SHORT TERM SCIENTIFIC MISSION (STSM) SCIENTIFIC REPORT

Action number: CA17110

STSM start and end date: 12/01/2020 to 18/01/2020
Grantee name: ELEFTHERIOS MELETIS

STSM title - WG3 | STSM2:
Evaluation of machine learning methods for the estimation of a probability of freedom from infection from data generated by disease control programmes

PURPOSE OF THE STSM:

The main objective of the SOUND control project is summarized in this figure:
Disease control programme → Data input → Model → Output

Main purpose of this STSM was the evaluation of machine learning methods in order to compare surveillance output from different control programmes, i.e. data input.
To do so, main focus of this STSM, was investigating how factors influence the test outcome and testing the dependencies between factors, in a country free of infection. In particular, influence or importance of different attributes of the surveillance program should had to be integrated.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

Bayesian networks (BNs) were used to represent the set of variables and their conditional probabilities. BNs are considered to be an example of machine learning structure input in relation to output information. To learn a BN the data has to comprise presumably influential factors and certain output. Due to lack of reports of the test outcomes, from the original test dataset for the surveillance of BVDV infection in cattle, artificial data were generated. A new binomial variable, representing the result of the performed test was generated. Selected factors were super imposed by a proportion probability to create positive test outcomes. Using the input dataset, the structure between the factors was represented by a fitted BN. Four factors, Herd_type {Dairy, Beef}, Sample_type {Bull tank milk, Blood}, Sample_size {numbers of animals examined}, Herd_size {factor with 5 scales} were examined on how they influence the result {Negative, Positive} of the performed test. By adding a factor on representing Country {A, B}, two different surveillance programs were mimicked using different factor weights. Specifically for Country A factors Sample_type and Sample_size were assumed to have more weight in the outcome, while on Country B Sample_type and Herd_size. In summary, we had two artificial datasets representing the output of two different surveillance programs. The learned dependencies were explored in detail by selecting individual factor constellations.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

Screenshots of the BNs illustrating the associated factors with the outcome are provided below. BNs can be used to examine the relationship between the factors and the output, which is the main purpose of this approach. Detailed examination yields subtle difference between alternative factor constellations. However, two main problems are not yet fully understood. First, even with substantial differences with the factor manipulations the approach is not simple to discriminate the difference on the distribution of positive and negative outcomes. Second, the output, as a proportion of positive and negative outcomes, might be insufficient to reflex such subtle differences between programs. On the other hand, this similarity of the output for differently assumed surveillance protocols might be used to further develop discriminatory criteria, which could not be finalized in this STSM.

For example, on the below BNs we can see how the probability of having a positive outcome differs between the beef herds of the two countries. The probability of having a positive outcome from the beef herds of Country A is 11.51 % (Figure 1), while in Country B is 20.98% (Figure 2).

Figure 1

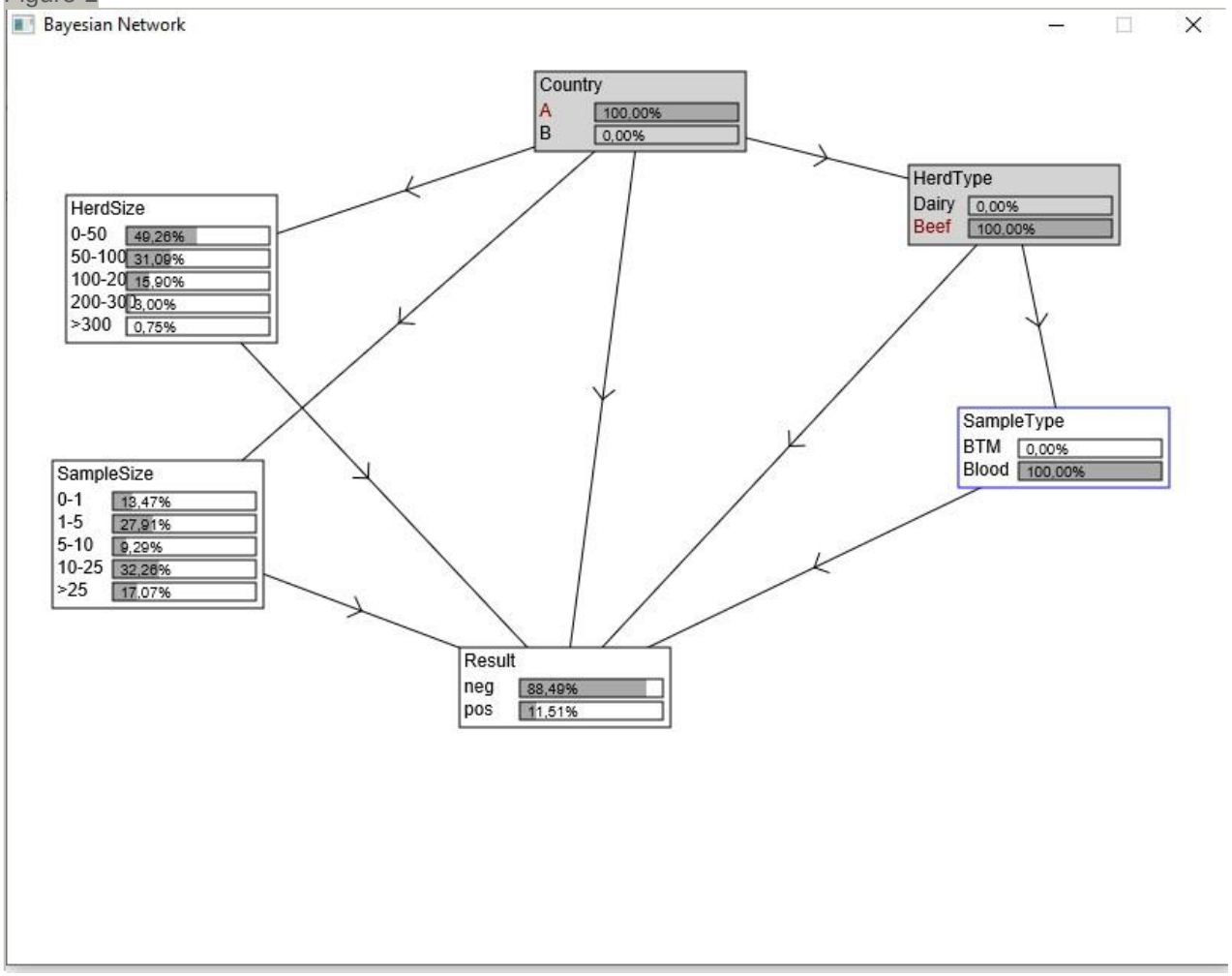
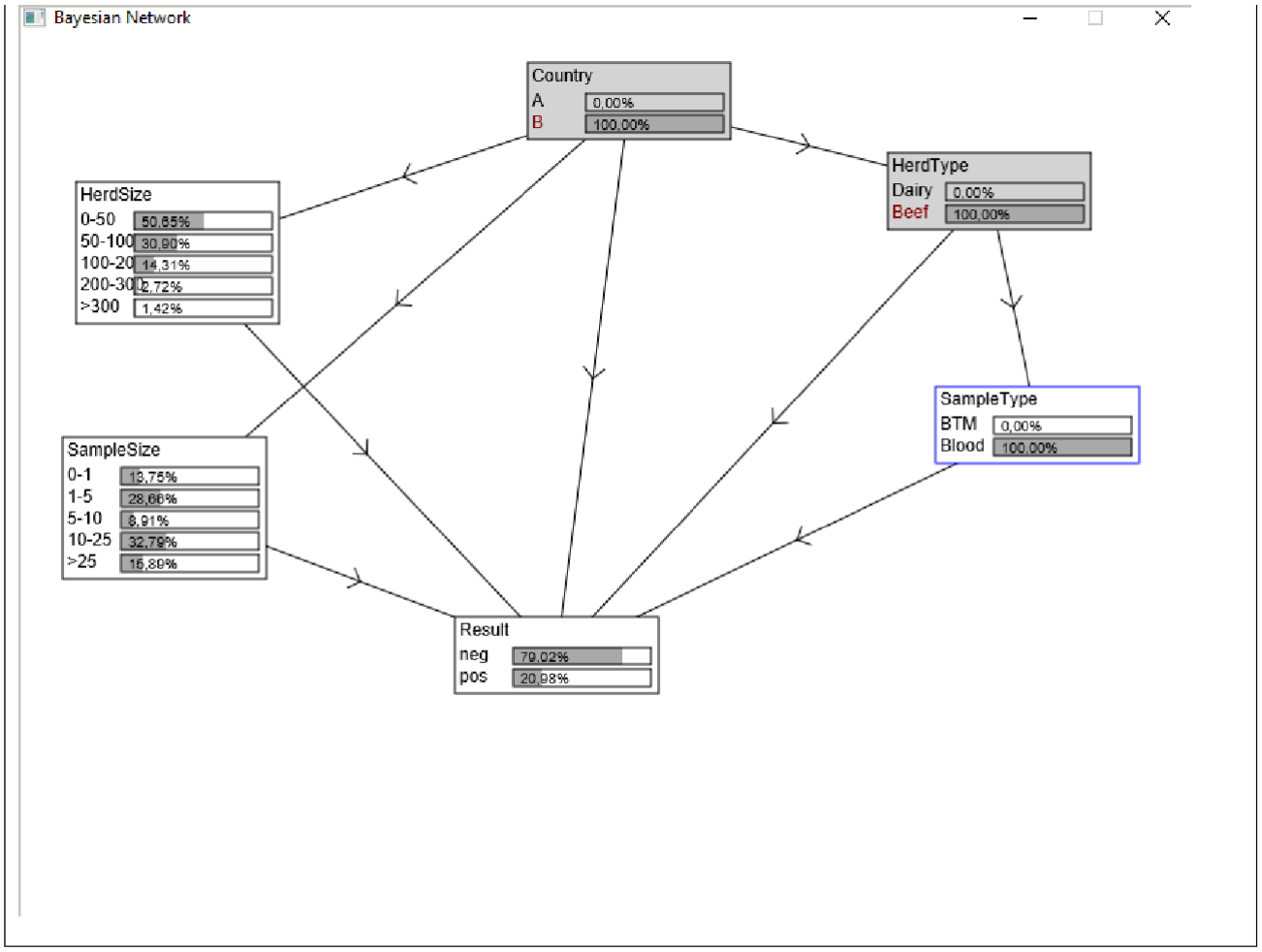


Figure 2


Bayesian Network
Country
A 0,00%
B 100,00%
HerdSize
0-50 50,65%
50-100 30,90%
100-20 14,31%
200-30 2,72%
>300 1,42%
HerdType
Dairy 0,00%
Beef 100,00%
SampleType
BTM 0,00%
Blood 100,00%
SampleSize
0-1 18,75%
1-5 28,66%
5-10 8,91%
10-25 32,79%
>25 10,88%
Result
neg 79,02%
pos 20,98%