Bilag 15. Eksempel på beslutningsstøtte system

GEUS: Hans Jørgen Henriksen

Ved hjælp af en strukturel læringsanalyse af strukturer i et datasæt ('structural learning') kan der findes frem til hvilke sammenhænge i datasættet der er mest afgørende for følsomheden overfor udvaskning af pesticid i sandjordsområder.

I projektet er der ved modelsimuleringer skaffet kvantitative oplysninger om hvilke forhold og jordegenskaber der er afgørende for udvaskningen af pesticid. Blandt andet er variabiliteten af jordegenskaberne og korrelationen mellem dem undersøgt. I en idealsituation, hvor værdien af alle jordegenskaber af betydning for pesticidudvaskning er kendt overalt, ville transporten af pesticid gennem den umættede zone kunne beregnes absolut. Med virkelighedens spredte datagrundlag er det åbenbart at dette ikke er muligt, hvorfor der er taget udgangspunkt i en vurdering af, dels hvad der er fundet at være de mest betydningsfulde jordegenskaber, dels i en generalisering af resultaterne.

Undersøgelserne af sandjorde har således vist at det er muligt at karakterisere særligt pesticidfølsomme arealer ved hjælp af oplysninger om et begrænset antal jordegenskaber. Resultaterne viser at indholdet af organisk kulstof, ler og silt i den øverste meter af jordprofilet (indenfor de pedolgiske A-, B- og C-horisonter), kan beskrive hovedparten af jordens følsomhed overfor udvaskning af pesticid, men at der kan være samme følsomhed ved kombinationer af forskellige værdier af jordegenskaber.

Derfor demonstreres der her et beslutningsstøtte system, der kan håndtere den forskellige vægtning af jordegenskaber, idet der benyttes de data fra kvadratnetsprofiler i sandjordsområder, som er blevet brugt til at foretage multivariat korrelation mellem simuleret relativ udvaskning af pesticid og indholdet af organisk kulstof, ler og silt.

Datasættet er analyseret med henblik på at identificere "struktur" (retningsorienterede sammenhænge mellem systemvariable og betingede sandsynlighedstabeller, CPTs). Analysen omfatter:

- Strukturel analyse og læring for at opbygge et Bayesiansk net (BN) bestående af systemvariable og retningsorienterede links (pile)
- Bestemmelse af CPTs for de fastlagte strukturer (BNs) udfra datasæt
- Eksempler på anvendelsen af BNs som beslutningsstøttesystem for identifikation af betydende zoneringskriterier og pesticidsårbare profiler.

Metode

Strukturer i datasættet kan analyseres ved hjælp af redskabet "Hugin Learning Wizard". Denne algoritme er indbygget i Hugin som er et software der kan anvendes til at konstruere BNs og som på baggrund af Bayes' sætning er i stand til efterfølgende at regne ('propagation') på nettene, givet at en eller flere variable er kendte (f.eks. målte). Der er to af disse algoritmer: en NPC ('Necessary Path Condition')og en PC ('Path condition') algoritme. Sidstnævnte benyttes i det følgende.

PC algoritmen fungerer i følgende trin:

- Parvis statistisk analyse af alle variable af om de er uafhængige (undtagen for par af variable som er tillagt en begrænsende betingelse)
- Tilføjelse af retningsløse forbindelser mellem de par af variable, hvor der ikke er fundet nogen betinget uafhængighed. Den resulterende graf med angivelse af retningsløse forbindelser kaldes "skelettet" i de strukturelle sammenhænge.
- Der identificeres herefter sammenstød ('colliders'). Sammenstød er par af retningsbestemte links der mødes i et knudepunkt (systemvariabel).
- I næste trin retningsbestemmes de links, hvis retning kan udledes på baggrund af de betingede uafhængigheder og identificerede sammenstød.
- Til sidst genereres retninger til de resterende retningsløse sammenhænge, idet det sikres at retningsbestemte links ikke går i ring (en forudsætning som brugen af BNs skal opfylde for at beregningsalgoritmen kan finde en løsning).

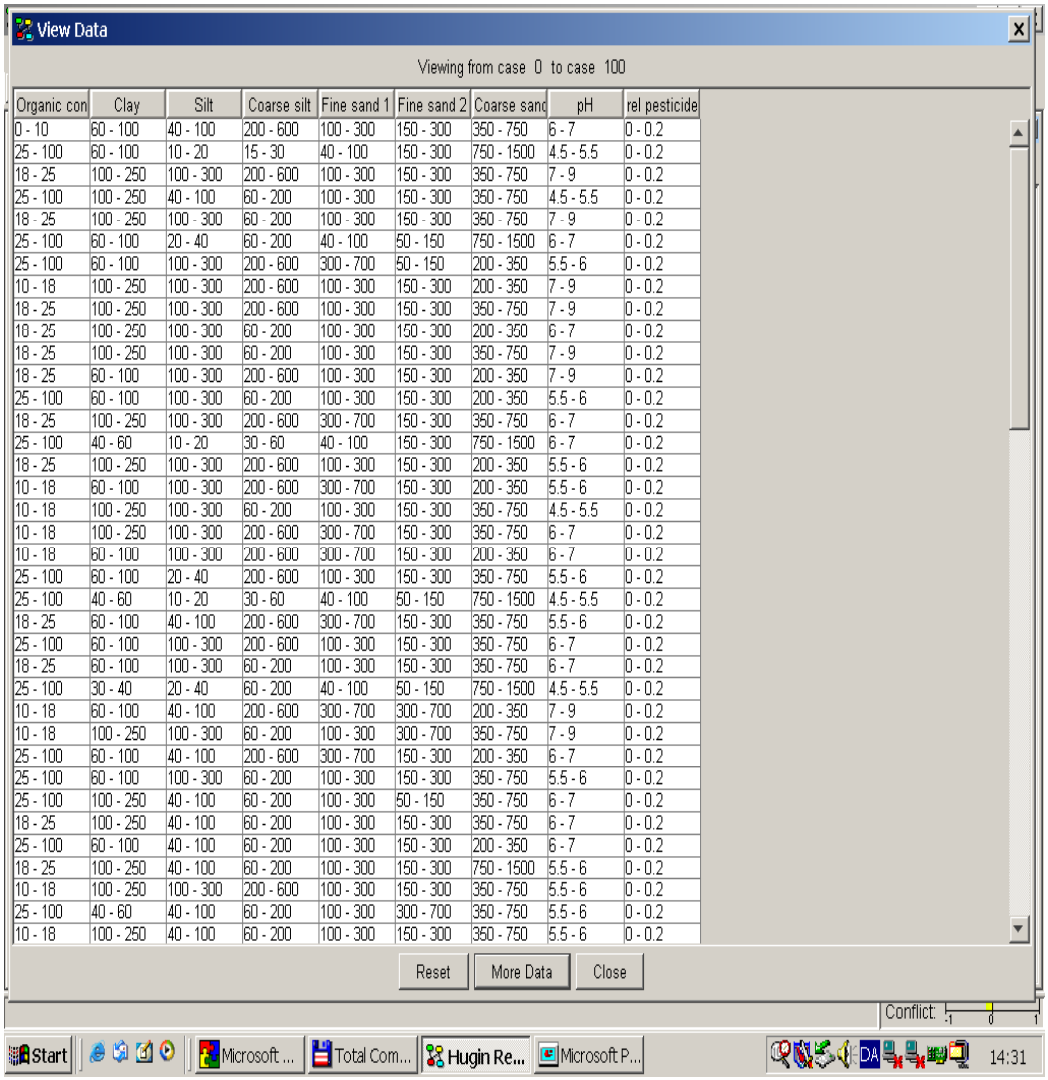
Normalt vil PC algoritmen ikke være i stand til at fastlægge retning på alle variable, hvorfor nogle forbindelser (pile retninger) vil blive genereret tilfældigt. Det må derfor bedømmes om nogle af de strukturelle sammenhænge, som er fundet, virker ulogiske. Hvis dette er tilfældet kan man forsøge at gentage den strukturel sammenhængsanalyse, idet det er muligt *a priori* selv at definere sammenhænge og retninger, udfra en ekspertviden om mest logiske eller realistiske årsagvirkningssammenhænge.

Det er dokumenteret at traditionelle strukturelle læringsalgoritmer med begrænsende betingelser giver korrekte sammenhænge under forudsætning af at datasætttene er uendeligt store, at testene (målinger) er perfekte og at der ikke forekommer retningsorienterede links som går i ring (der kræves en såkaldt 'directed acyclic graph', DAG). Hvis datasættene derimod er begrænsede, giver disse læringsalgoritmer imidlertid ofte for mange udsagn om betingede uafhængigheder, og kan fejlagtigt undlade at identificere vigtige sammenhænge. Der skal ofte mange tusind datapunkter til en sikker bestemmelse af en BN struktur alene ud fra data, men kombineret med apriori definerede sammenhænge/links udfra ekspertviden, kan et mere begrænset datasæt som f.eks. kvadratnetsdataene give brugbare strukturer.

PC algoritmen arbejder relativt hurtigt, men for den langsommere NPC algoritmen er den resulterende graf generelt en bedre beskrivelse af de betingede uafhængighedsrelationer i data. Dette gør sig især gældende for små datasæt hvor NPC algoritmen bør foretrækkes.

De første trin i en praktisk analyse af strukturel sammenhæng består i:

- Udvælgelse af systemvariable og data som skal indgå i analysen
- Definition af tilstande for hver enkelt systemvariabel og organisering af data i samlede datasæt f.eks. udfra de intervaller som definerer de forskellige tilstande (se Figur 15.1)
- Analyse af retningsbestemte links og tilhørende CPT'er, se Figur 15.2, 15.3 og 15.4)

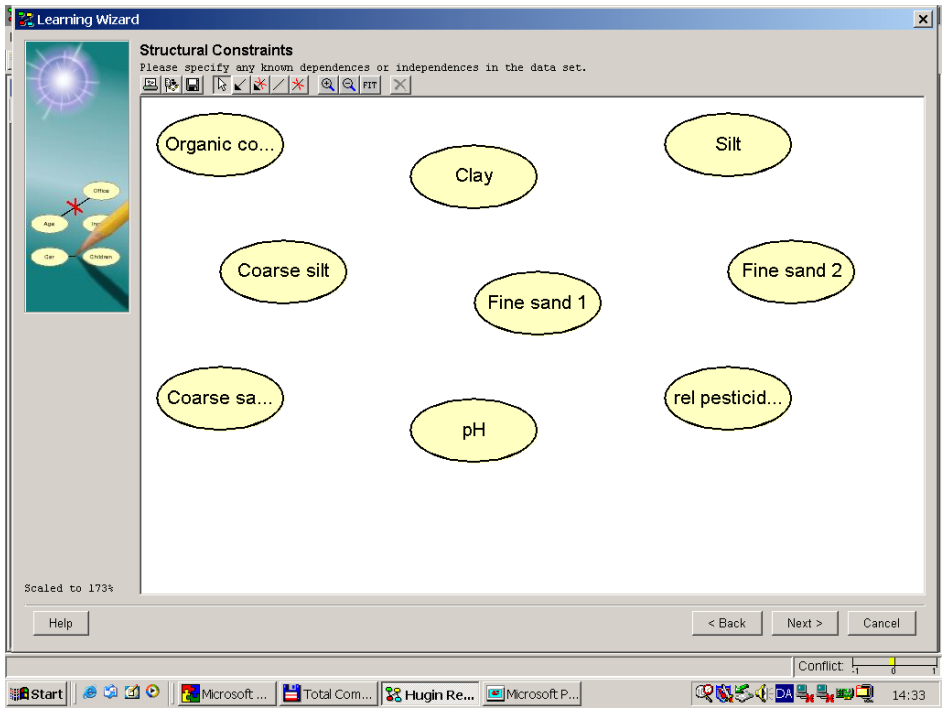


- **Figur 15.1**. Første skridt i den strukturel analyse er at udvælge systemvariable (organisk stof, ler, silt, groft silt, fint sand 1, fint sand 2, groft sand, pH og relativ pesticidudvaskning) og gruppere data i tilstande udfra fastlagte intervaller (f.eks pH 4.5-5.5, 5.5-6, 6-7 og 7-9). Konkret inddeles data for hver variabel her i fire intervaller. Dette er gjort interaktivt ved hjælp af "Hugin", og med det viste resultat. Valgte af intervaller har stor betydning for den strukturelle læring.

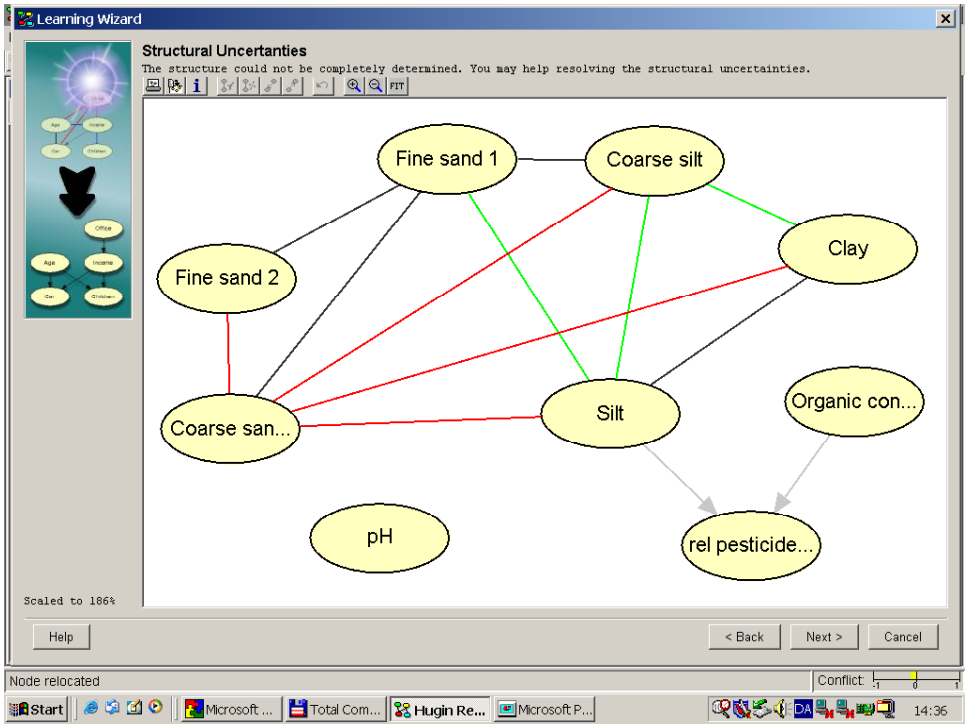
Indledningsvis vises blot variablene (figur 15.2), hvorefter man som bruger kan tilføje kendte afhængigheder eller uafhængigheder. Hugins forslag til sammenhænge afhænger af de valgte intervaller, hvorfor programmet fx. vil kunne foreslå usandsynlige sammenhænge mellem silt- og lerindhold. I sådanne tilfælde må brugeren selv tilføje en uafhængighed mellem variablene. En anden mulighed er at brugeren ønsker at analysere betydningen af nogle parametre mens relationen mellem andre fastholdes (fx. kan de vigtige parametre organisk kulstof, ler og silt fastholdes i en relation til simuleret pesticidudvaskning mens betydningen af de øvrige parametre undersøges).

Med det relativt lille antal datasæt på ca. 150 kvadratnetsprofiler kan den statistiske analyse resultere i noget tilfældige sammenhænge frem for logiske og reelle. Derfor bør resultaterne

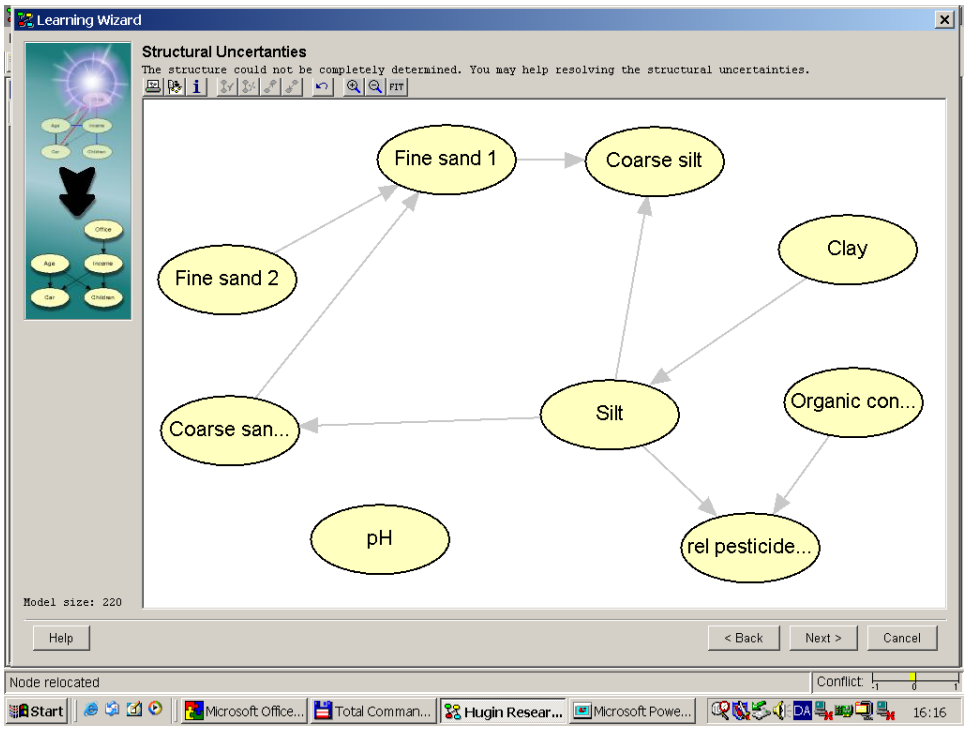
bedømmes kritisk undervejs mens strukturen fastlægges, evt. gennem gentagelser af operationerne i den strukturelle læring med nye forudsætninger (illustreret i figur 15.3 og 15.5), indtil et tilfredsstillende resultat foreligger, figur 15.4.



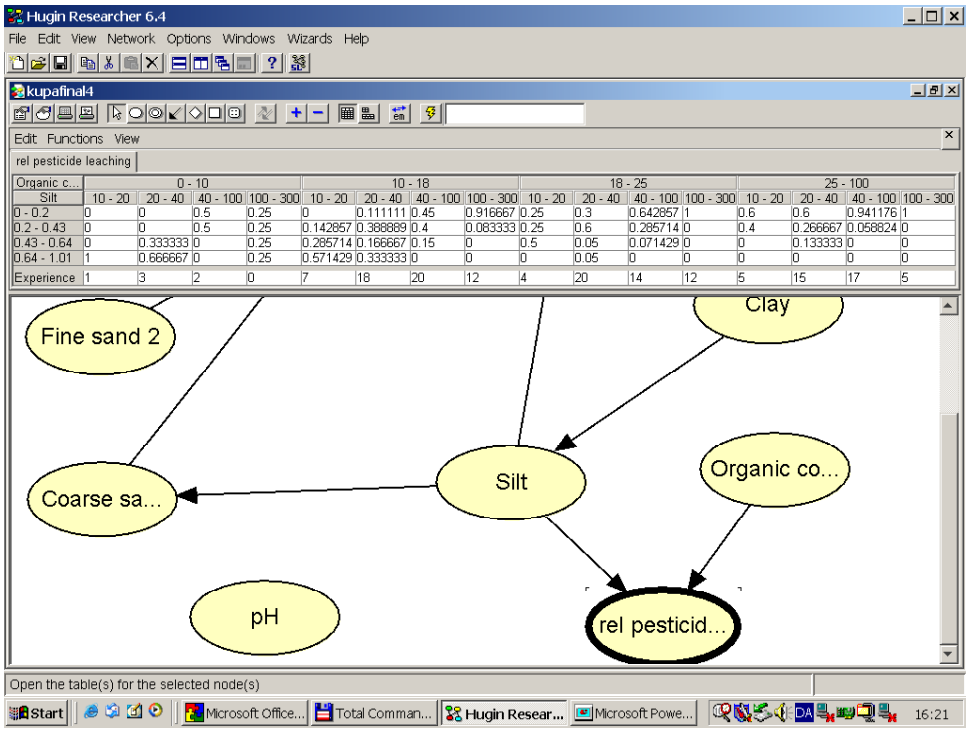
***Figur 15.2**. Illustration af systemvariablene før der er fastlagt sammenhænge eller uafhængigheder udfra foreliggende datasæt.*



***Figur 15.3**. Programmet viser stærke retningsbestemte links (grå pile) hvor der udfra datasæt er stor årsag-virknings afhængighed mellem nogle af variablene, mens der for øvrige sammenhænge må foretages en manuel tilføjelse af retningsbestemte links.*



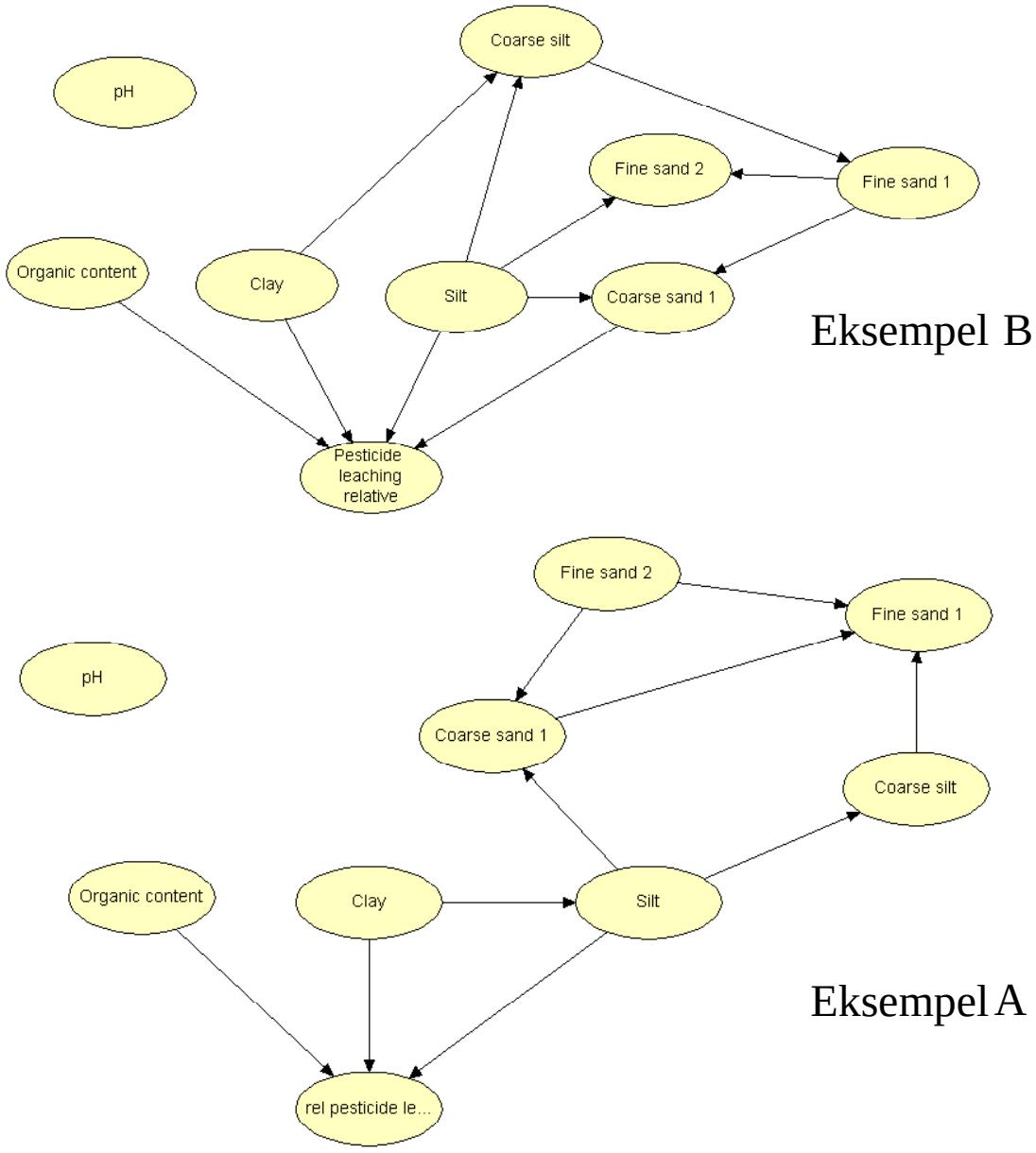
***Figur 15.4**. Det resulterende BN efter at alle retningsbestemte links er definerede enten udfra informationsværdien i datasæt af læringsrutinen i Hugin eller ved manuelle apriori definitioner af brugeren. Der er ingen sammenhænge til pH fordi pH tilsyneladende varierer uafhængigt af alle øvrige variable.*



***Figur 15.5**. Variablen 'rel pesticid…' (relativ pesticid udvaskning) har to styrende "ophavsparametre": 'Silt' and 'Organic co…' (organisk kulstof). Nogle af sammenhængene er ret svage*

(hvilket også fremgår af tælleren "Experience" i CPT'en). CPT'en ville antagelig kunne forbedres ved også manuelt at indlægge sandsynligheder udfra ekspert viden, f.eks. der hvor tælleren viser at CPT-kolonne er baseret på et fåtal af datasæt (f.eks. organisk stof i intervallet 0-10 og silt 100-300 som har 'experience' = 0 => p = 0.25 for de fire tilstande).

Eksempler på anvendelse af BBN som beslutningsstøtte system med henblik på beskyttelse af grundvand mod udvaskende pesticid.



Figur 15.6 A og B. *To forskellige resultater af strukturelle sammenhænge i data. I A er der tre parametre der har betydning for den relative pesticid udvaskning mens der i B er fire influerende parametre. pH har ikke direkte betydning hverken i A eller B. Dette skyldes antagelig at udvaskningen er fastlagt på grundlag af jordens bindings- og transportegenskaber, mens der*

ikke er taget hensyn til den mere pH-afhængige nedbrydning i modelsimuleringerne (MACRO).

Det første eksempel, A i figur 15.6, er BN etableret ved en strukturel læring, hvor der er tillagt afhængigheder mellem hver af systemvariablene: 'organisk stof', 'ler' og 'silt' og 'relativ pesticidudvaskning'. Alle andre sammenhænge er blevet bestemt af Hugins strukturelle læringsalgoritme og under sideløbende interaktive input fra eksperter. I det andet eksempel, B i figur 15.6, er BN etableret under antagelse af uafhængighed mellem 'silt' og 'ler', støttet med bestemmelser ved strukturel læring af Hugin støtte af ekspertudsagn vedrørende hvilke sammenhænge, der bør inkluderes og hvilken retning de har.

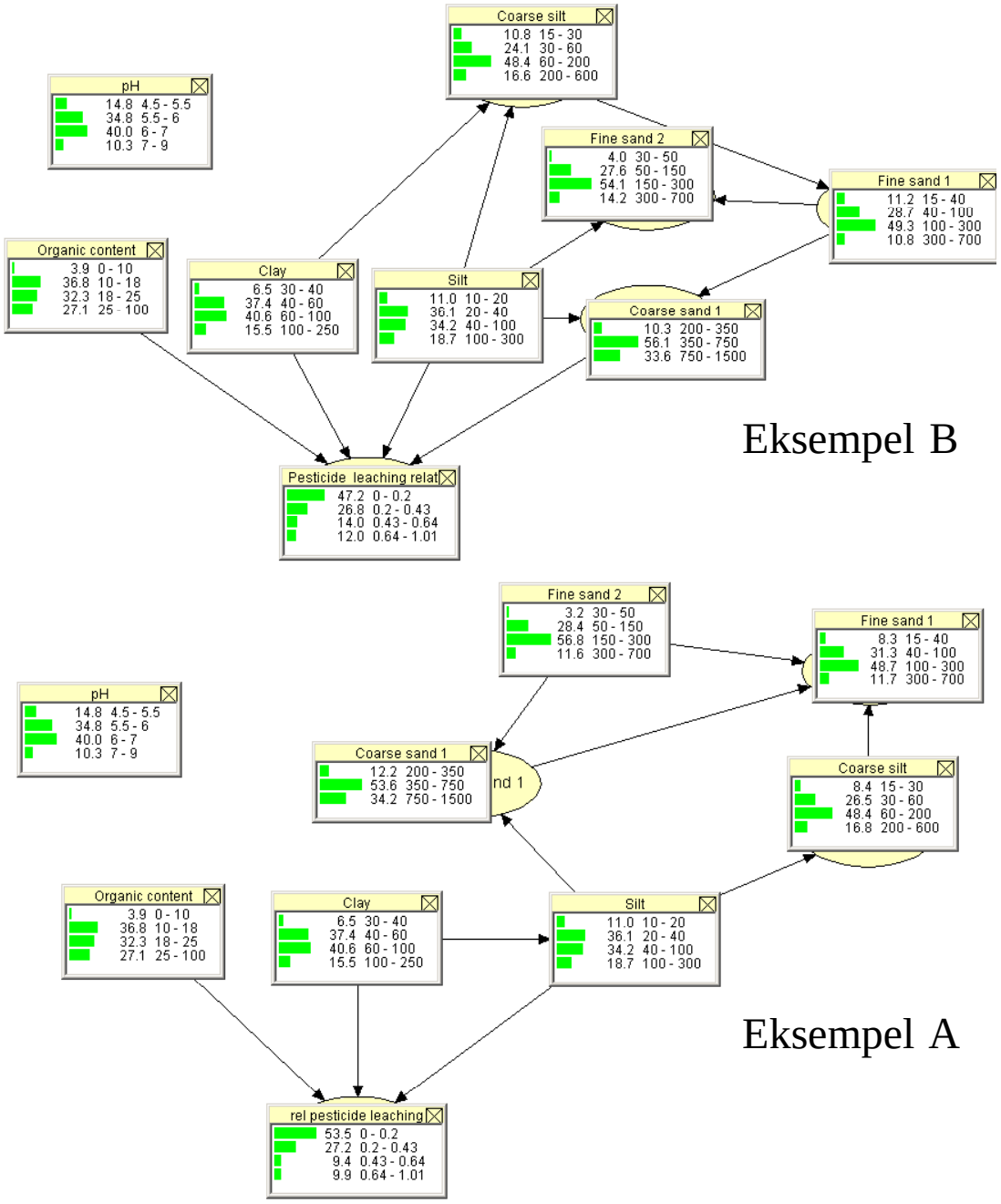
I eksempel A i figur 15.6 er der kun de tre parametre 'organisk stof', 'ler' og 'silt', der har direkte indflydelse på den relative simulerede udvaskning af pesticid. 'ler' og 'silt er imidlertid ikke uafhængige af hinanden, idet lerindholdet influerer på siltindholdet, som vist med pilen. Siltindholdet influerer på andre parametre: 'fint sand 1 & 2', 'groft sand' og 'groft sand 11'. Da disse variable er forholdsvis nemme at måle vil oplysninger om dem eventuelt kunne erstatte data vedrørende siltindholdet.

I eksemplet, figur 15.6 B, er der ingen sammenhæng mellem indholdet af 'Ler' og 'Silt'. Under denne forudsætning resulterer strukturanalysen i en mere kompleks sammenhæng, hvor fire variable har indflydelse på den relative pesticidudvaskning. Også i dette eksempel viser analysen at 'silt', 'ler' og 'organisk stof' er væsentlige for forudsigelse af udvaskningen og dermed væsentlige kortlægningsparametre, men eksempel B viser yderligere at 'groft sand 1' variablen må tages i betragtning, når strukturanalysen forudskikker at der ikke er indbyrdes afhængighed mellem parametrene 'ler' and 'silt'. Variablen 'groft silt' er afhængig af både 'ler' og 'silt', og influerer selv på variablen 'fint sand 1'.

Figur 15.7 viser resultatet af den strukturelle analyse (systemvariable og retningsbestemte links) og resulterende sandsynlighedsfordelinger for samtlige variable i nettene for eksemplerne A and B (figur 15.6). Der kan nu eksperimenteres med følsomheden overfor ændringer fra at en variabel er usikkert bestemt i form af en sandsynlighedsfordeling for at antage en af de forskellige intervaller, til at tilstanden er kendt f.eks. udfra en måling og følgeeffekterne heraf for alle andre variable kan så beregnes ved hjælp af Hugin, se eksperimenter for eksempel A i figur 15.8 – 15.10.

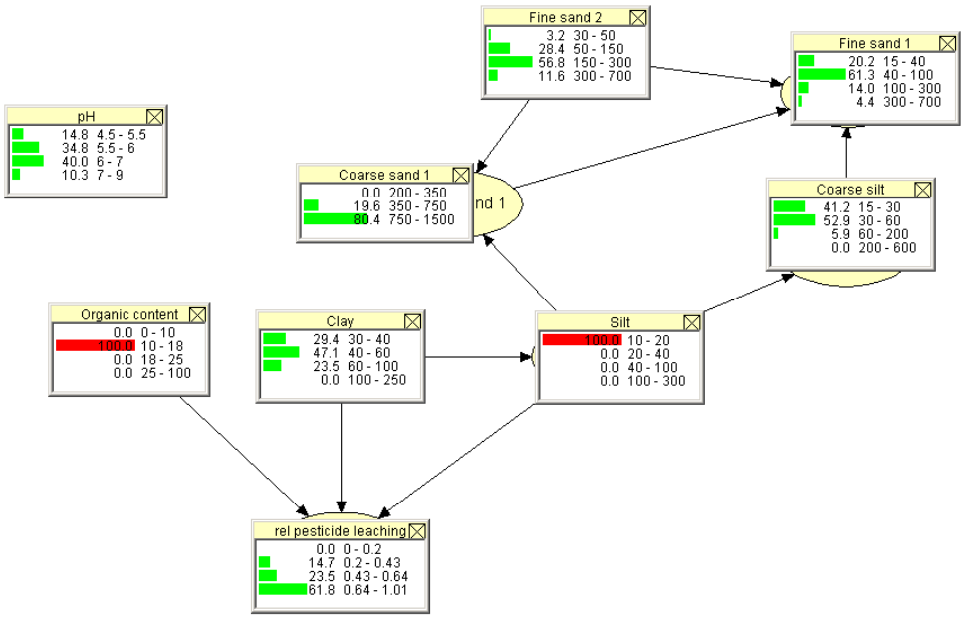
Hvis for eksempel tilstanden for variablen 'silt' (figur 15.8) udfra målinger kan fastsættes til at ligge i det laveste interval (10-20 kg/m^2) og tilstanden på 'organisk stof' til det næst laveste interval (10-18 kg/m^2) så medfører det en form for alarm, hvor der er 61,8% risiko for at den relative simulerede pesticidudvaskning fra jordtyper, hvor systemvariable har de valgte tilstande (intervaller), ligger i den mest sårbare klasse, og 23,5% risiko for at sårbarheden af jordtypen ligger i den næsthøjeste klasse. Figuren viser også at de tillagte 'silt'-værdier har ændret sandsynlighedsfordelinger for 'ler', 'fint sand 1', 'groft sand 1' og 'groft silt'. Kun den oprindelige fordeling af 'fint sand 2' er uændret.

Alternativt kan BN i eksempel A bruges til at undersøge den afledte effekt i de øvrige variable ved en bestemmelse af de parametre som er lettest at måle i laboratoriet ('fint sand 1 & 2', 'groft sand 1' og 'groft silt'), figur 15.9. I det viste eksempel influerer ændringen alle variable undtagen 'organisk stof'.

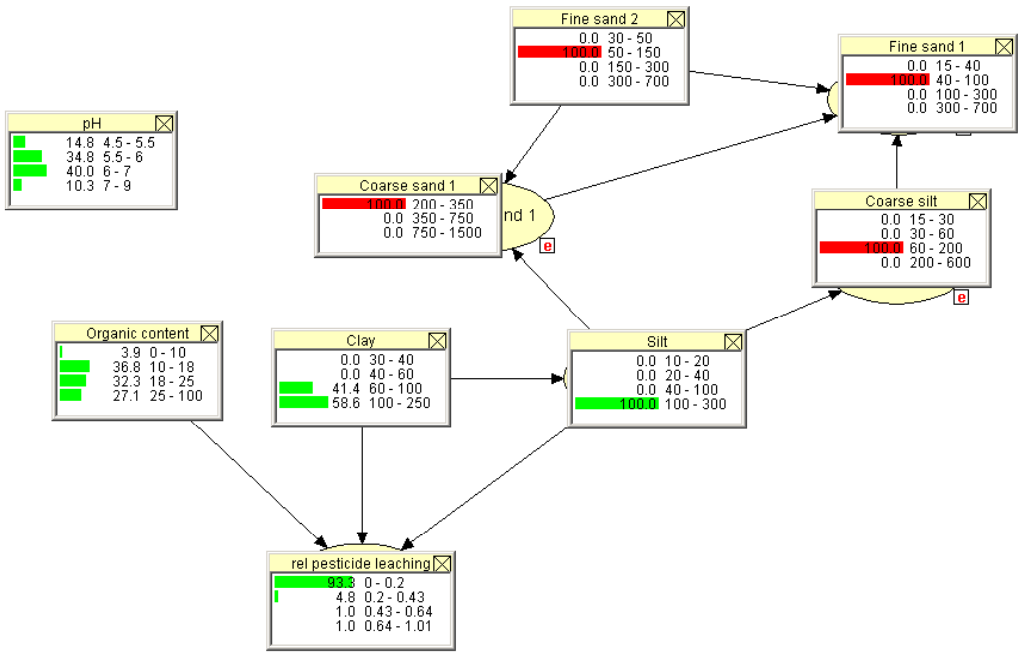


Figur 15.7 A og B. Strukturel læring (systemvariable og retningsbestemte links) og simulerede sandsynlighedsfordelinger for tilstande for de to eksempler i figur 6.26. De jordprofiler som er mest sårbare overfor udvaskning af pesticid ligger i intervallet 0.64-1.01 (9,9 % af profilerne i eksempel A). Også følsomme, om end i mindre grad, er profiler som hører til intervallet 0.43-0.64 (9.4 % i eksempel A). Der er altså i denne beregning 18 % af de analyserede profiler som falder i de to mest pesticidfølsomme kategorier.

Den relative pesticidudvaskningsindikator ('rel pesticide leaching') viser, med oplysninger som er lagt ind i den strukturelle analyse, at der kun er ringe risiko for at jorden er følsom overfor udvaskning (= 0.01 eller 1 pct.). Der er således basis for at beslutte, hvorvidt en sådan lav procentvis risiko er acceptabel eller om der skal indsamles yderligere dokumentation om enten 'ler' eller 'organisk stof' for at gøre sandsynligheden for at jordtypen er meget sårbar endnu mindre end 1 pct.

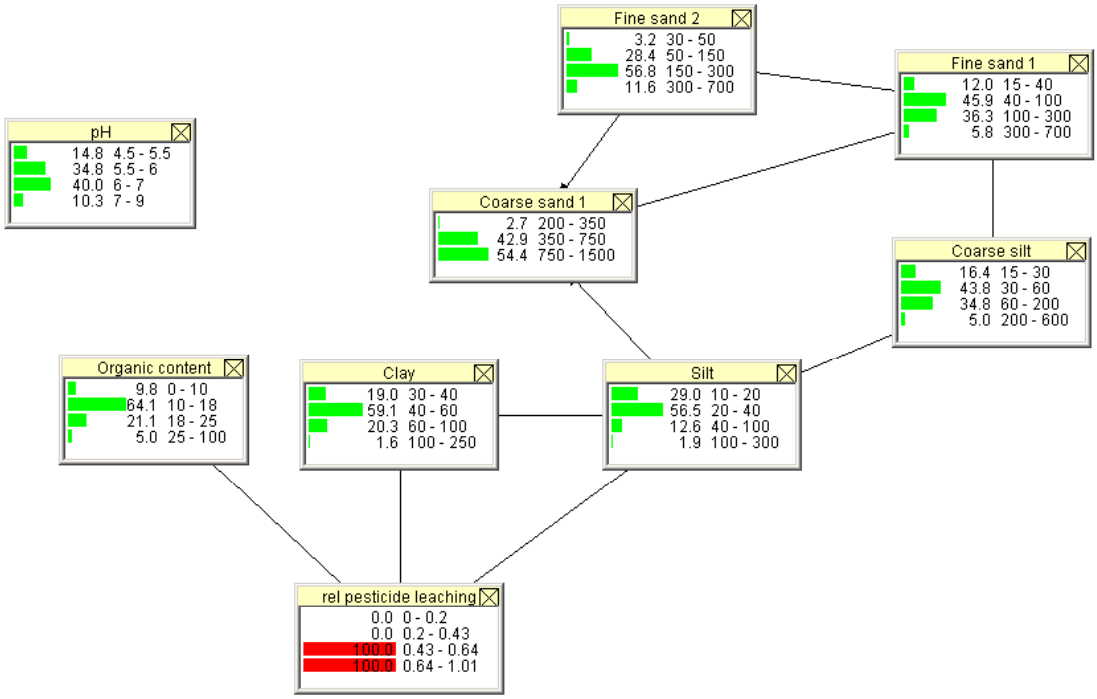


***Figur 15.8**. Eksempel på hvordan der kan eksperimenteres med effekten af at tilføje evidens (kendskab til udvalgte systemvariable udfra fx målinger - røde bjælker) og opdatering af øvrige sandsynligheder i nettet og prognosen for sårbarheden overfor pesticid udvaskning med denne sikre viden (med udgangspunkt i BN for eksempel A).*



***Figur 15.9**. BN for eksempel A hvor der er tillagt kendt tilstand for 'groft sand 1', 'fint sand 2', 'fint sand 1' og 'groft silt'. Effekten er at der er 41,4 % chance for at 'ler'-indholdet ligger i intervallet 80-100 og 58,6 % chance for at det ligger i intervallet 100-250. De simulerede forhold i dette eksperiment medfører dermed at der er ringe risiko for udvaskning af pesticid.*

Endelig er det muligt at indføje en sandsynlighedsfordeling (likelihood) for en eller flere udvalgte systemvariable i BN og analysere effekten på de øvrige variable under denne reviderede antagelse. Dette er eksemplificeret i figur 15.10 for BBN i eksempel A, idet det antages at der er det højeste eller næsthøjeste niveau af relativ pesticidudvaskning (enten 0.43-0.64 eller 0.84-1). Her får man et indtryk af hvilke 'intervaller' der er mest sandsynlige for de øvrige variable f.eks. 'organisk stof', 'ler' og 'silt' som i dette tilfælde har størst sandsynlighed for at være i næstlaveste interval (56 - 64 %) for de tre variable.



Figur 15.10*. Eksempel på effekten af at indføje likelihood fordeling (fifty – fifty for næsthøjeste og højeste sårbarhedsinterval) (BN for eksempel A).*

Sammenfatning vedrørende anvendelse af strukturel analyse som beslutningsværktøj

Eksperimenterne med BN og algoritmen for strukturel læring i dette bilag har haft til formål at demonstrere et muligt værktøj til beslutningsstøtte. Eksemplerne er således ikke projektets resultater. De præsenterede BN-eksempler (A og B) er to alternativer ud af et større antal forsøg med forskellige sammenhænge/manglende sammenhænge. I tilfælde hvor der ikke aktivt indlægges afhængigheder mellem variable resulterer de strukturelle analyser i at to ('silt' og 'organisk stof') eller tre ('silt', 'ler' og 'organisk stof') variable til beskrivelse af relativ pesticid udvaskning, afhængig af de tillagte værdier og forbindelser/manglende forbindelser og afhængighedsretninger.
Eksemplerne er beregnet med PC algoritmen selv om NPC algoritmen muligvis ville kunne give et endnu bedre resultat på grund af det relativ begrænsede datagrundlag (150 datasæt).
De BNs som er udviklet her bør vurderes nøjere forud for en eventuel praktisk anvendelse, idet CPT'erne for nogle af sammenhængene er bestemt på et ret svagt grundlag (baseret på <5 oplysninger). Trods dette forbehold vurderes BNs imidlertid at være et fleksibelt værktøj til at analysere pesticidudvaskning idet der er mulighed for at trække på dels vidensgrundlaget fra kvadratnetsprofilerne og samtidig opdatere nettene med nye målinger fra et konkret område. Hertil kommer at kvalificerede erfaringer også kan indbygges i form af likelihood. BN ud-

gør dermed et anvendeligt værktøj til at afdække og formidle et bedste estimat for pesticidsårbarheden, og samtidig illustrere usikkerheder på dette skøn.

Det er af særlig betydning i forbindelse med zonering af særligt pesticidfølsomme arealer at BN eksplicit kan kvantificere og formidle usikkerhed. Derved er det muligt at bedømme den relative udvaskning af pesticid på grundlag af alle data indenfor et givet areal og i tilslutning hertil at kombinere denne vurdering med andre datasæt fra de landsdækkende kvadratnetsprofiler og yderligere parametre (fx. landbrugspraksis, nedbrydning af pesticider, udbringning af pesticider, grundvands moniteringsdata og socio-økonomiske forhold).

BNs er frem alt velegnet som et dialogværktøj der kan give overblik og integration, og samtidig påpege hvilke yderligere data der kan være behov for at indsamle