



Estimating missing data in water quality time series

Aart H Smits¹, Paul K Baggelaar² and Peter G Stoks¹

¹RIWA, Nieuwegein, Netherlands

²Icastat Consultancy, Hengelo, Netherlands

Where we are



IAWR

Umbrella organization
of 3 Associations

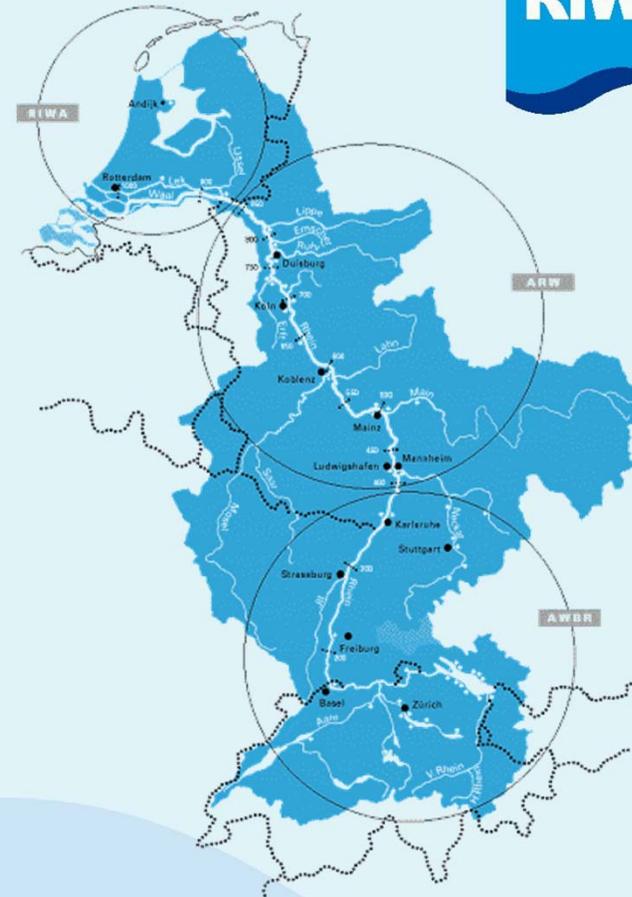
RIWA: Netherlands

ARW: lower Germany

AWBR: upstream Germany,
Switzerland

120 utilities
30 million consumers

Mission: source water quality should allow drinking water production
using near-natural treatment only





RIWA / IAWR

- **Initially “Pressure group” fighting water pollution**
- **Confronting polluters / decision makers with WQ data and demands**
 - *Strategy: actions based on sound science / hard evidence only!*
 - *Gradual shift from confrontation to cooperation*
- **Several Rhine memoranda (latest 2008)**
 - *WQ objectives for pollutants of concern*
 - *recommendations on pollution reduction*



WQ monitoring network

- **Cooperation with Nat'l Dutch and German water authorities**
 - *Harmonized program (WQ variables, methods, data exchange,...)*
- **Four locations in the Dutch part of the Rhine basin**
 - *German-Dutch border, intake sites*
- **28 locations in total**
 - *10 main sites (mostly intakes)*
 - *freq \geq 13; > 300 variables*
- **Trend detection and compliance testing**
 - *Legal standards & DMR Threshold values*
 - *Chemical & biological*

Sampling sites





WQ monitoring network

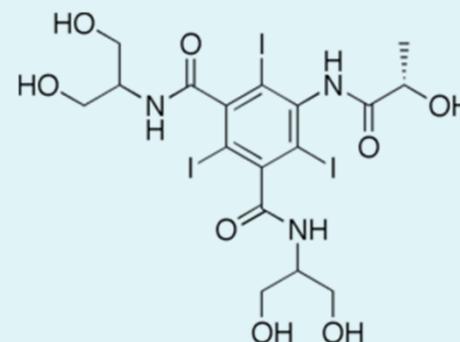
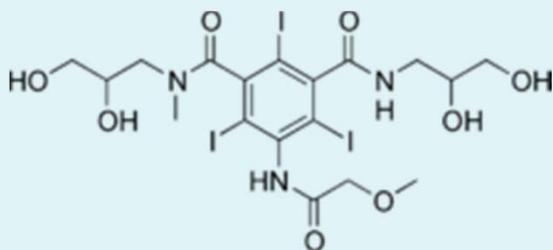
- **Emphasis on “emerging pollutants”**
 - *artificial sweeteners, anti-corrosion agents (dishwashers), fragrances, sunblocks, pharmaceuticals, foam stabilizers (fire extinguishers), insect repellents...*
- **DMR threshold 1 ug/L for stable, polar substances, 0.1 ug/L for biologically active substances**



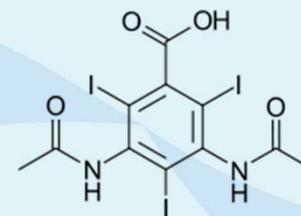
...emerging pollutants...



X-Ray contrast media

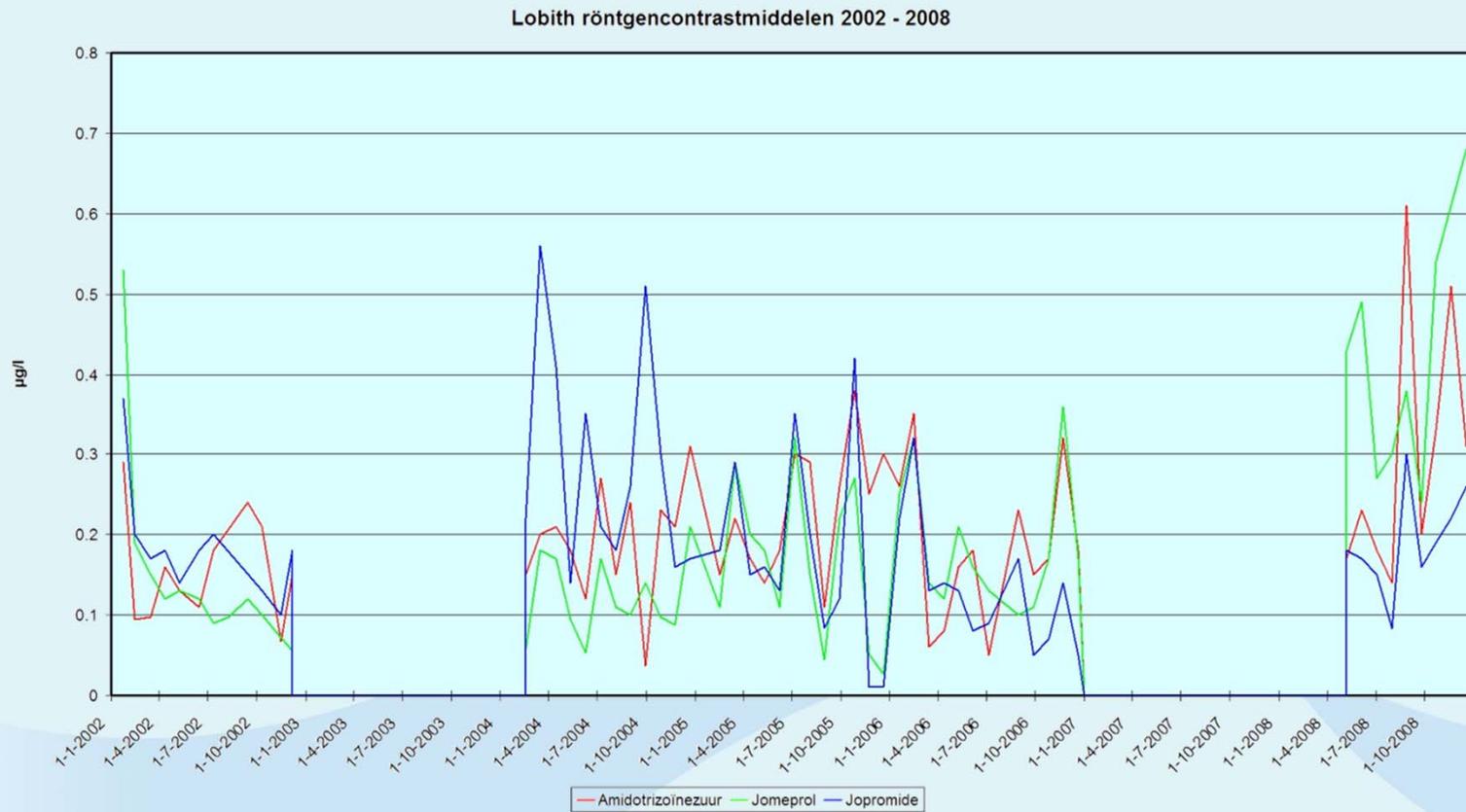


- High production volume, high consumption
- Very stable, very polar
- Not retained in “simple treatment”, difficult to remove in advanced treatment
- Levels well over 0.1 ug/L in Rhine basin





The original time series





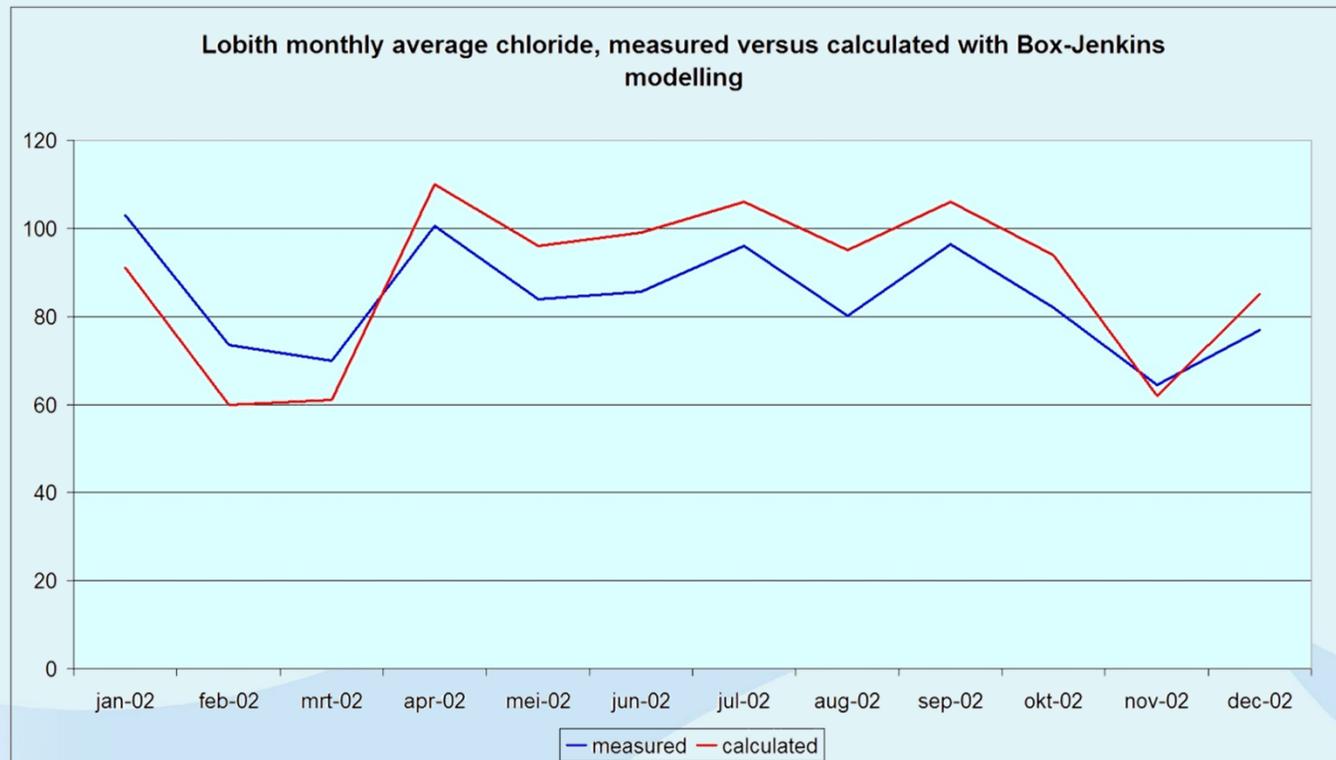
The approach – 1

Box-Jenkins transfer modeling

- Existing time series of chloride at Lobith
- One year of monthly averages taken out
- Recalculation using data from upstream site (Cologne) and water discharge data at Lobith
 - *Average difference ~ 10 mg/l*



Chloride at Lobith: measured vs calculated





The situation





Conclusion

Results promising, however, X-ray agents unacceptable:
Cologne, too, had data gaps

Additional approaches

Trial with Düsseldorf data failed: data set too small

Interpolation between Düsseldorf and Nieuwegein failed: water discharge very variable → correlation “vague”

- *input tributaries*
- *different bypass routes under varying discharge conditions*

Situation





Neural Network approach

Available software package*, runs on pc

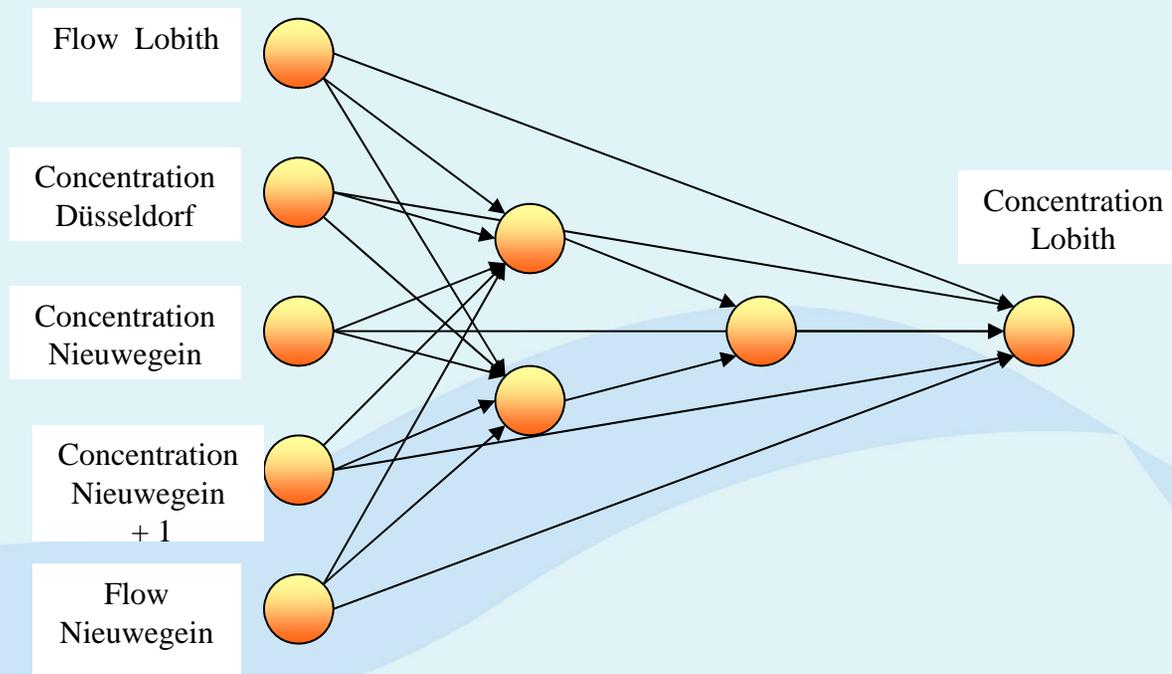
Input variables: X-ray agent data Düsseldorf (upstream) and Nieuwegein (downstream) and discharge data Lobith (target site, inbetween) and Nieuwegein; and X-ray agent data Nieuwegein shifted 1 sampling period

- under low flow conditions these X-ray agent data contain information about Lobith 1 sampling period earlier

* MBP 2.2 (Lopez&Ribero, 2010)

Neural Network approach

More capable of describing non-linear relationships than time series or interpolation methods



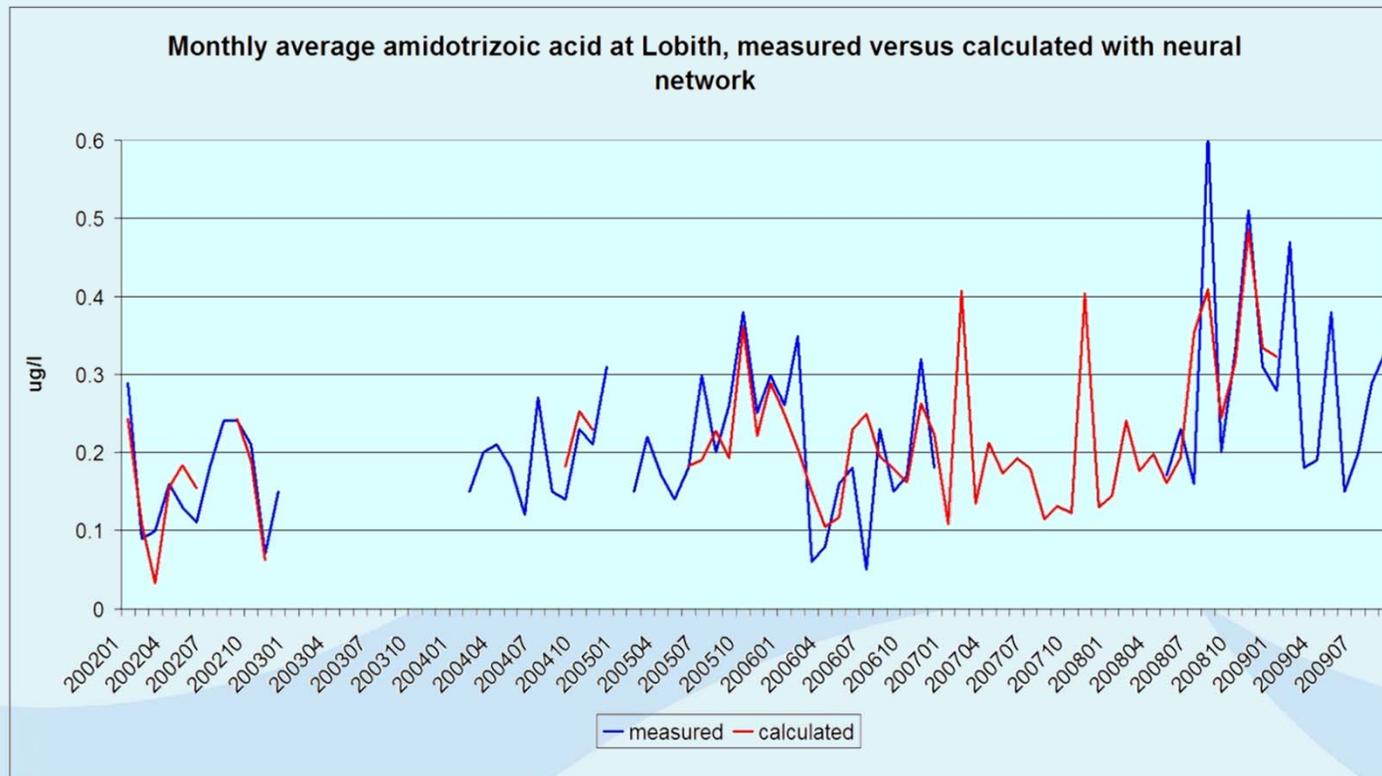


The situation





Results



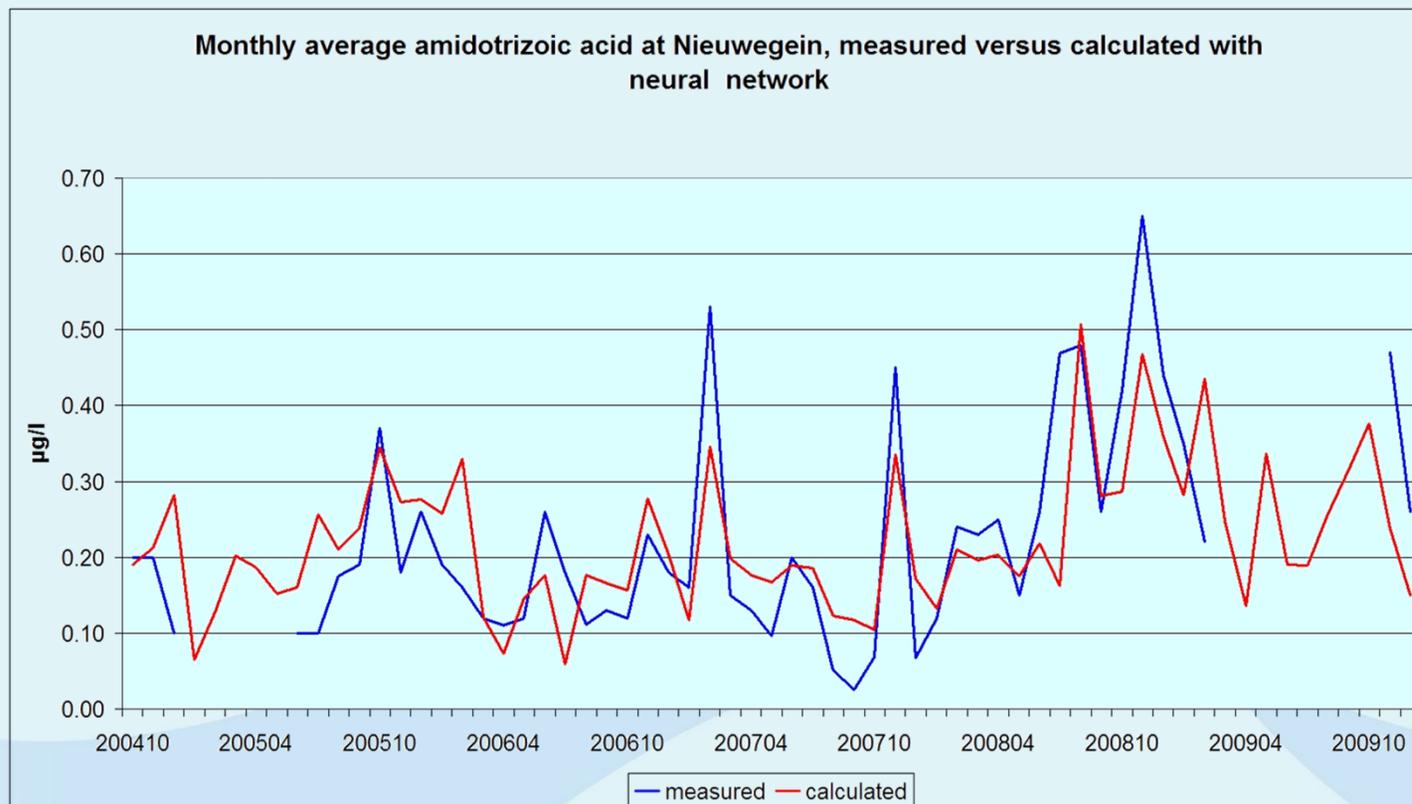


Additional calculations: gap Nieuwegein

Input variables: water discharge at Lobith and Nieuwegein, X-ray agent data at Lobith and these data, one sampling period earlier

→ as before, under low flow conditions these X-ray agent data contain information about Lobith 1 sampling period earlier

Results





Conclusion

- X-Ray media measurements 13 / yr for 5 years at 4 sites would cost ca \$ 35000
- Insufficient data make trend assessments difficult→ Lower costs but virtually useless
- Trend estimation possible by estimating missing data
- Estimated data have reasonable precision but have to be marked in the database as “calculated, not measured”



Current work

- Database contains some 4000 dataserries (variables) with real data $>$ MDL (and some 3500 with mostly “ $<$ “)
- Around 130 dataserries missing 1 quarter year of data & some 200 missing 2 quarter years
- Cluster analysis on those series to find variables with similar behaviour
- Filling gaps in those series