$$P^*_A = 0.01,$$

$$RR\_High = 5,$$

$$AR\_Low = 1 / (RR\_High \times Pr\_High + 1 - Pr\_High),$$

$$AR\_High = AR\_Low \times RR\_High,$$

$$EPIA\_Low = AR\_Low \times P^*_A,$$

$$EPIA\_High = AR\_High \times P^*_A,$$

where *AR_Low* and *AR_High* are the adjusted relative risks in the *Low* and *High* zones, where "adjusted" means that the relative risks are "adjusted to ensure that the (weighted) average adjusted risk for the section of the *SSC* reference population represented by the risk node is 1" [(4)]. For simplicity and generality these calculations can be represented in matrix format. If we extract the submatrix **r** as a column vector of relative risks for each zone, and construct **c** as a row vector of the proportion of animals by zone, then

$$EPIA_i = \frac{r_i P_A^*}{\sum_i c_i r_i}, \; i = 1 \ldots 2, \tag{7}$$

and so the effective probabilities of infection which appear as elements in **P** is

$$\boldsymbol{EPIA} = P^*_A \, \mathbf{r(c\ r)–1}. \tag{8}$$

and there is no need to calculate adjusted risks as an intermediate step.

The node diagram shows that there are three classification steps. Hence, given matrix **P**, the sensitivity of the surveillance system can be calculated using

$$\mathbf{P}^3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.0095 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{9}$$

The component unit sensitivity is $CSeU = \mathbf{P}^3(1, s) = 0.0095$, where s is the dimension of **P**. The surveillance system sensitivity (assuming independence) is then calculated as

$$SSC = 1 - (1 - 0.0095)^n, \tag{10}$$

where n is the number of samples tested.

In this particular case, there is no particular advantage in developing the matrix formulation other than that provided by the relative simplicity of the associated node

diagram (Figure 2b). Indeed, because we consider only positive outcomes, and the effective probabilities of infection must conform to the design prevalence, *CSeU* can be readily computed as the product

$$CSeU = P^*_A \,.\, Se = 0.0095. \quad (11)$$

A feature of the matrix formulation, however, is that it clearly shows that *CSeU* is linear in the parameters of equation 6, and so variation in these parameters has no effect on the expected value of *CSeU*. Hence, if the mean of *CSeU* is the only parameter of interest, simulation is not required for its estimation.

Code for calculation of *CSeU* in this case study is provided as an Appendix.

6.2 Bayesian belief network representation for the FMD study

Structurally, the BBN for the FMD case study is trivial (Figure 3a). A *Zone* node displays the probability that a sample derives from the *High* or *Low* zone; a *Disease* node enforces the design prevalence; and a *Test result* node allows for the sensitivity of the ELISA test. Some care needs to taken to populate the conditional probability tables (CPTs) of the nodes: the CPT of the single marginal node (*Zone*) is populated using *Pr_High* = 0.3 (and 1 – *Pr_High* = 0.7); the CPT of the *Test result* node is populated using *ELISASe* and our assumption that *ELISASp* = 1 (i.e. that the specificity is 100%); and the CPT of the *Disease* node is populated using equation 8.

[Figure 3 hereabouts]

The value of the BBN formulation lies in the clear presentation of dependency and outcomes in Figure 3. As the arrows imply, the presence of disease depends only on the zone and the result of testing depends only on the disease status. A check that *EPIA_Low* and *EPIA_High* have been calculated correctly is provided by the *Disease* node which (when the network is evaluated) displays our belief that a randomly selected sample comes from a diseased animal, $P^*_A = 0.01$. The value we seek, *CSeU*, is displayed in the *Test result* node as the probability of the state *Positive*, from which we derive *CSe* using equation 5.

The structural simplicity of a BBN makes extensions of the basic model simple. For example, chronic diseases such as paratuberculosis of cattle are more likely to be found in older animals, and so the design prevalence should be distributed across age classes [(14)]. Accounting for zone-specific age distributions and age-specific probabilities of infection in a scenario tree version of the FMD model would require *az* branches to represent the zone and age classes (where *a* is the number of age classes and *z* is the number of zones). By contrast, a BBN version requires just one extra node (Figure 3b). In both versions, recursive application of equations like 8 can provide the effective probabilities of infection.

Our next case study highlights both the expository value of the alternative formulations and the use of the BBN for calculating *CSe* for both representative and targeted sampling schemes.

7. Case study 2. Danish CSF

This case study is based on a real example of Danish surveillance for CSF. Here we compare the scenario tree analysis of Martin *et al.* [3] with analyses using a transition matrix and a BBN. A complete description of the system is provided by Martin *et al.* [3], including a depiction of the full scenario tree (their Figure 2). In brief, the surveillance system uses a serological test of blood samples collected from adult pigs at abattoirs. The samples are classified by county of origin, herd status and farm type. We focus here on calculation of *CSe* for this system.

6.3 Matrix formulation for the Danish CSF study

Construction of a simplified node diagram—using the branch pruning algorithm—yields Figure 4. The node diagram shows that the expected proportion of positive samples at the design prevalence can be partitioned from the complete set of samples in six steps.

1. Herds are selected from either South Jutland (high risk) or other (low risk) counties.
2. Of these, a proportion of herds from each zone could be expected to be infected given herd prevalence P^*_H and the relative risk of infection in each zone.
3. Infected herds may be breeder or slaughter herds. The proportion of each type differs between zones.
4. All pigs in slaughter herds are growers, but breeder herds have both adults and growers and the proportion of each differs between the two zones.
5. Blood samples are taken from adult and grower pigs which have different risks of infection. The design animal prevalence, P^*_A, and relative risk of infection allow us to determine how many of each age class would be infected.
6. A proportion (*ELISASe*) of all blood samples from infected pigs return positive results. This proportion does not depend on age class.

The nodes of Figure 4 are numbered sequentially left to right and top to bottom with numbers corresponding to the entries in the description of the transition matrix, **A** (Table II).

[Figure 4 hereabouts]

[Table II hereabouts]

Most of the transition probabilities are calculated following the methods in Martin *et al* [3], but our matrix formalism provides a shortcut for calculating the effective probabilities of infection for each age class. Slaughter herds have only growers and so the effective probability of infection for them is P^*_A. In breeder herds, the effective probabilities of infection depend on the design animal prevalence, P^*_A, and the number of pigs of each class in this type of herd. We are given P^*_A, but the number of pigs in each age class must be calculated using the earlier transition probabilities. This is most easily achieved

by computing the matrix power $\mathbf{P}_r^4$, where $\mathbf{P}_r$ is the reduced matrix obtained by removing the last two rows and columns of the full transition matrix, **P**. We then obtain the proportion of samples of these classes using columns nine and ten of row one of $\mathbf{P}_r^4$ (since these nodes represent the proportions in each age class after the earlier classification steps). This yields proportions **c** = [0.42, 0.58] and thence, using equation 8 with **r** = [5, 1], effective probabilities of infection = [0.0933, 0.0187] (Table II). Finally, using the completed matrix, we compute $\mathbf{P}^6$ to obtain a *CSeU* for representative sampling of 0.000475.

As in the previous example, the matrix formulation (displayed as a node diagram, Figure 4) provides a relatively simple overview of the sampling scheme. The diagram is complete, all parameters contributing to the calculations are displayed, and derived and fundamental (where fundamental refers to assumed or externally estimated values) parameters are clearly distinguished.

Bayesian belief network representation for the Danish CSF study

The BBN for the CSF study is presented as a network diagram in Figure 5a. The unshaded nodes indicate states of the population of samples, while the shaded *Sampling scheme* node indicates the choice of either *Representative* or *Targeted* sampling. When the *Sampling scheme* is *Representative* (as shown in the figure) the probabilities of each state in the *County*, *Age* and *Herd type* nodes (Table III) match the expected proportions for non-targeted sampling of the Danish pig population (see Table II). When the *Sampling scheme* node is set to *Targeted*, the CPTs ensure that only adult breeders are sampled and the probability of a sample being from South Jutland matches the observed proportion (Figure 5b).

[Figure 5 hereabouts]

[Table III hereabouts]

County and *Age* CPTs are populated directly using Table II, and the *Herd status* CPT is populated using equation 8. Setting values of the remaining CPTs is relatively complicated and is dealt with below. For now, we note that the display of the network in the Netica™ software [15] (Figure 5) provides a rough check that these calculations have been performed correctly: for the state *Infected*, the *Herd status* node displays the herd design prevalence, P^*_H, and the *Animal status* node displays (as a percentage) $P^*_H.P^*_A$ = 0.0005.

To set the CPT of the *Animal status* node, we first set *Sampling scheme* to *Representative* and *Herd status* to *Infected*. We then set the belief in *Herd type* to *Breeder* and compiled the network to obtain the distribution amongst age classes. This yields the same proportions obtained using the matrix method, **c** = [0.42, 0.58] and we then use equation 8 to obtain effective probabilities of infection $\boldsymbol{EPIA}_{\text{breeder}}$ = [0.0933, 0.0187]. Repeating the process after setting *Herd type* to *Slaughter* yields $\boldsymbol{EPIA}_{\text{slaughter}}$ = [0.25, 0.05] (see Table IV). Notice that setting beliefs in the BBN reflects the assumption in Martin *et al.* [3] that the prevalence in a herd—given that it is infected—is P^*_A regardless of the

herd type. The differences between the effective probabilities of infection between *Breeder* and *Slaughter* herds arise because of differences between age distributions of the two *Herd types* and the requirement to jointly satisfy the relative risks and design prevalence, P^*_A. As the arrow from *Herd type* to *Animal status* in Figure 5 implies, under this assumption *Animal status* is only conditionally independent of *Herd type.*

[Table IV hereabouts]

An alternative analysis is possible by removing this relationship, reflecting a new assumption that *Herd type* affects *Animal status* only through differences in *Age class*. For representative sampling, this yields the *Age classes* **r** = [0.204, 0.796] (Adults, Growers), from which we obtain ***EPI*** = [0.138, 0.0276]. While preserving the overall animal-level design prevalence, this assumption means that the level of infection is assumed to be higher in *Breeder* herds which contain more susceptible (*Adult*) animals. This increases the estimate of the sensitivity of the surveillance system—which happens to target *Breeder* herds. This assumption may be realistic, but the decision to allow variation in P^*_A at this level will often be dictated by international standards and the requirements of trading partners[4].

The final step is to calculate the *CSe* for the targeted sampling scheme. For this we require two quantities for each herd:

1. SeH_i: The herd sensitivity is the probability that one or more positive outcomes will be obtained when n_i animals are sampled and the herd is infected at the (animal-level) design prevalence P^*_A.
2. $EPIH_j$: The effective probability of infection of a herd in county *j*.

With the BBN representation of the sampling scheme, the conditional probability of any particular state can be calculated by belief updating. Given two events A and B where A precedes or causes B, our belief in B given knowledge of A is

$$\Pr(B) = \Pr(A).\Pr(B|A), \tag{12}$$

while knowledge of B gives us

$$\Pr(A) = \Pr(B)/\Pr(B|A). \tag{13}$$

An algorithm to calculate SeH_i and $EPIH_j$ for each herd can therefore be devised using two applications of 12:

1. Set the state of the *Sampling scheme* node to *Targeted*
2. Set the state of the *County* node to the county of origin of the herd (*South Jutland* or *Other*)
3. Read $EPIH_j$ as the belief that the *Herd status* is *Infected* (given that sampling is targeted and the county is known)
4. Now set the *Herd status* node to *Infected*

5. Read the probability that an individual sample yields a positive result in an infected herd in county *j* ($PUPos_j$) as the belief that *Serology* is *Positive* (given that sampling is targeted, the county is known and the herd is infected)
6. Set SeH_i equal to $1 - (1 - PUPos_j)^n$.

As shown in Martin *et al.* [3], the prior probability of a negative herd test is then

$$\Pr(\text{Herd } i \text{ in county } j \text{ is negative}) = 1 - SeH_i.EPI_j, \quad (14)$$

and the *CSe* for the targeted sampling scheme is

$$CSe = 1 - \prod_{j=1}^{2} \prod_{i=1}^{I_j} (1 - EPIH_j.SeH_i), \quad (15)$$

where I_j is the number of herds in the county from which samples have been processed.

The calculations above use knowledge of causes to infer probabilities of an effect (equation 12); the reverse calculation (13) is most useful when more than one sampling scheme targets a sampling unit or cluster. For joint sampling of, say, a herd we must calculate the posterior probability of a negative test in the first sampling scheme (see section 5.2. in [4]). In the BBN for CSF this is easily achieved by setting node states appropriately to account for ancillary information (e.g. *County*, *Age*, *Herd type*) and setting the *Serology* node to *Negative*. This yields the posterior estimate of $EPIH_j$ in the *Herd status* node (Figure 5b). If more than one sample is collected from a herd, we apply Bayesian revision—that is, the posterior estimate of $EPIH_j$ from each sample becomes the prior $EPIH_j$ for the next sample.

8. Discussion

The methods presented here provide compact overviews of surveillance systems and expand the range of tools that can be used for their analysis. The matrix method yields a simple diagram and provides a useful method for calculating the effective probabilities of infection for programmers and statisticians. Formulation as a matrix permits some kinds of automated analysis. Figure 2b, for example, was constructed (semi) automatically using an existing routine to draw diagrams based on demographic matrices [16]. The requirement to transcribe probabilities into the correct location in a transition matrix, however, can make implementation tedious, prone to errors, and difficult to audit.

The BBN is a more generally useful formulation. BBN node diagrams provide a compact representation of surveillance systems, such that the structure can be understood at a glance. The node diagram required for CSF surveillance (Figure 5) requires just seven nodes to represent the complex scenario tree equivalent. Indeed, the BBN version is actually a superset of the scenario tree shown as Figure 2 in Martin *et al.* [3] since the *Sampling scheme* node represents the cloning of a complete tree to model a targeted sampling scheme. As Figure 3b demonstrates, it is often a relatively simple matter to model additional complexity by adding structure to a BBN, especially where a single factor, such as age, affects more than one state or event.

BBN node diagrams provide a particularly clear representation of conditional independence. Our Figure 5 shows that *Herd type* is affected only by the *County* of a sample. *County* affects the distribution of *Age class* and *Herd type*, and whether an individual animal is infected in an infected herd depends on both of these factors. Our alternative analysis is easily depicted by dropping a single connection in the diagram (from *Herd type* to *Animal status*). Conditional relationships are high-level structures in the model—as opposed to the low-level structures in the CPTs—and such relationships are difficult to discern amongst the branches and leaves of scenario tree diagrams, for which branch labels are be required to depict some of the conditioning. A BBN diagram, therefore, provides a more visually accurate depiction of dependence relationships than a scenario tree.

BBN diagrams are particularly useful for exposition and auditing. Several software packages provide interactive tools for working with BBNs. Users of the software can set beliefs for any node and obtain instant feedback when beliefs for other nodes are calculated. This capacity, and the visual feedback provided by the display of probabilities both as numbers and "belief bars" (as illustrated in Figures 5a and 5b) makes it easy to check the BBN for consistency with data and assumptions—an essential step for reviewing claims of freedom from disease.

A BBN can be converted into an influence diagram, which can be used directly as a tool to assist decision makers [17]. In the CSF example, the *Sampling scheme* node could easily become a decision node for evaluation of alternative sampling strategies. Given sampling costs, and a value for the confidence in freedom, the decision node would rank strategies according to their expected utility. It then becomes a simple matter to place a value on additional information and determine the sensitivity of the network and its outputs to different surveillance components.

Calculation of component sensitivities using a BBN poses some practical difficulties. First, calculation of probabilities for multiple clusters (e.g., herds) where risk factors vary by cluster requires either automation by code or tedious pointing and clicking via the graphical user interface of a software program. However, the code required is quite straightforward. For the CSF example, we automated Netica™ via its COM interface [18], with each step in the algorithm requiring just one line of code. Automation by code is a generally useful technique, since it allows an existing BBN to be used as a component of a simulation program. This can obviate the need to discretize continuous variables—which can lead to errors in BBNs [19]—and also permits stochastic simulation.

A more serious problem with the BBN formulation of a surveillance problem is that some of the logic and relationships are hidden—not in the nodes and arrows of the diagram—but in conditional probability tables. These relationships are explicit in the scenario tree format at the expense of greater clutter in the resulting diagram. Scenario trees are rightly promoted in the OIE handbooks on import risk analysis [5, 6] as an aid to logical identification of pathways and information requirements. Their value for "clarifying ideas and understanding the problem" [6], however, diminishes with complex surveillance programs for which a BBN node diagram provides greater clarity. In the examples presented here, low-order relationships are encapsulated in the particular probabilities of the CPTs. These are populated using observed frequencies and application of equation 8 (Tables III and IV). In more complex applications, extra care may be required to ensure that CPTs are well documented and that they accurately represent data and assumptions.

For surveillance systems where systematically collected data is available, the BBN formulation is a bridge to a fully Bayesian analysis using Markov chain Monte Carlo methods to obtain posterior estimates of prevalence based on prior distributions of model parameters (e.g., [20-22]). The dependence on prior distributions makes such methods inherently controversial [23] but the scenario tree and BBN approaches have a similar credibility problem when there is uncertainty in detection probabilities and other parameters. In both approaches, many sources of uncertainty—e.g., in parameter values, bounds and model structure—are ignored for the sake of exposition and mathematical tractability. Burgman *et al.* [24] have recently demonstrated how information-gap theory [25] can be used to make robust decisions by accounting for such uncertainty in BBNs.

There is no fully satisfactory general methodology for proving disease freedom in international fora. For complex surveillance systems, BBNs and the calculation methods presented here extend the reach of the scenario tree methodology, making some claims easier to analyze and audit.

Table I. Case Study I. Fundamental parameters and derived quantities

Name	Value	Calculation	Explanation
Fundamental parameters			
PrP_High	0.30	Fixed value	Proportion of animals close to infected country (high risk)
RR_High	4	Fixed value	Relative risk of infection close to infected country
P^*_A	0.01	Fixed value	Animal design prevalence
ELISASe	0.95	Fixed value	ELISA sensitivity
Derived quantities			
AR_Low	0.526	$1/(RR\_High \times PrP\_High + 1 - PrP\_High)$	Adjusted relative risk of infection for animals far from infected country
AR_High	2.11	$AR\_Low \times RR\_High$	Adjusted relative risk of infection for animals close to infected country
EPIA_Low	0.00526	$AR\_Low \times P^*_A$	Effective probability of infection for an animal far from infected country
EPIA_High	0.0211	$AR\_High \times P^*_A$	Effective probability of infection for an animal close infected country

Table II. Transition probabilities for the Danish CSF case study. Proportions of herd types and age classes derived from Table 3 of Martin *et al.* [3]. Effective probabilities of infection calculated as described in the text.

From	To	Formula	Value	Explanation
1	2	*PrP_SJ*	0.0913	Proportion of herds in South Jutland (SJ)
1	3	1 – *PrP_SJ*	0.909	Proportion of herds in other counties
2	4	*EPIH_SJ*	0.0285	Effective probability of infection for a herd in SJ
3	5	*EPIH_Other*	0.00814	Effective probability of infection for herds in other counties
4	6	*PrP_SJ_Breeder*	0.472	Proportion of breeder herds in SJ
4	8	1 – *PrP_SJ_Breeder*	0.528	Proportion of slaughter herds in SJ
5	7	*PrP_Other_Breeder*	0.489	Proportion of breeder herds in other counties
5	8	1 – *PrP_Other_Breeder*	0.511	Proportion of slaughter herds in other counties
6	9	*PrP_SJ_Breed_Ad*	0.45	Proportion of adults in breeder herds in SJ
6	10	1 – *PrP_SJ_Breed_Ad*	0.55	Proportion of growers in breeder herds in SJ
7	9	*PrP_Oth_Breed_Ad*	0.41	Proportion of adults in breeder herds in other counties
7	10	1 – *PrP_Oth_Breed_Ad*	0.59	Proportion of growers in breeder herds in other counties
8	10	*PrP_Slaughter_Grower*	1	Proportion of growers in slaughter herds (all counties)
9	11	*EPIA_Breeder_Adult*	0.0933	Effective probability of infection of an adult in an infected breeder herd
10	11	*EPIA_Breeder_Grower*	0.0187	Effective probability of infection of a grower in an infected breeder herd
11	12	*ELISASe*	0.95	Sensitivity of the CSF ELISA

Table III. Case Study II. Conditional probability table for the (a) *Herd type* and (b) *Age* nodes. In each case the elements reflect observed proportions in the sampled population.

(a)

Sampling scheme	*County*	Breeder	Slaughter
Targeted	South Jutland	1	0
	Other	1	0
Representative	South Jutland	0.472	0.528
	Other	0.489	0.511

(b)

Sampling scheme	*County*	*Herd type*	Adult	Grower
Targeted	South Jutland	Breeder	1	0
		Slaughter	0	1
	Other	Breeder	1	0
		Slaughter	0	1
Representative	South Jutland	Breeder	0.45	0.55
		Slaughter	0	1
	Other	Breeder	0.41	0.59
		Slaughter	0	1

Table IV. Case Study II. Conditional probability tables for the (a) *Herd status* and (b) *Animal status* nodes. Calculation details are described in the text.

(a)

County	Infected (*EPIH*)	Uninfected (1 – *EPIH*)
South Jutland	0.0285	0.972
Other	0.00814	0.992

(b)

Herd status	*Herd type*	*Age*	Infected (*EPIA*)	Uninfected (1 – *EPIA*)
Infected	Breeder	Adult	0.0933	0.907
		Grower	0.0186	0.981
	Slaughter	Adult	0.25	0.75
		Grower	0.05	0.95
Uninfected	Breeder	Adult	0	1
		Grower	0	1
	Slaughter	Adult	0	1
		Grower	0	1

Figure 1. Node diagram representation of a scenario tree for a sampling scheme in which an infected population is subjected to a single test. It has three nodes being: (1) "All" the undifferentiated state before testing is completed, (2) "+" positive and (3) "–" negative outcomes.

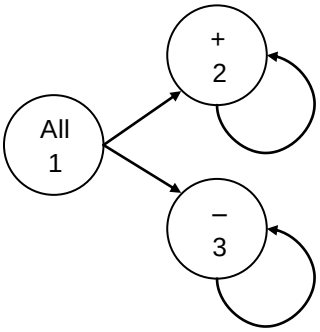


Figure 2. Scenario tree for Case study I (a) and corresponding node diagram (b) in which the width of the arrows approximates the transition probabilities—created using the Poptools [16] software.

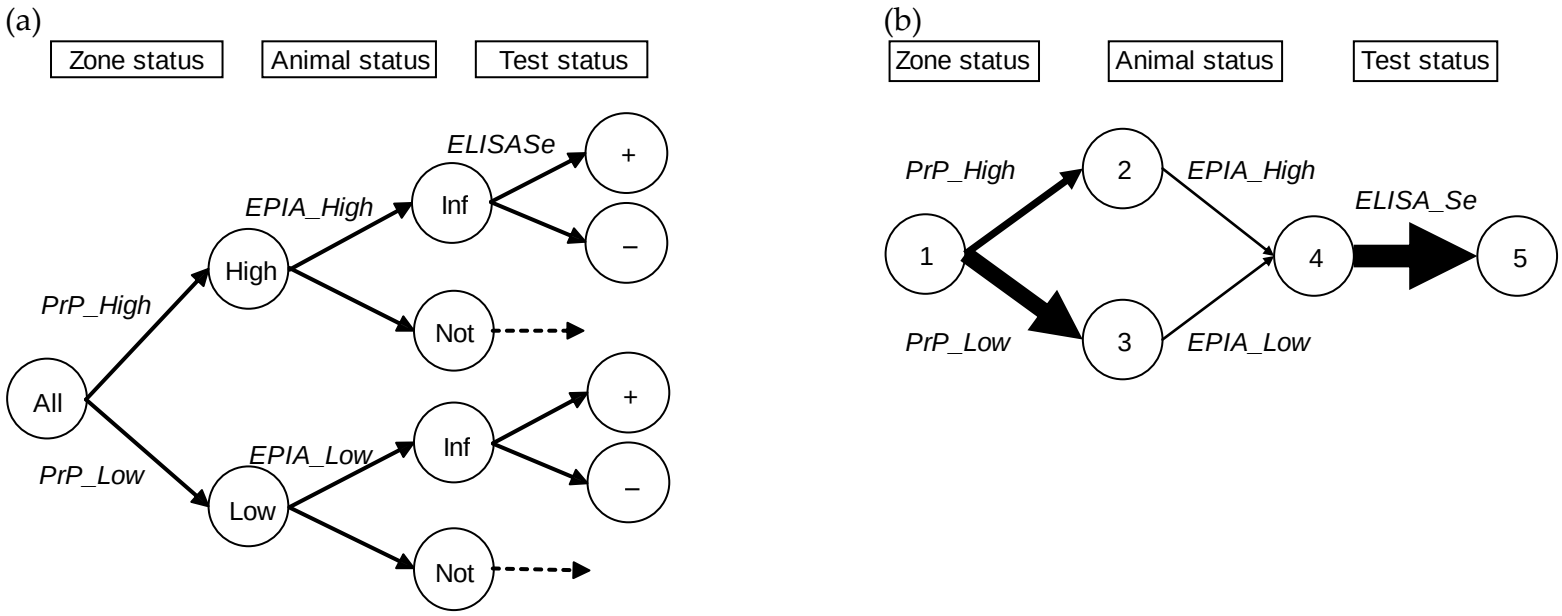
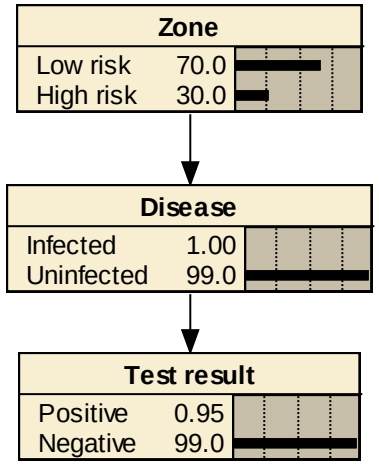


Figure 3. (a) Bayesian belief network for the FMD scenario tree as displayed by the Netica™ software [15], and (b) an extension of the network to include zone-specific age distributions and age-specific probabilities of infection.

(a)



(b)

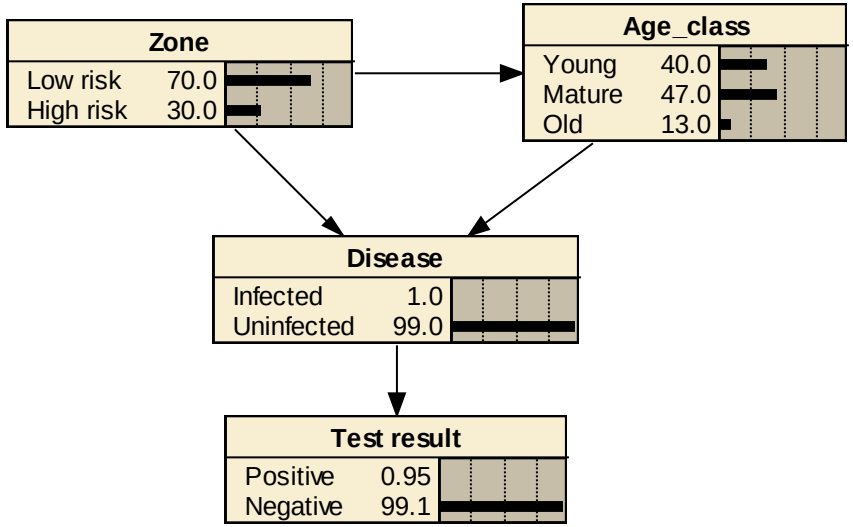


Figure 4. Node diagram for Case study II showing possible histories for positive samples. Fundamental parameters are shown in bold font. The width of the arrows approximates the transition probabilities.

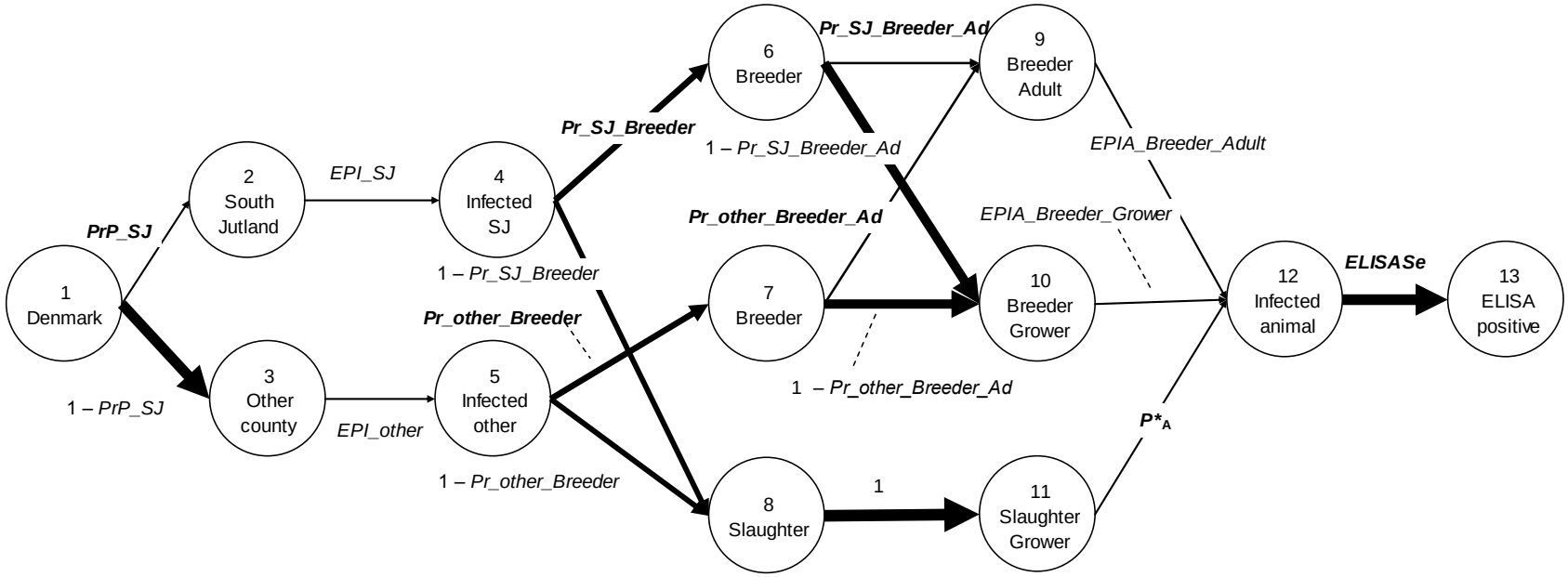
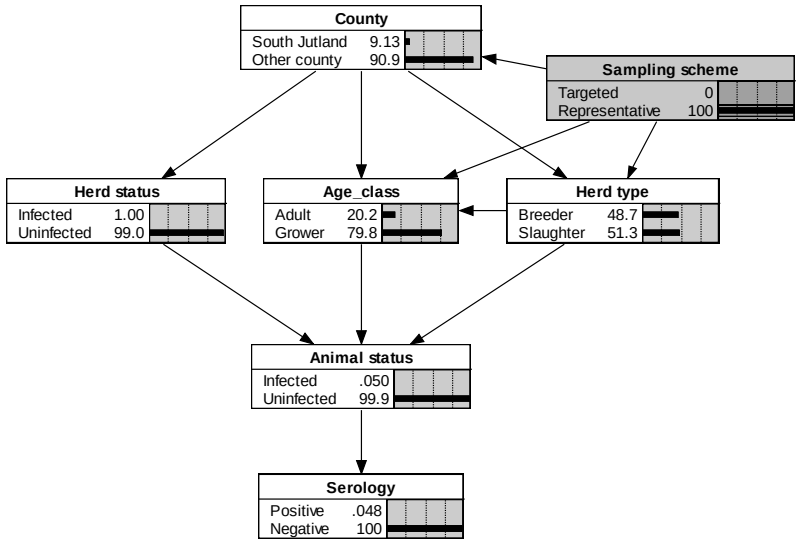
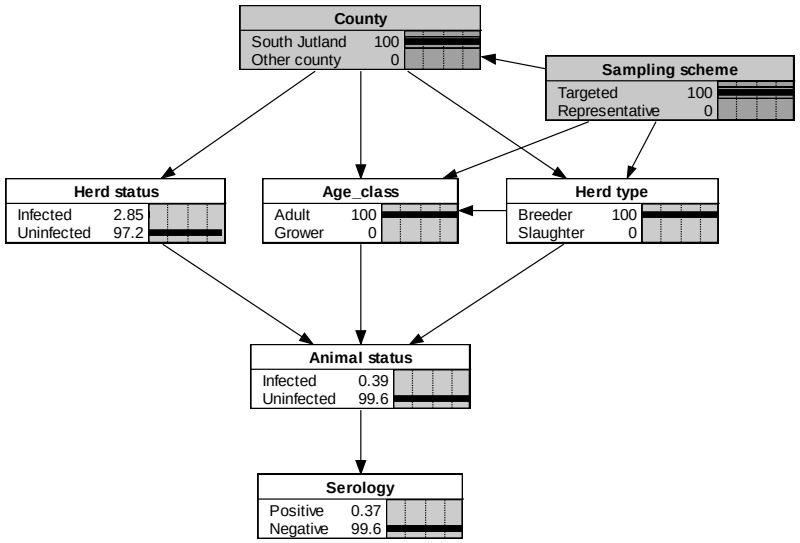


Figure 5. Bayesian belief network representation for the Danish CSF study in (a) the 'default' state in which the *Sampling scheme* node is set to *Representative* but no other information is know about a sample, and (b) in the state required to calculate a posterior probability of infection of a herd given a single negative sample from South Jutland in the targeted sampling scheme.

(a)



(b)



9. References

1. WTO. *Agreement on the Application of Sanitary and Phytosanitary Measures*. Geneva: World Trade Organization 1995.
2. Zepeda C, Salman M, Thiermann A, Kellar J, Rojas H and Willeberg P, The role of veterinary epidemiology and veterinary services in complying with the World Trade Organization SPS agreement, *Preventive Veterinary Medicine*, 2005; 67:125-140.
3. Martin P, Cameron A, Barfod K, Sergeant E and Greiner M, Demonstrating freedom from disease using multiple complex data sources 2: Case study–Classical swine fever in Denmark, *Preventive Veterinary Medicine*, 2007; 79:98-113.
4. Martin PAJ, Cameron AR and Greiner M, Demonstrating freedom from disease using multiple complex data sources: 1: A new methodology based on scenario trees, *Preventive Veterinary Medicine*, 2007; 79:71-97.
5. OIE, *Handbook on import risk analysis for animals and animal products. Volume 2: Quantitative risk assessment*, OIE - World Organisation for Animal Health, 2004.
6. OIE, *Handbook on import risk analysis for animals and animal products. Volume 1: Introduction and qualitative risk analysis, OIE - World Organisation for Animal Health*, 2004.
7. OIE, *Terrestrial Animal Health Code*. Paris, France: OIE (World Organisation for Animal Health), 2004.
8. Heckerman D, Bayesian Networks for Data Mining, Data Mining and Knowledge Discovery, *Data Mining and Knowledge Discovery*, 1997; 1:79-119.
9. Borsuk ME, Stow CA and Reckhow KH, A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis, *Ecological Modelling*, 2004; 173:219-239.
10. Hammond TR, A recipe for Bayesian network driven stock assessment, *Canadian Journal of Fisheries and Aquatic Sciences*, 2004; 61:1647-1657.
11. R Development Core Team. *R: A language and environment for statistical computing*, 2008. Available at: http://cran.r-project.org/, Accessed on November 15, 2008.
12. Bailey NTJ, *The Elements of Stochastic Processes with Applications to the Natural Sciences*. New York: John Wiley & Sons, 1964.
13. Bedford T and Cooke R, *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge University Press, 2001.
14. Sergeant ESG, Nielsen SS and Toft N, Evaluation of test-strategies for estimating probability of low prevalence of paratuberculosis in Danish dairy herds, *Preventive Veterinary Medicine*, 2008; 85:92-106.
15. Norsys Software Corp. *Netica version 4.08*, 2008. Available at: http://www.norsys.com, Accessed on December 1, 2008.
16. Hood GM. *PopTools version 3.0.5*, 2008. Available at: http://www.cse.csiro.au/poptools, Accessed on May 30, 2008.

17. Shachter RD, Evaluating Influence Diagrams, *Operations Research*, 1986; 34:871-882.
18. Norsys Software Corp. *Netica Help*, 2008. Available at: http://www.norsys.com, Accessed on December 1, 2008.
19. Parsons DJ, Orton TG, D'Souza J, Moore A, Jones R and Dodd CER, A comparison of three modelling approaches for quantitative risk assessment using the case study of Salmonella spp. in poultry meat, *International Journal of Food Microbiology*, 2005; 98:35-51.
20. Hanson TE, Johnson WO and Gardner IA, Hierarchical Models for Estimating Herd Prevalence and Test Accuracy in the Absence of a Gold-Standard, *Journal of Agricultural, Biological, and Environmental Statistics*, 2003; 8:223–239.
21. Ranta J, Tuominen P and Maijala R, Estimation of True Salmonella Prevalence Jointly in Cattle Herd and Animal Populations Using Bayesian Hierarchical Modeling, *Risk Analysis*, 2005; 25:23-37.
22. Suess EA, Gardner IA and Johnson WO, Hierarchical Bayesian model for prevalence inferences and determination of a country's status for an animal pathogen, *Preventive Veterinary Medicine*, 2002; 55:155-171.
23. Martin PAJ, Current value of historical and ongoing surveillance for disease freedom: Surveillance for bovine Johne's disease in Western Australia, *Preventive Veterinary Medicine*, 2008; 84:291-309.
24. Burgman MA, Wintle BA, Thompson CA, Moilanen A, Runge MC and Ben-Haim Y. *Qualitative modelling and Bayesian network analysis: Eliciting reliable expert judgement*. Melbourne: Australian Centre of Excellence for Risk Analysis
25. Ben-Haim Y, *Information-gap decision theory, 2nd ed*. London: Academic Press, 2006.