Directed acyclic graphs and Bayesian networks – promising tools to assess links between water quality time series

Photo: C. Hackmann, Krycklan (Sweden)

Benny Selle (Berlin and Tübingen) and Klaus-Holger Knorr (Münster)

Beuth University of Applied Sciences Berlin, Universities of Münster and Tübingen, Germany

Background

During the last decades, both dissolved organic carbon (DOC) and Fe concentrations as well as pH increased in many streams of northern America and Europe.

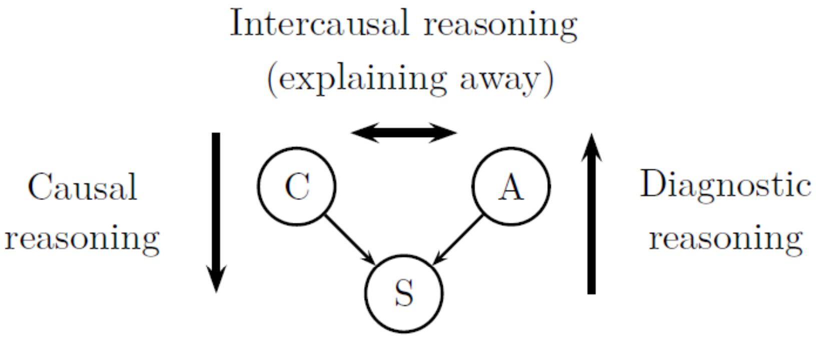
The interplay between the upward trends of Fe, DOC and pH remains scientifically disputed.

Increasing DOC concentrations are practically relevant for C storages of catchment soils and water quality of streams feeding drinking water reservoirs.

Organic carbon (OC) and Fe, bound as Fe-OC-associations in catchment soils, may be mobilised via redox processes and/or pH increase, which are know processes for mineral soils but they were less intensively investigated for organic soils.

Directed acyclic graphs (DAG) and Bayesian Networks are an active field of research in artificial intelligence with main applications in computer and life sciences; but there are a few promising studies on water quality too (e.g. Stow and Borsuk, 2003, Ecosystems; Alameddine et al., 2011, Environ Modell Softw).

DAGs and Bayesian networks
(example taken from A.L. Madsen, Hugin Expert A/S)

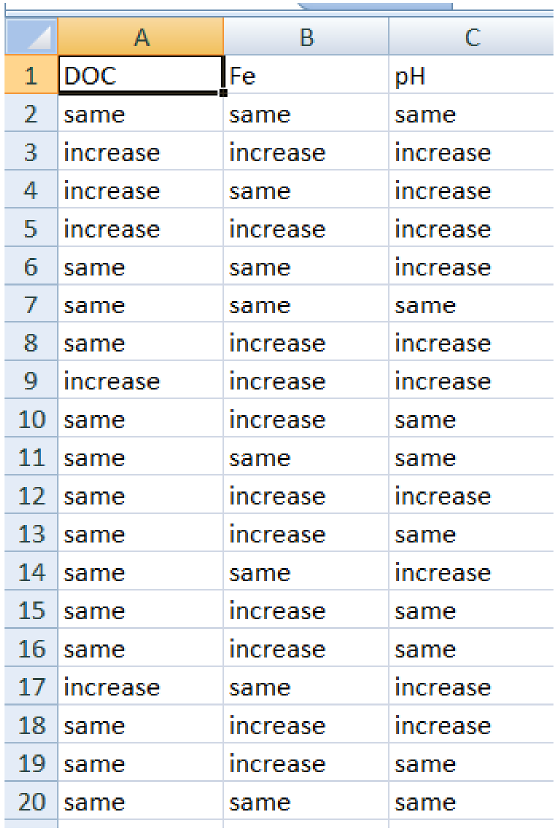


–Cold (C) and allergies (A) may cause sneezing (S), so S can be used to diagnose both C and A. If we observe someone to sneeze and we know that he/she has an allergy then this provides us with some information on whether or not he/she has a cold (intercausal reasoning, C and A become correlated if S is known but C and A are independent if S is unknown).

–A Bayesian network, i.e. a DAG plus a conditional probability table for each node (C,S,A), allows reasoning/inference using Bayesian calculus.

Our analysis was based on long term trends of DOC, Fe and pH for 73 German streams feeding drinking water reservoirs

(data taken from Musolff et al. 2017*)



	A	B	C
1	DOC	Fe	pH
2	same	same	same
3	increase	increase	increase
4	increase	same	increase
5	increase	increase	increase
6	same	same	increase
7	same	same	same
8	same	increase	increase
9	increase	increase	increase
10	same	increase	same
11	same	same	same
12	same	increase	increase
13	same	increase	same
14	same	same	increase
15	same	increase	same
16	same	increase	same
17	increase	same	increase
18	same	increase	increase
19	same	increase	same
20	same	same	same

“increase” means significantly increasing linear trends during the period 1993-2013 (Mann-Kendall test at 5% significance level)

whereas

“same” are insignificant or significantly decreasing linear trends during the period 1993-2013 (Mann-Kendall test at 5% significance level)

* Musolff et al. (2017), Global Change Biology, 23 (5), 1891–1901.

Observations and statistical tests can be used to construct DAGs and subsequently Bayesian networks, e.g.* with 3 variables (A,B,C) the following DAGs correspond to a set of conditional (|) dependence ($\not\perp\!\!\!\perp$) and independence ($\perp\!\!\!\perp$) statements

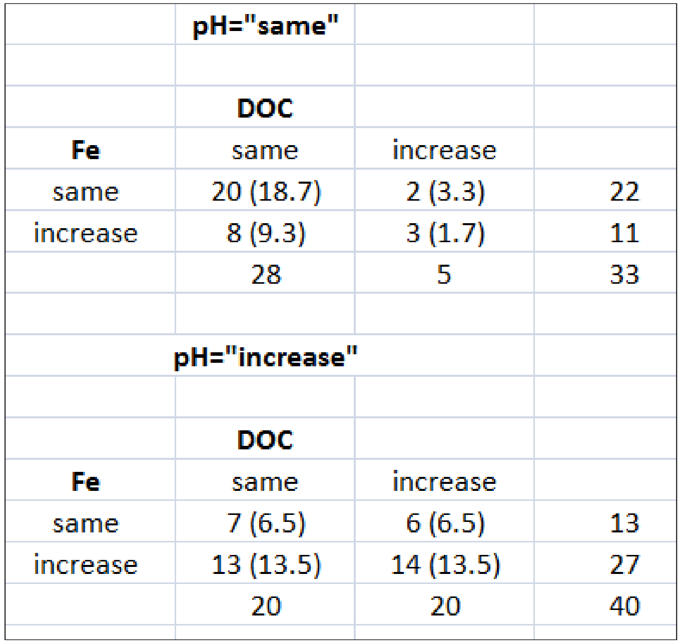
$A \rightarrow B \rightarrow C$	$A \not\perp\!\!\!\perp B, A \not\perp\!\!\!\perp B \mid C, B \not\perp\!\!\!\perp C, B \not\perp\!\!\!\perp C \mid A, A \not\perp\!\!\!\perp C$ $A \perp\!\!\!\perp C \mid B$
$A \leftarrow B \leftarrow C$	$A \not\perp\!\!\!\perp B, A \not\perp\!\!\!\perp B \mid C, B \not\perp\!\!\!\perp C, B \not\perp\!\!\!\perp C \mid A, A \not\perp\!\!\!\perp C$ $A \perp\!\!\!\perp C \mid B$
$A \leftarrow B \rightarrow C$	$A \not\perp\!\!\!\perp B, A \not\perp\!\!\!\perp B \mid C, B \not\perp\!\!\!\perp C, B \not\perp\!\!\!\perp C \mid A, A \not\perp\!\!\!\perp C$ $A \perp\!\!\!\perp C \mid B$
$A \rightarrow B \leftarrow C$	$A \not\perp\!\!\!\perp B, A \not\perp\!\!\!\perp B \mid C, B \not\perp\!\!\!\perp C, B \not\perp\!\!\!\perp C \mid A, A \not\perp\!\!\!\perp C \mid B$ $A \perp\!\!\!\perp C$

* example taken from A.L. Madsen, Hugin Expert A/S.

Note that the first three models above are equivalent model structures, i.e. DAGs compatible with the same set of statistical (in)dependencies in the data; the corresponding Bayesian networks have the same model quality (e.g. same Akaike information criterion) and they behave the same in terms of diagnostic and prognostic reasoning.

Tests of statistical independence for data set (Chi-squared tests at 5% significance level)

1. DOC depends on pH (if DOC is “same” then pH tends to be “same”, likewise for increasing pH and DOC tends).
2. Fe depends on pH.
3. Fe depends on DOC.
4. DOC depends on pH given the state of Fe, i.e. if Fe is set to “same” or “increase”
5. Fe depends on pH given DOC.
6. Fe and DOC are *independent* given pH (tables on the right, expected counts - if DOC and Fe were independent - are given in brackets).



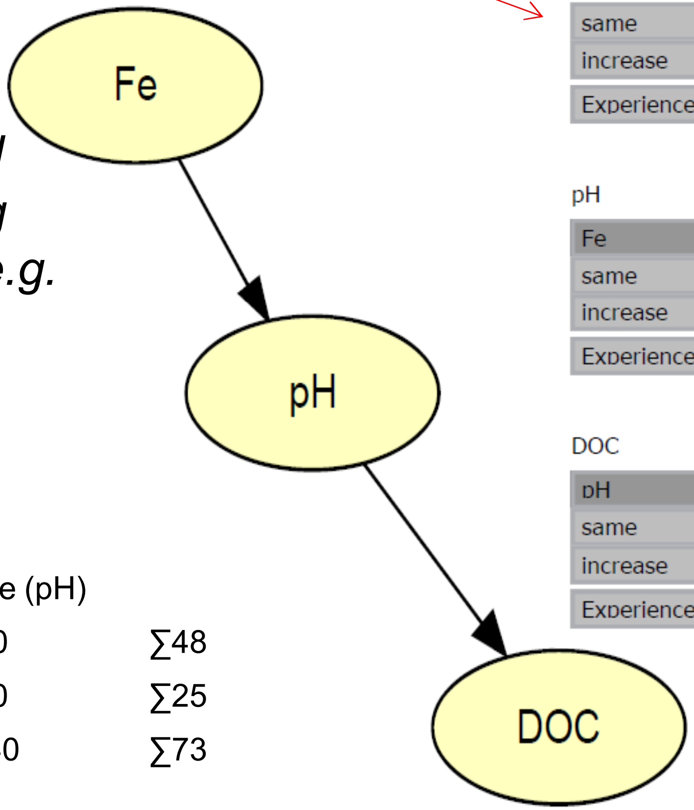
	pH="same"		
	DOC		
Fe	same	increase	
same	20 (18.7)	2 (3.3)	22
increase	8 (9.3)	3 (1.7)	11
	28	5	33
	pH="increase"		
	DOC		
Fe	same	increase	
same	7 (6.5)	6 (6.5)	13
increase	13 (13.5)	14 (13.5)	27
	20	20	40

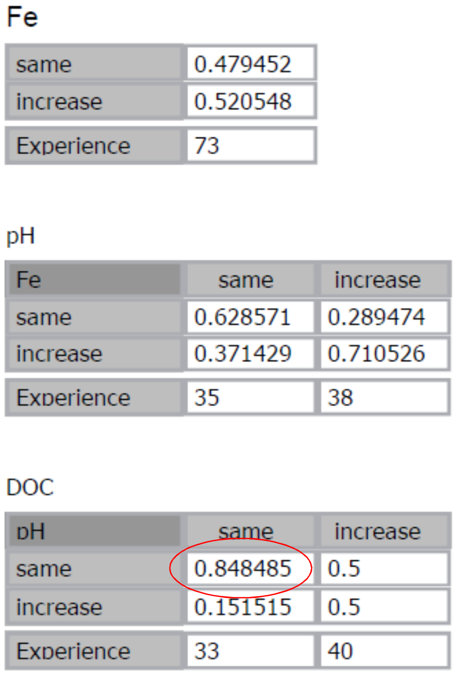
Test results (1. to 6.) imply a DAG of either Fe → pH → DOC or
Fe ← pH → DOC or (implausible)
Fe ← pH ← DOC (DOC is unexplained)

Setup of Bayesian Network (= DAG + Conditional Probability Tables)

Entries of conditional probability tables (on the far right) computed from data (see table below) using maximum likelihood estimation, e.g.
P(DOC=same|pH=same) =
P(¬DOC|¬pH) = 28/33
= 0.848485

	pH		
DOC	same (¬pH)	increase (pH)	
same (¬DOC)	28	20	∑48
increase (DOC)	5	20	∑25
	∑33	∑40	∑73





Fe

same	0.479452
increase	0.520548
Experience	73

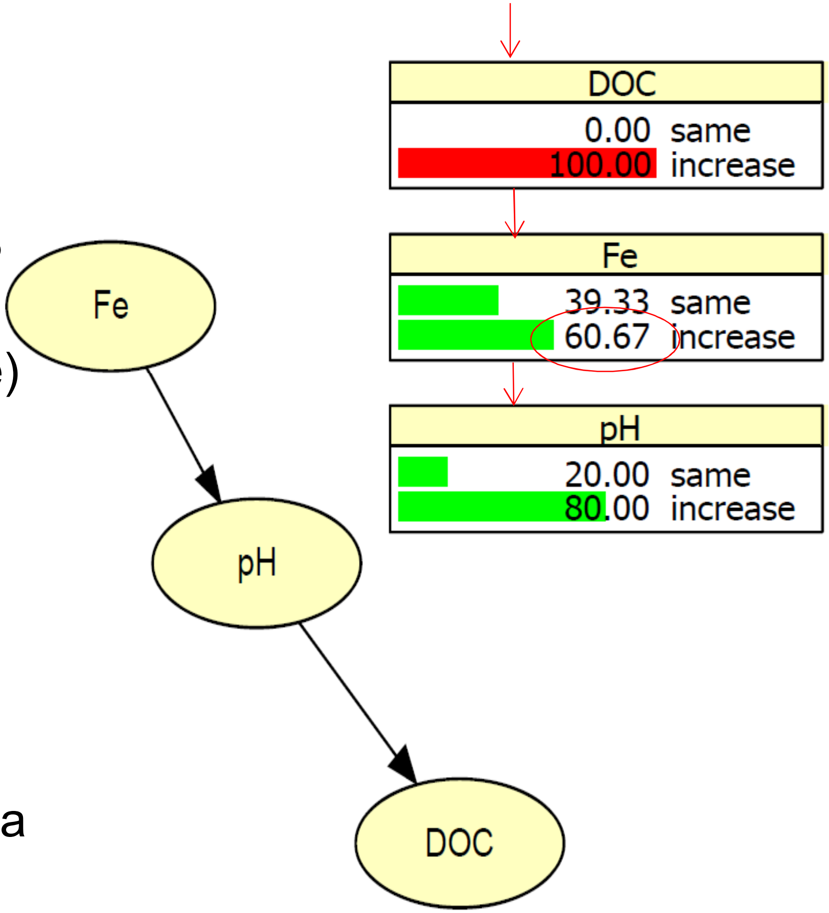
pH

Fe	same	increase
same	0.628571	0.289474
increase	0.371429	0.710526
Experience	35	38

DOC

pH	same	increase
same	0.848485	0.5
increase	0.151515	0.5
Experience	33	40

Inference using Bayesian Network (diagnostic reasoning),
e.g. probabilities of states of both pH and Fe given DOC="increase"



Example: calculation of P(Fe|DOC)

1. marginalisation

$P(DOC,Fe) = P(DOC|pH) \times P(pH|Fe) \times P(Fe) + P(DOC|\neg pH) \times P(\neg pH|Fe) \times P(Fe) = 0.2078$

$P(DOC,\neg Fe) = P(DOC|pH) \times P(pH|\neg Fe) \times P(\neg Fe) + P(DOC|\neg pH) \times P(\neg pH|\neg Fe) \times P(\neg Fe) = 0.1347$

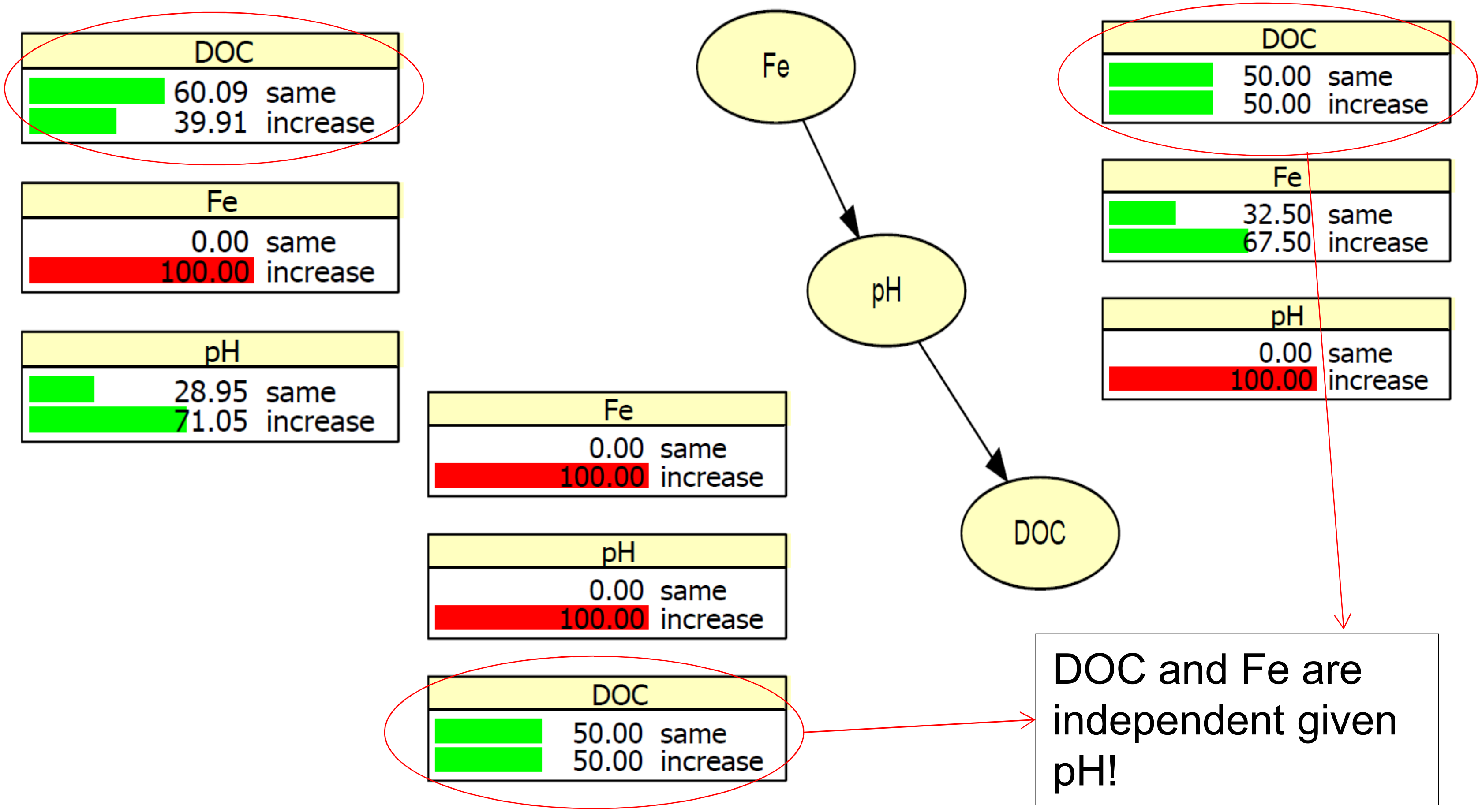
2. normalisation:

$P(DOC) = P(DOC,\neg Fe) + P(DOC,Fe) = 0.3425$

3. conditioning:

$P(Fe|DOC) = P(DOC,Fe)/P(DOC) = 0.1347/0.3425 = \underline{0.6067}$ = "probability to find a stream with an increasing Fe-trend if an increasing trend for DOC was observed"

Bayesian Network: prognostic reasoning, link between both increasing Fe and pH trends and DOC trends appears to be relatively weak



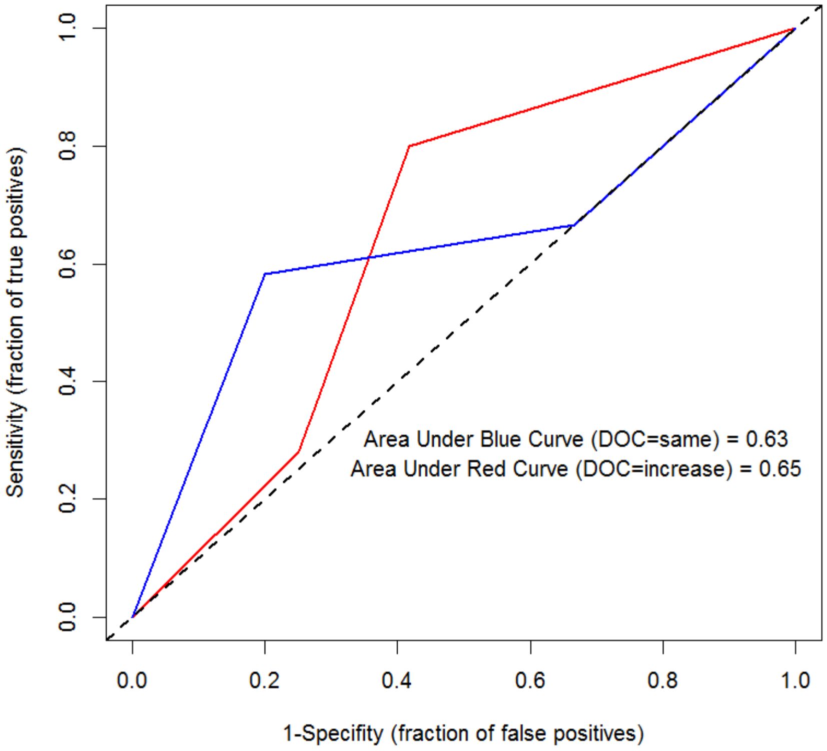
DOC and Fe are independent given pH!

Quality of Bayesian Network: Confusion Matrix for DOC, predicted state is the one with the highest modelled probability (≥0.5)

	Observation		
Model	*increase*	*same*	
increase	7	12	∑19
same	18	36	∑54
	∑25	∑48	∑73

Error Rate = (18+12)/73 = 0.41, i.e. fraction of streams for which DOC trends were incorrectly modelled
Sensitivity (for DOC="increase") = 7/25 = 0.28, i.e. fraction of streams with observed positive DOC trends that were correctly modelled
Specifity (for DOC="increase") = 36/48 = 0.75, fraction of streams that were *not* observed to have a positive trend but were correctly modelled, i.e. DOC="same".

Receiver Operating Characteristic for modelling DOC



Fractions of true and false positives are plotted for different cut-off levels (modelled probability at which particular state is predicted). Area under curve (AUC) is a measure of model quality with a value of 0.5 being a pure random model (dashed line) and a value of 1 for a perfect model. Values of 0.63/0.65 are relatively poor.

Summary and Conclusions

A series of statistical tests generated a set of possible DAGs linking the investigated water quality variables in three possible ways. One DAG (Fe → pH → DOC) was considered useful to explain DOC trends, whereas the other two DAGs were evaluated to be either implausible (Fe ← pH → DOC) or impractical (Fe ← pH ← DOC) as the latter did not explain DOC trends. However, the Bayesian network Fe → pH → DOC only poorly explained increasing DOC trends (low sensitivity and AUC). Furthermore, in the Bayesian network both increasing pH and Fe trends were relatively weakly linked to DOC increases. Nevertheless, the approach taken may provide novel information on the interplay between Fe, pH and DOC if it is modified in the following way. The data set used to build and test the model was, however, so far relatively incomprehensive (only trends) and heterogeneous with catchments from all over Germany. A set of water quality data - also relating to seasonality and discharge dependence of relevant solutes - from catchments with similar characteristics such as sub-catchments of Krycklan in Sweden could give more reliable and interpretable results. For the analysis with additional variables, structure learning algorithms need to be applied.