Use of structural learning for assessment of sandy soils vulnerable to pesticide leaching

The overall objective of the KUPA project (www.kupa.dk) is to develop an operational concept for identifying areas where aquifers were vulnerable to pesticide contamination. In the part I of KUPA, the focus has been on a concept for sandy soils.

Studies had shown that pesticide degradation occur predominately in the presence of oxygen (aerobic conditions) (Albrechtsen et al., 2001). Throughout Denmark aerobic conditions occur mainly in the unsaturated zone above the groundwater table. Therefore, understanding of the fate of pesticides in the unsaturated zone is crucial for determining the risk of underlying aquifers to pesticide contamination.

The transport of pesticides through the unsaturated zone is largely controlled by the residence time in the aerobic zone, and the degradation and sorption characteristics of the soil. Residence times are a direct reflection of the hydraulic properties of the soil (hydraulic conductivity and porosity) and the thickness of the unsaturated zone together with the sorption capacity of the soil. Degradation rates are represented by the half-life (DT50) and sorptive characteristics of the pesticide and soil.

The focus in the KUPA study was to quantify the magnitude of the parameters controlling the transport of pesticides through the unsaturated zone. In order to develop a meaningfull classification scheme, the variability in the magnitude of individual parameters and correlation between parameters were investigated. Ideally, if all parameters important to pesticide leaching were known everywhere, the mass flux of pesticides through the unsaturated zone could be quantified. Obviously this is not the case and therefore, it is necessary to focus on determination of dominating parameters and generalization of results.

The results of KUPA for sandy soils showed that it was possible to identify areas vulnerable to pesticide leaching based on a limited number of soil parameters. Subsequently in second phase it is possible to further evaluate which soils are the most vulnerable based on modelling and/or correlation, again based on simple and easy collectable data. The analysis in KUPA show that three parameters are important in the first phase: ‘organic content’, ‘clay’ and ‘silt’. These parameters were accumulated for the A-, B- and C-horizon from the surface to 1 meter’s depth.

The example in this section illustrate an analysis of a nationwide dataset (‘kvadratnets data’ approximately 150 cases) which was used in KUPA to establish multi-variant correlations between ‘organic content’, ‘clay’ and ‘silt’ with predicted ‘relative pesticide leaching’ using the MACRO unsaturated zone solute transport pesticide model.

In the present example structural learning has been applied in order to analyse the 150 cases to identify ‘structure’ (variables and links) and CPT’s. The example includes:

- Structural learning in order to construct BBNs based on data (from 150 cases)
- Examples of use of BBNs for decision support system for groundwater management

6.5.4.1 Structural learning based on KUPA dataset

Structure learning can be performed via the Hugin Learning Wizard which allows data to be read from databases, to be preprocessed, etc or by activating one of the structural learning algorithms directly. Two algorithms are available for structural learning: The PC algorithm and the NPC algorithm. In this example we have used the PC algorithm.

The Hugin PC algorithm, which is a variant of the original PC algorithm due to Spirtes, Glymour & Scheines (2000), belongs to the class of constraint-based learning algorithms. The basic idea of these algorithms is to derive a set of conditional independence and dependence statements by statistical tests.

The algorithm performs the following steps:

- Statistical tests for conditional independence are performed for all pairs of variables (except for those pairs for which a structural constraint has been specified)
- An undirected link is added between each pair of variables for which no conditional independences were found. The resulting undirected graph is referred to as the skeleton of the learned structure.
- Colliders are then identified, ensuring that no directed cycles occur. (A collider is a pair of links directed such that they meet in a node.)
- Next, directions are enforced for those links whose direction can be derived from the conditional independences found and the colliders identified.
- Finally, the remaining undirected links are directed randomly, ensuring that no directed cycles occur.

One important thing to note about the PC algorithm is that, in general, it will not be able to derive the direction of all the links from data, and thus some links will be directed randomly. This means that the learned structure should be inspected, and if any links seem counterintuitive, one might consider using the Learning Wizard which provides a means of specifying structural domain knowledge.

Traditional constraint-based learning algorithms produce provably correct structures under the assumptions of infinite data sets, perfect tests, and directed acyclic graph (DAG) faithfulness i.e., that the data can be assumed to be simulated from a probability distribution that factorizes according to a DAG. In the case of limited data sets, however, these algorithms often derive too many conditional independence statements. Also, they may in some cases leave out important dependence relations.

Generally, it is recommended to use the NPC algorithm, as the resulting graph will be a better map of the conditional independence relations represented in the data. In particular, when the data set is small, the NPC algorithm should be the one preferred. The NPC algorithm, however, has longer running times than the PC algorithm.

The initial steps in structural learning are the following:

- select which variable should be included in the test
- define states for the problem and pre-process raw data into aggregated datasets (see Figure 6.21)
- analyse relationships (links and CPT's, see Figure 6.22, 6.33 and 6.24)

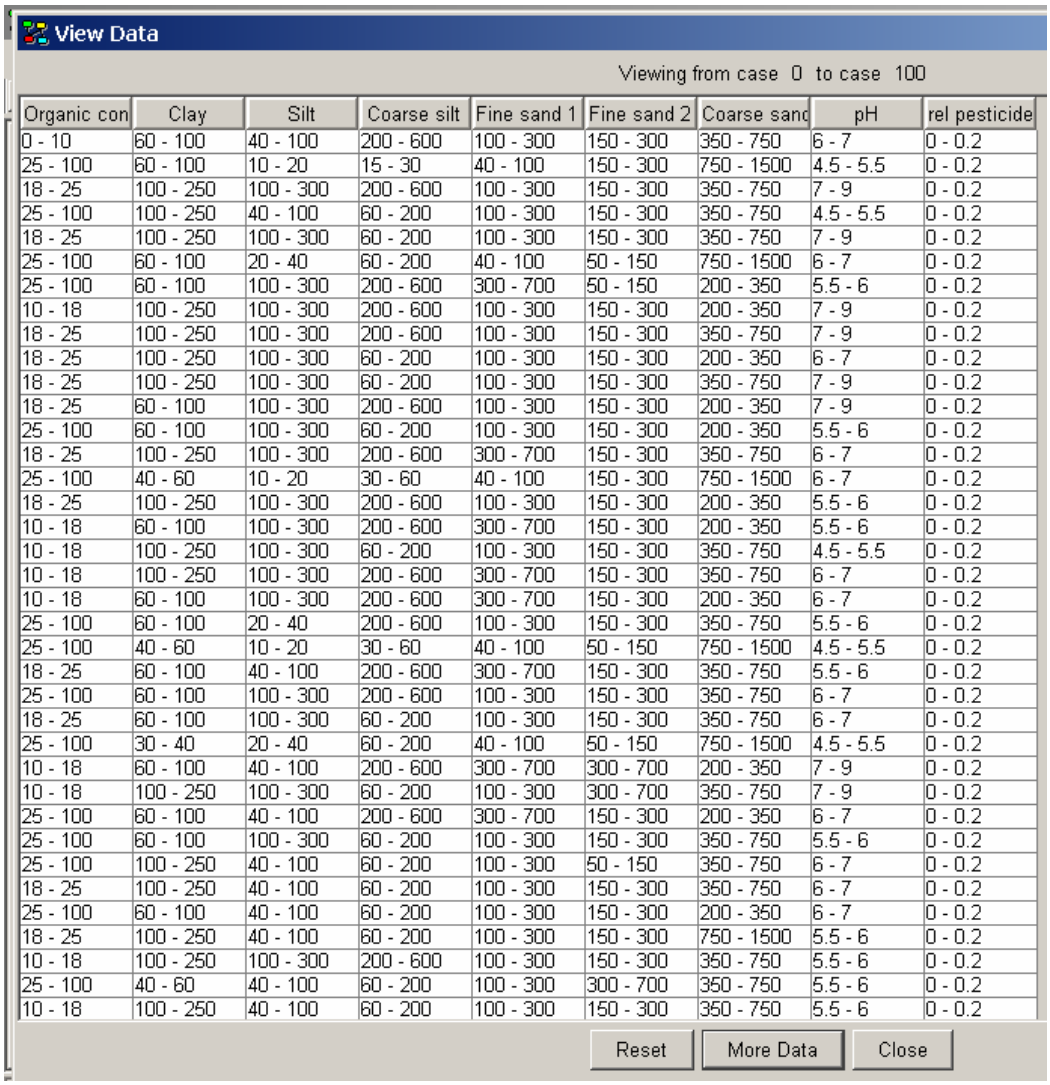

Figure 6.21 The first step in structural learning is analyse the data and to define states. In the example with the KUPA data set each variable is assigned one of four states. This can be done interactively using the Learning Wizzard in Hugin giving the results shown above.

First the variables are simply shown (Figure 6.22). The user are allowed to define any known dependencies or known independencies. Depending on the selected states Hugin may suggest relationship between clay and silt which the expert don't want, and in that case, the user should define such independencies. Alternatively, the user could want a certain structure at least for a small number of governing parameters, e.g. 'organic content', 'silt' and 'clay' with a given relationship to 'relative pesticide leaching' and analyse the other relationships given this predefinition.

When working with a limited number of cases, in our example approximately 150, the statistical analysis could easily result in coincidental structures rather than physical correct or logical structures. So the user should always be critical about what is produced, and eventually go a step back and repeat the structural learning process with other dependencies/independencies (the step shown in Figure 6.23). For the example below we used the PC algoritm, alternatively the NPC algoritm could have been selected.

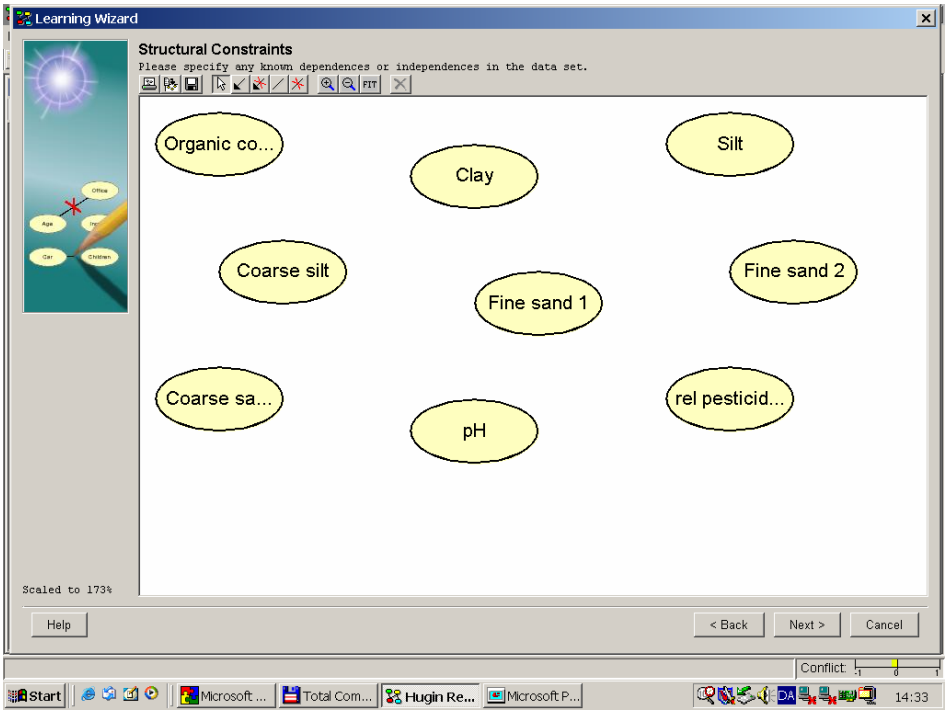

Figure 6.22 The Learning Wizzard ask for known dependencies or independencies between variable. In this example we didn't define any dependencies/independencies

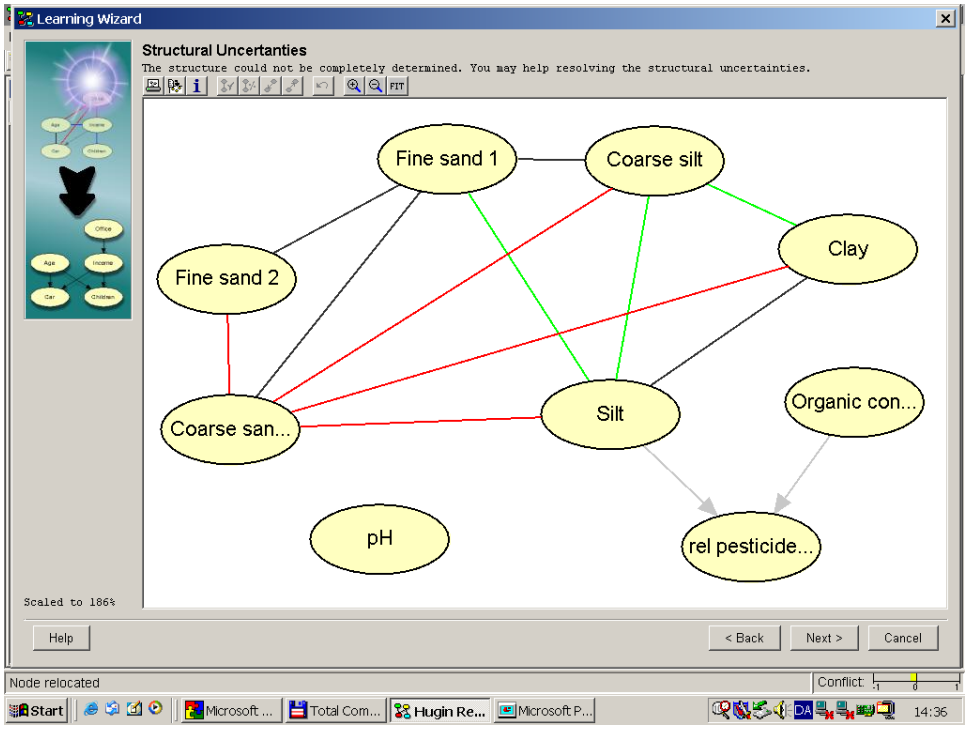

Figure 6.23 The Learning Wizzard defined 'strong' links between some variables shown with arrow. Other links had to be defined by the user

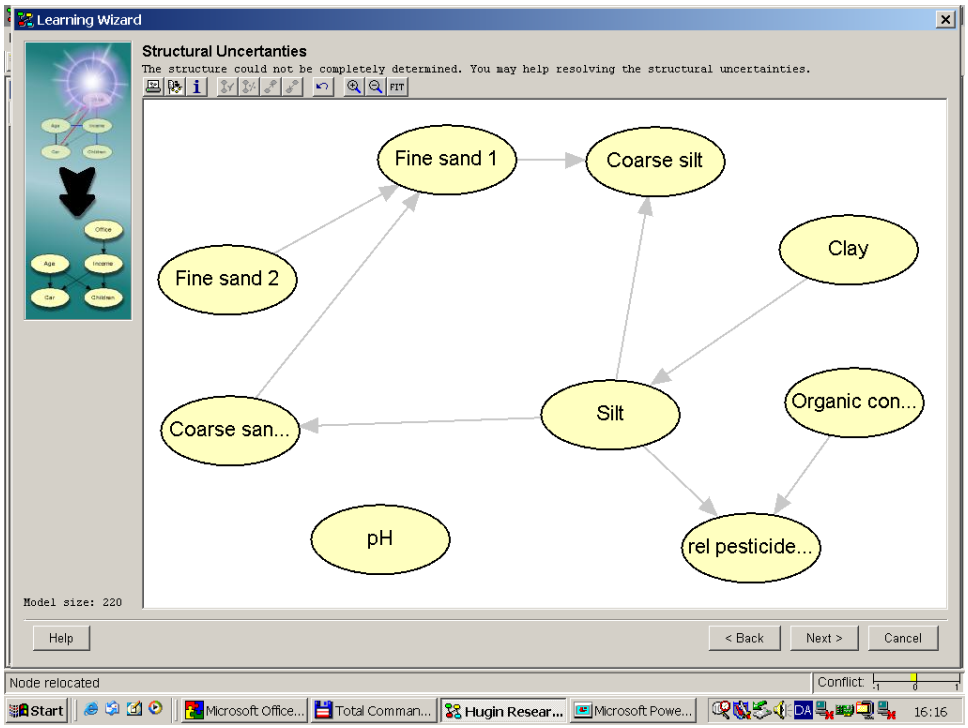

Figure 6.24 Final BBN after definition of all links. There is not a relationship between pH and any other variables in the BBN because pH.

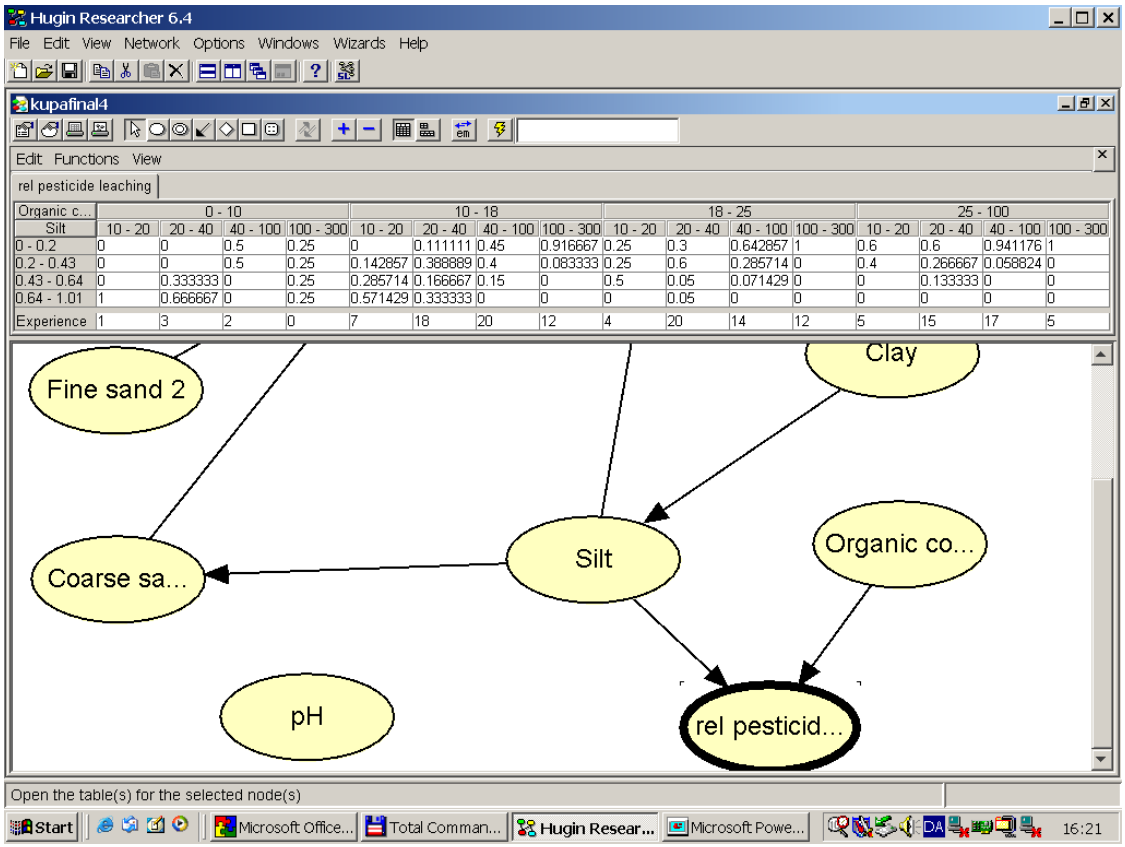

Figure 6.25 The variable 'rel pesticid…' (relative pesticide leaching) has two parent variables: 'Silt' and 'Organic co…' (organic content, in Danish: humus). Note that some of the relationships are rather weak (shown by the indentifier: "Experience" in the CPT). A more reliable CPT could be developed if also expert knowledge were incorporated.

6.5.4.2 Examples of use of BBNs for decision support system for groundwater management

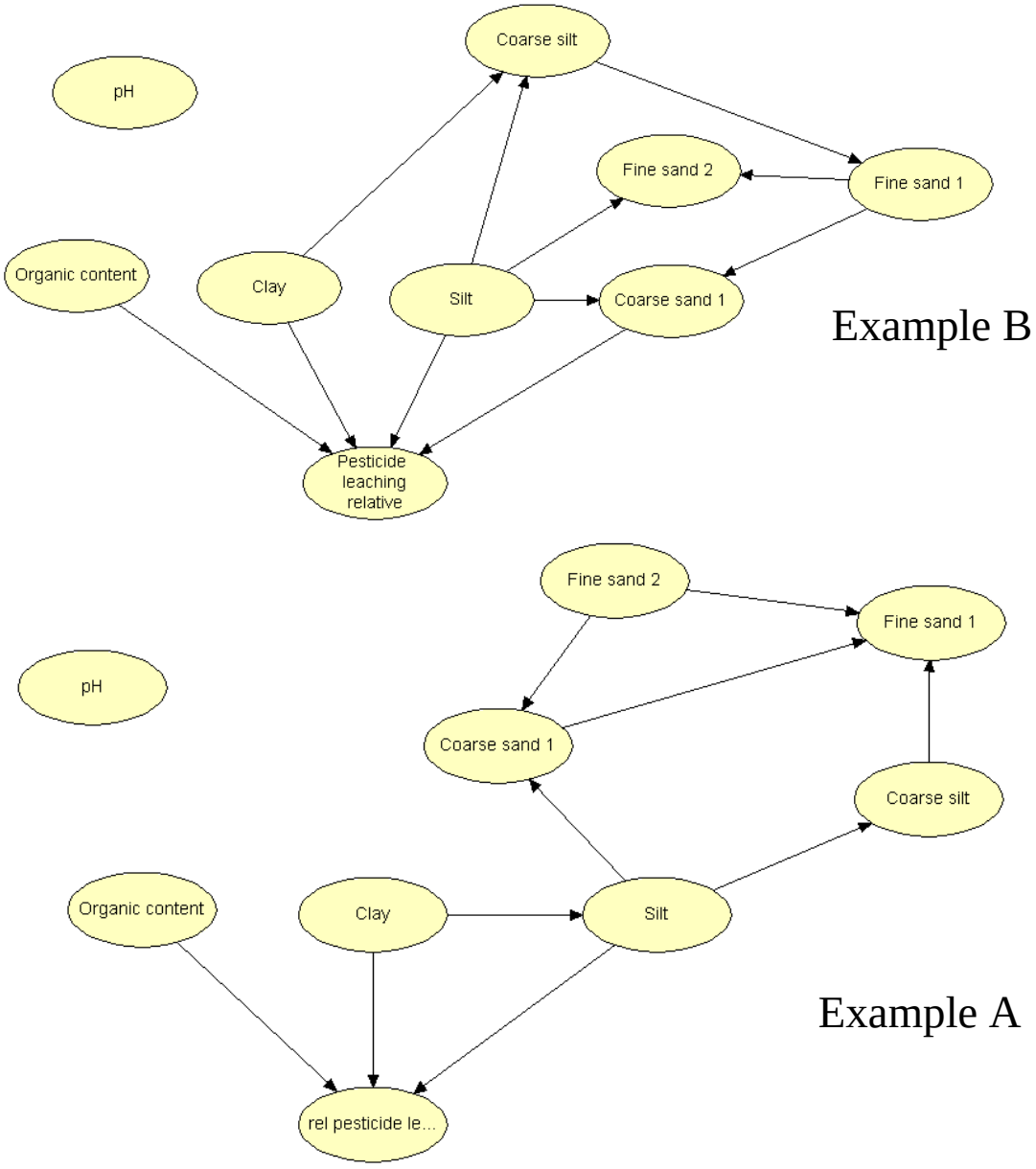

Figure 6.26 Two alternative results of structural learning based on the KUPA dataset. In the example A below three parameters had influence on relative pesticide leaching, whereas in the example B above, four parameters had influence on relative pesticide leaching. In none of the examples pH had direct influence (only hydraulics and sorption is described not degradation, and that is probably the reason why pH do not influence on relative pesticide leaching.

In the first example A the BBN was generated using structural learning and assuming dependencies between ‘Organic content’, ‘Clay’ and ‘Silt’ to ‘Relative pesticide leaching’. All other links were estimated interactively by Hugin and additional expert inputs. In the second example B the BBN was generated assuming independencies between ‘silt’ and ‘clay’ and with additional expert inputs in process of defining the which links should be included and direction.

The example A in figure 6.26 illustrates that only the three parameters have direct influence on relative pesticide leaching: ‘Organic content’, ‘Clay’ and ‘Silt’. However, ‘Clay’ and ‘Silt’ are interrelated which is indicated by the link between these variables. The more easily measured parameters ‘Fine sand 1 & 2’, ‘Coarse silt’, ‘Coarse sand 1’ influences the ‘Silt’ variable. Hence, instead of collecting data on ‘Silt’, it could be possible to collect data on ‘coarse sand’ or ‘coarse silt’ or even ‘Fine sand 1 & 2’. The latter data are easier to measure compared to clay, silt and organic content.

In the example in Figure B the structural learning resulted in another calculated structure which is a bit more complex with a total of four variables influencing relative pesticide leaching as parent variables. Again the test with Hugin confirm that results from KUPA regarding ‘Silt’, ‘Clay’ and ‘Organic Content’ as useful parameters for phase I vulnerability mapping for pesticides on sandy soils. But here the structural learning result in an additional parameter: ‘Coarse sand 1’ that has to been included, when the learning is based on the assumed independencies between ‘Clay’ and ‘Silt’.

A variable like ‘Coarse silt’ in example B has both ‘Clay’ and ‘Silt’ as parent variables and ‘Fine sand 1’ as a child variable. Below in Figure 6.27 is shown the result of structural learning and CPT’s calculated based on KUPA dataset for example A and B.

Note that evidence on ‘Silt’ in Figure 6.28 resulting in this variable in the lowest interval (10-20) and evidence on ‘Organic content’ in second lowest interval (10-18) brings the relative pesticide leaching indicator in a state of ‘alarm’. Under these conditions there is a 61,8 % chance of most vulnerable soil type and 23,5 % chance of second highest vulnerability interval. Note also that the evidence for ‘Silt’ has altered the ‘Clay’ probabilities and probabilities for ‘Fine sand 1’, ‘Coarse sand 1’ and ‘Coarse silt’. Only ‘Fine sand 2’ is not changed.

An alternative way of using the BBN in example A could have been to measure some of the easy observable sand and silt parameters (‘Fine sand 1 & 2’, ‘Coarse sand 1’ and ‘Coarse silt’) and simply analyse the ‘knock on’ effect of that evidence on the other variables. This is shown in Figure 6.29.

The only variable that was not influenced by the entered evidence in Figure 6.29 is the ‘Organic content’.

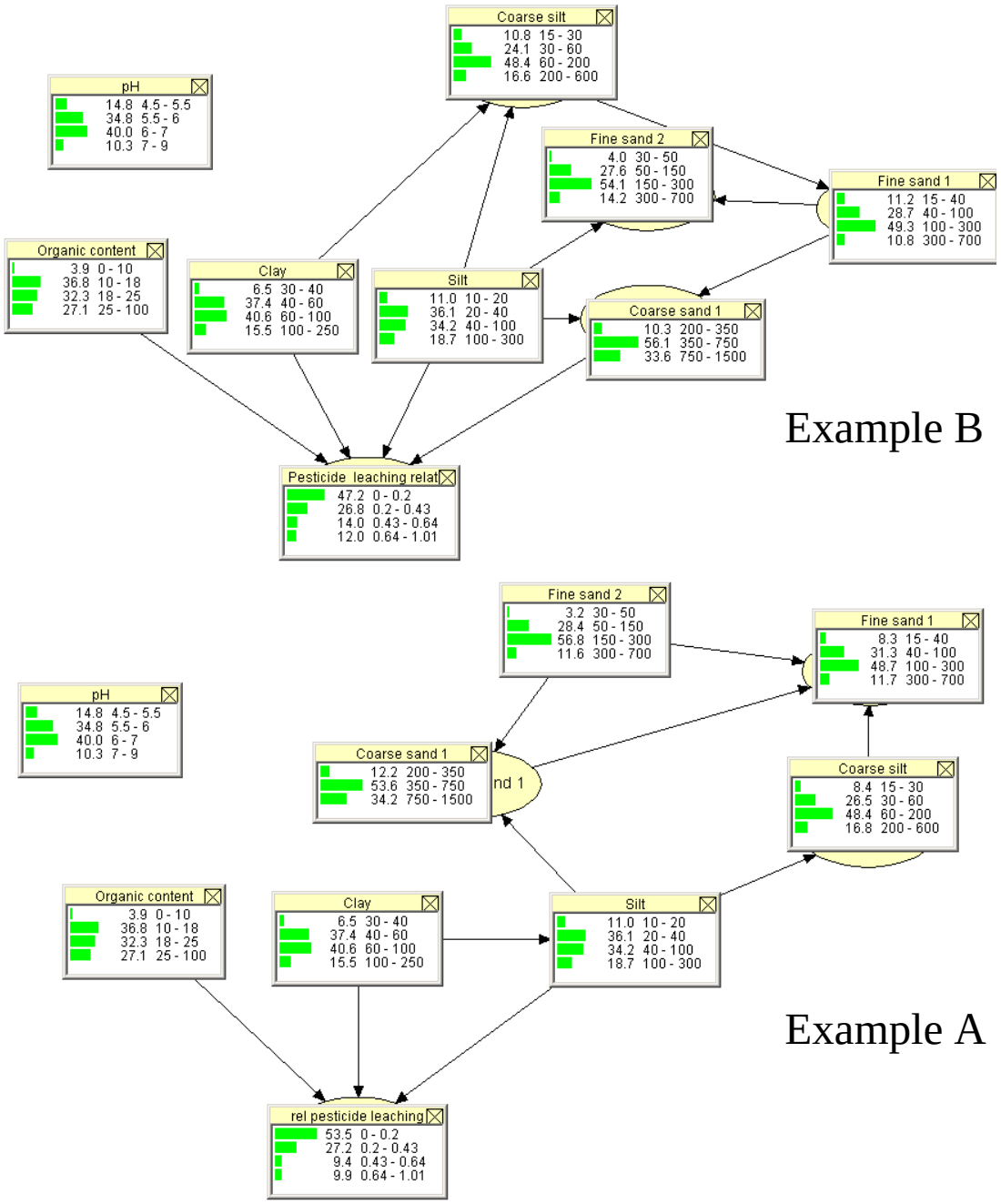

Figure 6.27 The most pesticide vulnerable areas belong to the interval 0.64-1.01 (9,9 % in example A). Vulnerable, but a bit less, are also soils belonging to the interval 0.43-0.64 (9.4 % in example A). Thus a total of 18 % belongs are relative pesticide leaching vulnerable.

Note that the 'Relative pesticide leaching' indicator based on the entered evidence now show a result of only a very little chance that the soil is vulnerable to leaching (= 1 % only). Depending on socio-economic framing conditions it could be decided to accept a low risk of pesticide leaching or to collect additional data on either 'Clay' or 'Organic content' in order to determine more precisely the risk of 'vulnerability to pesticide leaching'.

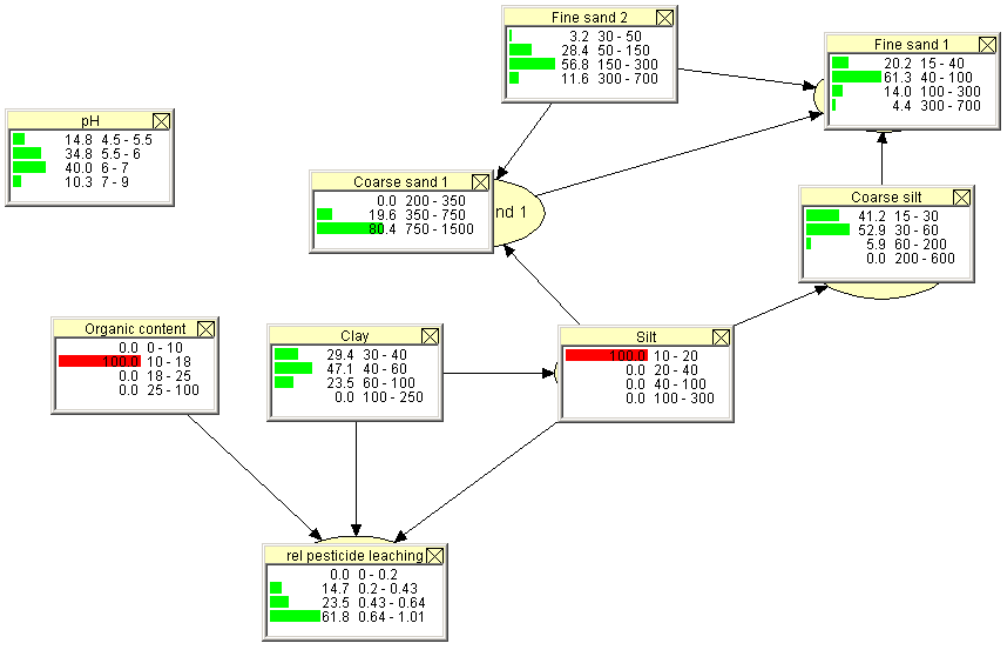

Figure 6.28 Example of adding evidence (red bars) to selected variables and updating the prognosis for pesticide leaching vulnerability for BBN example A.

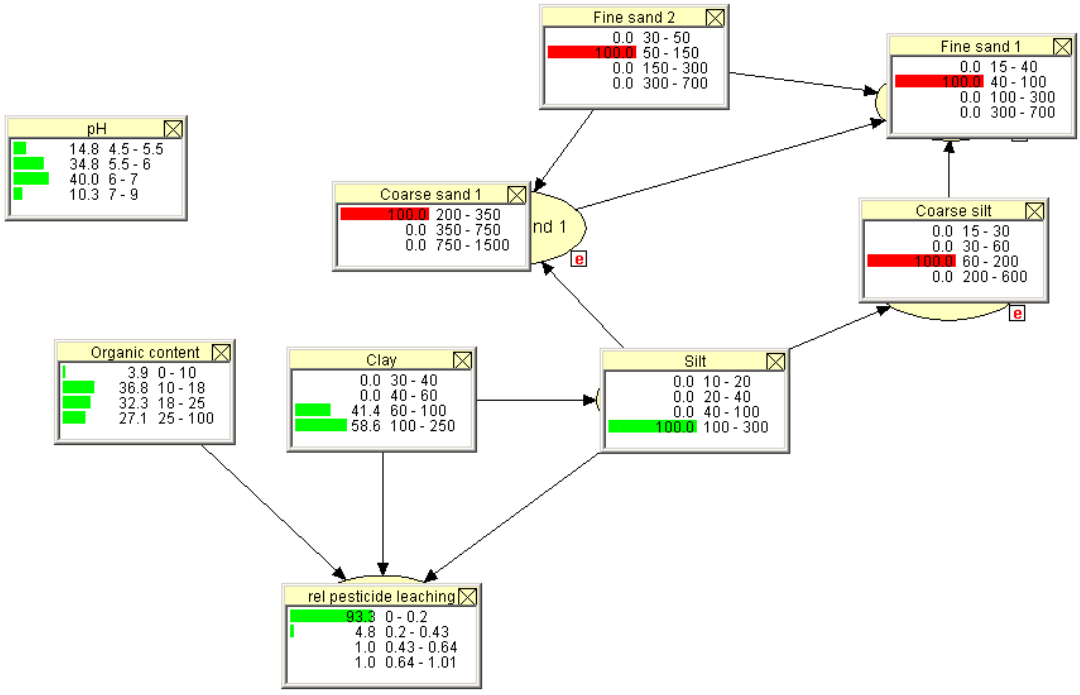

Figure 6.29 BBN example A with collected data for 'Coarse sand 1', 'Fine sand 2', 'Fine sand 1' and 'Coarse silt'. There is a 41,4 % chance for 'Clay' in the interval 80-100 and a 58,6 % chance of 'Clay' in the interval 100-250. Low risk of pesticide leaching in this case.

Another capability of a BBN, is the possibility of entering likelihood and analysing influence on other variables, given this assumption (likelihood). In figure 6.30 an example of entering likelihood for BBN example A, assuming highest or second highest relative pesticide leaching (either 0.43-0.64 or 0.84-1).

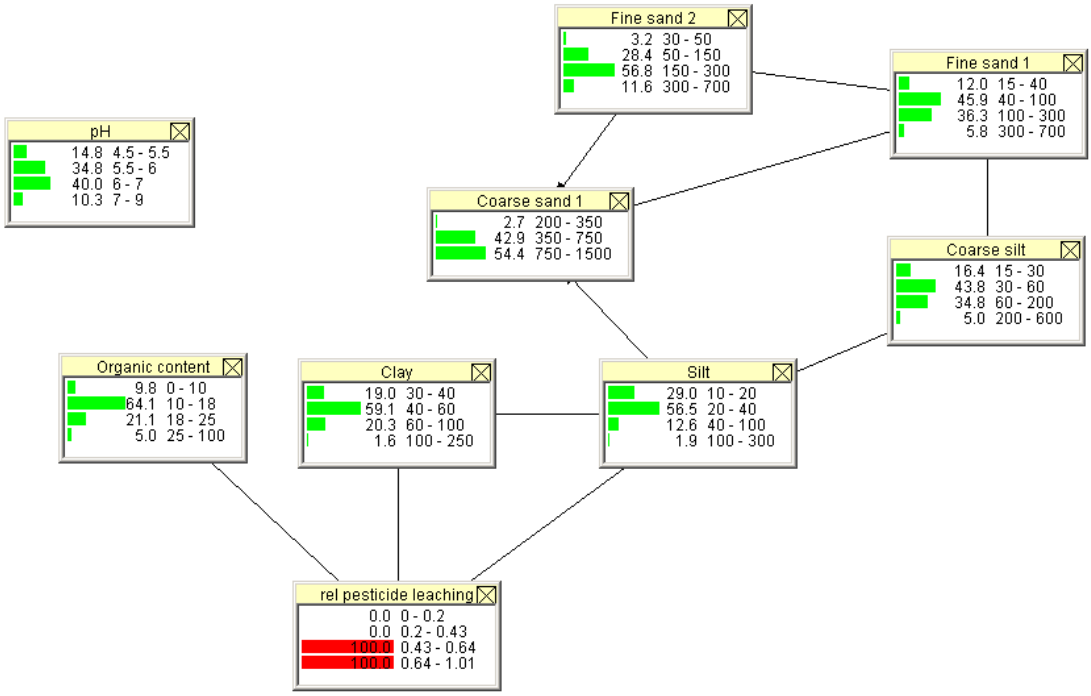

Figure 6.30 Example of entering likelihood of highest/high leaching risk in BBN example A

6.5.4.3 Summary of structural learning based on KUPA data set

The test of BBN development using structural learning algorithm on the KUPA dataset is only for demonstration purposes. Two alternative example BBNs: A and B was developed with different dependencies/independencies however many other structures was tested in the process. When no dependencies were defined, some structures resulted in two parameters: 'Silt' and 'Organic content' as parent variables to 'Relative pesticide leaching' others in three parameters: 'Silt', 'Clay' and 'Organic content' depending on which states were defined and which decisions were made when including/excluding other links or choice of direction of different links in the BBN. For examples the PC algorithm was used, even though the NPC algorithm may have been a better choice. The developed BBNs should be further evaluated before they are used in the real world because some of the CPT's for some of the relationships are rather weak (experiences < 5). However, BBNs are a flexible tool for analysing a dataset and combined with exper7t judgement useful for understanding and communication.

A main advantage of the BBNs, is the explicit quantification of uncertainty, which is useful in relation to groundwater management and protection. Furthermore, the possibility of assessing the 'Relative pesticide leaching' based on all available data for a given area and in addition, to combine this assessment with other data sets (e.g. degradation, pesticide application, groundwater monitoring and socio-economic conditions) is promising and should be further investigated. The example BBNs demonstrated in this section is limited to sandy soil conditions. Relative pesticide leaching for clay conditions should also be incorporated before the example BBNs are included in the more general BBN for groundwater protection described in earlier sections of this chapter.