

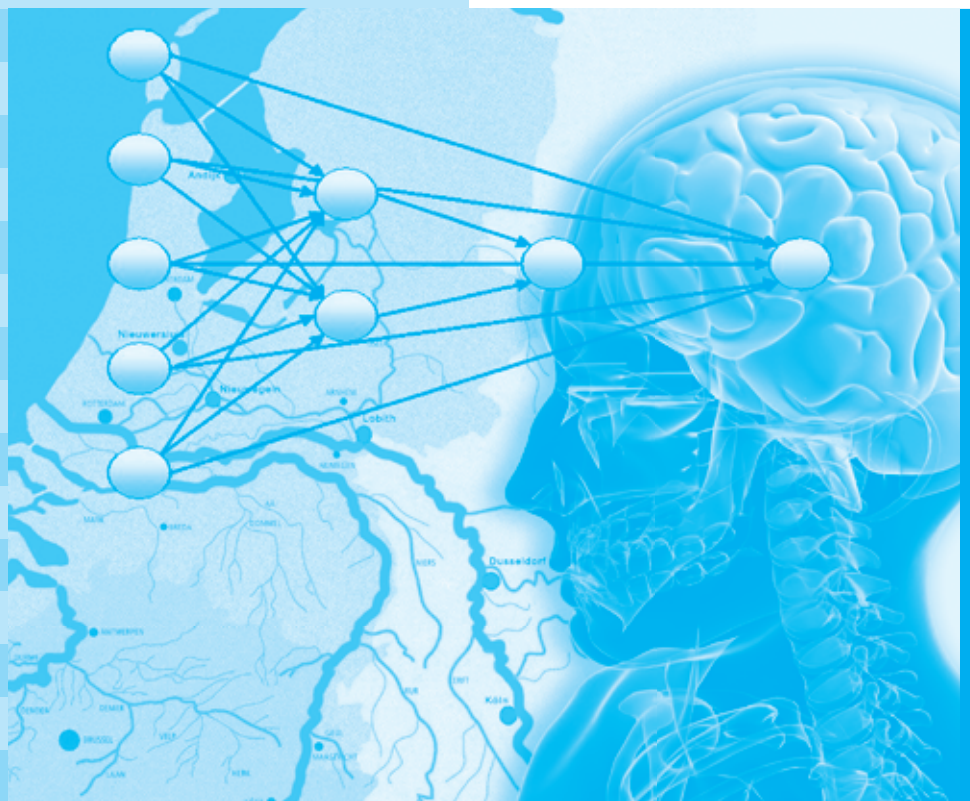
Estimating missing values in time series

RIWA
Rhine Water Works
The Netherlands



A. Smits
Drs. P.K. Baggelaar

Estimating missing values in time series



A. Smits
Drs. P.K. Baggelaar

Table of Contents

1. Introduction	3
2. The problem	5
3. The search for the solution	7
4. The artificial neural network	13
5. Results and conclusions	17
6. Riwa pict	20

Introduction

1

RIWA – the Association of River Water Companies – was founded almost 60 years ago as a cooperative organization by the Dutch water supply companies which use surface water for the preparation of drinking water. Three independent sections, for the rivers Rhine, Meuse and Scheldt, were united within RIWA into an umbrella organisation from 2002 onwards. Each section promotes the interests of drinking water production in its own catchment area: quality, research, reporting, information and action. These activities are determined, financed and implemented for each catchment area.

The section RIWA Rhine collaborates with its German, Swiss and French colleagues in the IAWR, the International Association of Water works in the Rhine area. This umbrella organisation, which was founded in 1970 by RIWA, ARW (Arbeitsgemeinschaft Rhein-Wasserwerke) and AWBR (Arbeitsgemeinschaft Wasserwerke Bodensee-Rhein), covers the entire Rhine river area.

The RIWA strives for a quality level in the surface water of the Rhine catchment area such that simple purification is sufficient to produce flawless drinking water.

The high quality levels, which drinking water must meet in Europe, require a preventive protection of the surface water. As science provides more insight into the hazards which threaten the health of human beings, higher demands are being set for drinking water. Drinking water should be free of unnatural and pathogenic substances. The surface water should be of such a quality level that it is possible to produce drinking water using natural / simple purification methods, such as slow sand filtration, rapid filtration, bank filtration or sedimentation. An important requirement is also that the water is in ecological balance.

The RIWA tries to achieve its objectives with information based on high-level research. One of the most important information sources is the RIWA monitoring network, which collects data from various locations in the Rhine area. The data is stored in a database – the RIWA base - so that information such as various statistics, compliance or non-compliance to water quality standards and trends can be derived.

Water quality variables can fluctuate during the year, due to changes in load, temperature, discharge, breakdown, re-aeration, etc. Therefore, the RIWA monitoring network measures the relevant variables at least every 4 weeks, in order to get a good picture over the whole year.

Apart from the fluctuations during the year there are also long-term fluctuations, mainly caused by “dry” and “wet” years, so that large differences may arise in the discharge between successive years, that influences the concentrations of most water quality variables.

The RIWA analyses for trends over a minimum period of 5 years, in order to gather information about long-term changes in the concentrations of the water quality variables.



The problem

2

RIWA is managing and maintaining a water quality monitoring network to identify (undesired) changes in quality, testing water against the IAWR target values and underpinning goals and requirements. The necessary data is taken from member companies, governments and, to a limited extent, from our own additional research. The monitoring network with the sampling locations in the Netherlands is the downstream part of the IAWR network which starts in Switzerland and continues through Germany where members of AWBR and AWR are collecting data which is exchanged with RIWA. The network consists of a so-called basic program, where a fixed spectrum of variables is analyzed at all sampling sites, and an additional program, where a wider range of variables is analyzed at so-called main sites (such as national borders). This additional program may be varied in width of variables every 3 years, but the sampling frequency is 13/year as a minimum (map 1).

Map 1, the monitoring locations along the river Rhine, mentioned in this article.



Despite the rather strict design criteria, time series of water quality data are regularly interrupted so that it is more difficult to obtain statistically sound statements. This can be due to a multitude of causes, such as changes in analytical methodology, switching between laboratories doing the analysis, miscommunication, or (temporary) financial cuts. The missing values cause difficulties in producing reliable and sound statements about the variables concerned.

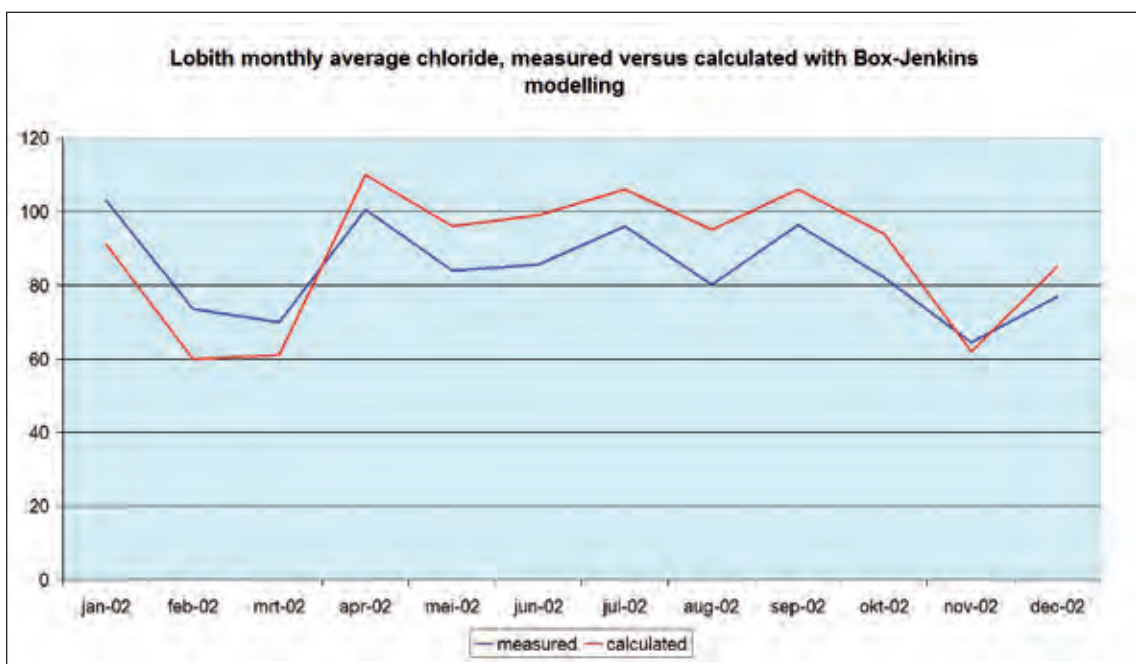
The X-ray contrast agents were included in the RIWA monitoring network in 2004, as the concentration of some of these substances were found to be very frequently above the IAWR quality target of 0.1 µg/l in earlier survey analyses. The time series of these substances were interrupted for various reasons. For Lobith (see map 1) the period from January 2007 to April 2008 (16 months) is missing and for Nieuwegein, Nieuwersluis and Andijk there are no reliable results in the periods January 2005 to May 2005 and March 2009 to October 2009 (a total of 13 months).

The search for the solution

3

An initial survey showed that estimating missing values with time series analysis (Box-Jenkins transfer modelling) might provide good results. This was tested with simulations, using series of monthly averages of chloride. A full year of data was omitted from a series of data from Lobith and then the missing values were estimated using chloride data from the upstream site of Cologne and discharge data from Lobith. This gave promising results, using a total series length of 6 years (Graph 1).

Graph 1



To obtain an objective yardstick for judging the precision of the estimation result, the method of extrapolating the autocorrelation function to a time lag of zero was used [Muskens, 1978*].

This method separates the total standard deviation into a part caused by variations of the process in the river and a part caused by sampling, analysis and processing errors (hereafter this error part is referred to as the SAP deviation). The minimum expected SAP deviation from the monthly average appeared to be 8.5 mg/l Cl. The standard deviation of the difference between the measured and simulated data is 10.5 mg/l, which is, therefore, in the order of magnitude of the expected SAP deviation.

* Doctoral thesis, Radboud University Nijmegen, 1978

A model is not necessarily less precise than measurements

Generally, a measured value of a water quality variable differs from the real value because of errors, caused by sampling, lab analysis and processing (such as the transcription of the value). In some cases there will be information about the probability distributions of these errors. From these distributions and using the law of propagation of errors, we can deduce the confidence interval around a measured value. This interval will contain the real value with a certain predefined confidence (such as 95%). However, the exact realisation of the error of a specific measured value will remain unknown, so as a consequence there always is uncertainty about the underlying real values that we are trying to measure.

The uncertainty about the real value of a water quality variable at a certain point in time can be reduced in various ways. One way is to estimate the real value as the average of a number of measurements taken at that point in time. Another way is to estimate the real value with a model for the time process that generated that real value, such as a Box-Jenkins model, or a neural network. The model can be derived using the time series of the measured values and the time series of measured values of related factors. If the model is a correct representation of the time process and the measurement errors are mostly random, it can even lead to a more precise estimate of a real value than a measurement, because it estimates the average of an infinite number of measured values taken at a given point in time. However, it is not possible to establish if a model is more precise than measurements, as we do not know the real values.

The chloride levels at Nieuwegein and Andijk also appeared reconstructable with this method. However, a good model could not be identified for the Nieuwersluis sampling point, so no useful results were achieved there. The most likely reason for this is the specific location of that sampling site in a fairly stagnant canal.

Based on the positive results with chloride data, a Box-Jenkins-time series model was used for the X-ray contrast agent amidotrizoic acid. However, this did not appear to be successful, partly because the data series from Cologne was not complete. ARW, the sister organisation of the RIWA, has complete data series of X-ray contrast agents at Düsseldorf, but these also appeared insufficient to estimate a sufficiently reliable model, with statistically significant model parameters.

The minimum expected SAP deviation for amidotrizoic acid at Düsseldorf was $0.05 \mu\text{g/l}$, but the Box-Jenkins-model resulted in a standard deviation of the difference between the measured and simulated data of at least $0.16 \mu\text{g/l}$, possibly because of the large fluctuations of the time series.

The following options were then investigated:

1. Using the X-ray contrast agent data from Düsseldorf to estimate the missing values from Lobith. Düsseldorf is only 138 kilometres upstream from Lobith, but between these two locations the Lippe and the Ruhr join the Rhine. The standard deviation of the difference between the measured and simulated data is $0.09 \mu\text{g/l}$.

Map 2, situation of the Rhine estuary



2. Linear interpolation between Düsseldorf and Nieuwegein. This results in a standard deviation of the difference between the measured and simulated data of 0.09 µg/l. With low river flows (<2000 m³/s), the stretch Lobith-Nieuwegein can be traversed in different ways, firstly via the Nederrijn/Lek and secondly via the Waal, the Betuwepand and the Lek. Locally at the Waternet intake point in Nieuwegein, the water can come from the North via the Amsterdam-Rhine Canal or from the South via the Beatrix locks. The flow travelling times between Lobith and Nieuwegein can vary between several days and several weeks along these different stretches. This causes the relationship between the concentrations at Lobith and Nieuwegein to be sometimes vague. With high discharges (>2000 m³/s), the stretch between Nederrijn/Lek/Beatrix locks is more clearly defined. The influence of these different situations on the standard deviation of the difference between the measured and simulated data was examined: with a discharge >2000 m³/s the standard deviation was 0.06 µg/l and with low discharges this was 0.10 µg/l.
3. Using a neural network. The possibilities of using an artificial neural network were examined, considering there might be non-linear relationships between the concentrations at Lobith and Nieuwegein. Neural networks are more capable of describing non-linear relationships than for example time series models or interpolation methods. Input variables were the monthly averaged concentrations of amidotrizoic acid at Düsseldorf and Nieuwegein, the discharge at Lobith and Nieuwegein and the concentration of amidotrizoic acid at Nieuwegein a month later. This latter concentration can contain information about the concentration at Lobith a month earlier at low water flow conditions. The estimated standard deviation of the difference between the measured and simulated data is 0.07 µg/l with this method.

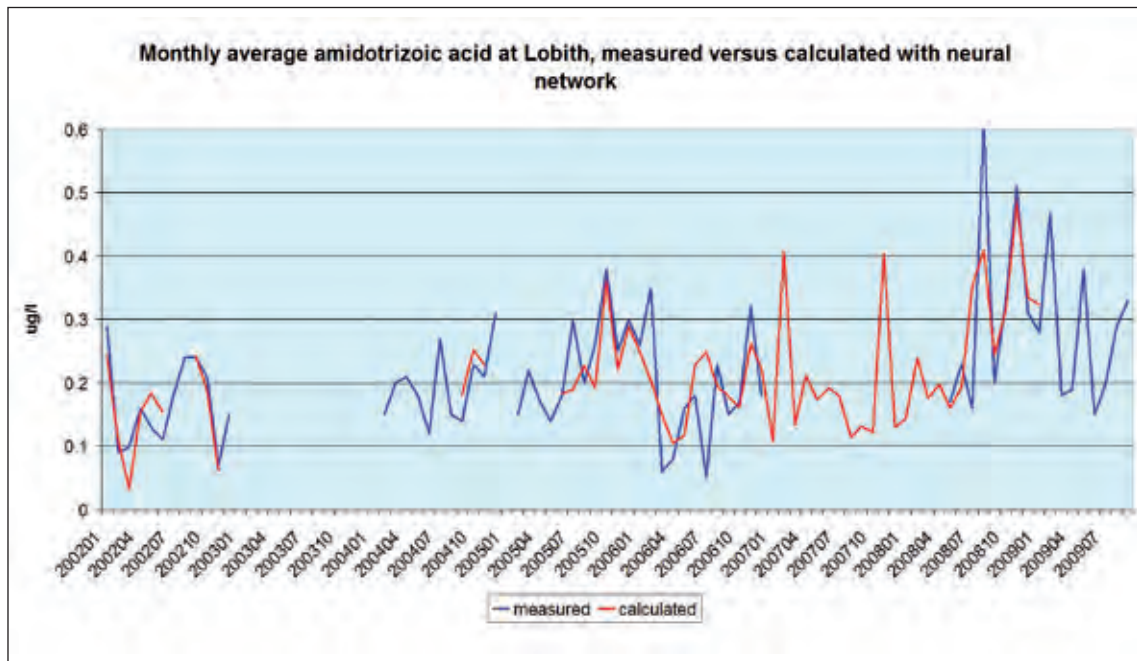
Table 1, summary of the estimation errors of amidotrizoic acid at Lobith in µg/l:

SAP deviation Düsseldorf	Box & Jenkins time series analysis	Düsseldorf instead of Lobith	Interpolation	Interpolation >2000 m ³ /s	Interpolation <2000 m ³ /s	Neural network
0.05	0.16	0.09	0.09	0.06	0.10	0.07

Based on these results, we decided to use the artificial neural network. Its results are acceptable and it also has a built-in flexibility to model both linear and non-linear relationships.

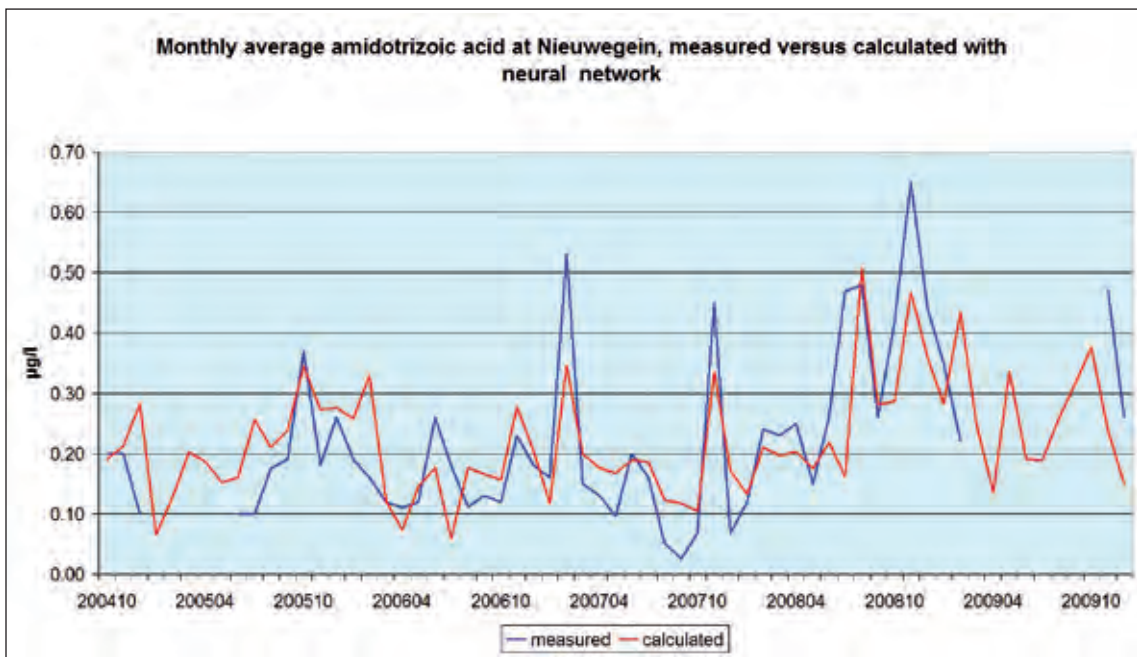
The result of using the artificial neural network for the “gap” at Lobith 2007/2008 for amidotrizoic acid is shown on the graphs below.

Graph 2



Missing values could also be easily constructed for the “gaps” in the series at Nieuwegein. The input of the model was the discharge at Lobith and Nieuwegein, the concentration at Lobith and this concentration one month earlier. On the basis of the extrapolated autocorrelation function, the minimum expected SAP deviation in the measurement error for amidotrizoic acid was approximately $0.10 \mu\text{g/l}$, whilst the standard deviation of the difference between the measurement value and the estimate using the neural network was $0.09 \mu\text{g/l}$.

Graph 3



Source: Rijkswaterstaat

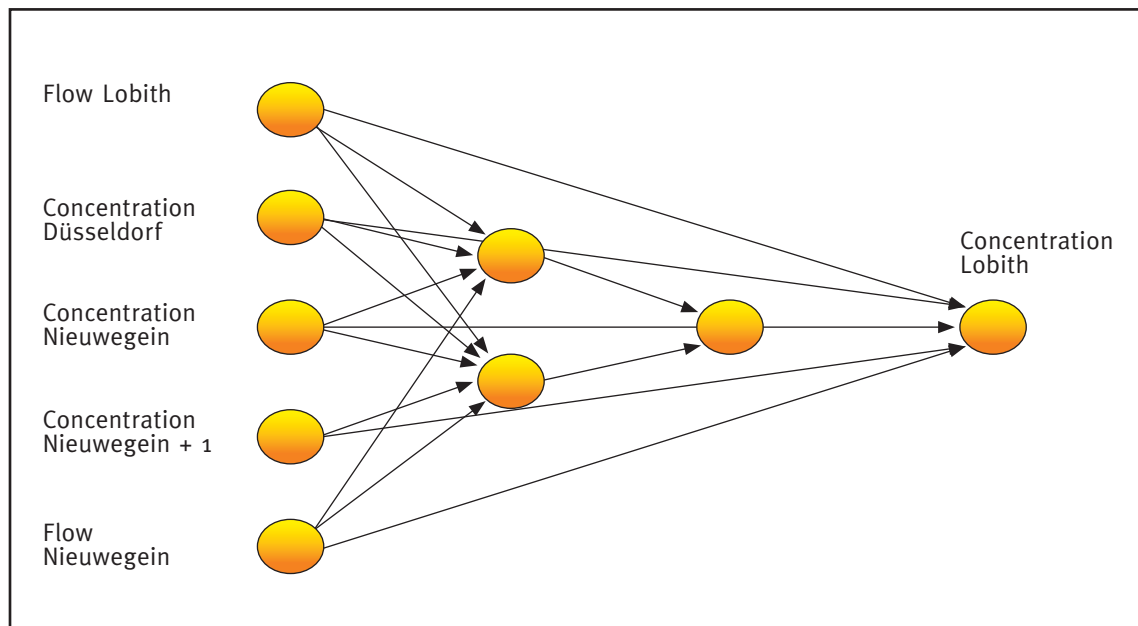


The artificial neural network

4

Artificial neural networks are a simplified copy of the networks in the human nervous system and brain. A biological neuron has several dendrites that receive information from other neurons via connections formed by an axon and synapses. These connections are very numerous. They are able to conduct one-way traffic due to the structure of the synapse. A nerve cell is activated if the sum of the received information pulses exceeds a certain threshold limit and then in its turn passes on this information to the cells to which it has many connections. This makes the signal transfer very complex. Artificial neural networks, although of much simpler construction, have analogously a construction of groups of neurons that are connected together. In the neural network used these neurons are grouped in layers, so there is an input layer and an output layer and one or more (in this case two) intermediate or hidden layers between them, by which the input and output neurons are connected.

Fig 1, representation of the neural network



Mathematically, these connections act as transfer functions, in which the function parameters are optimised during the “learning” of the network on the basis of input and output series.

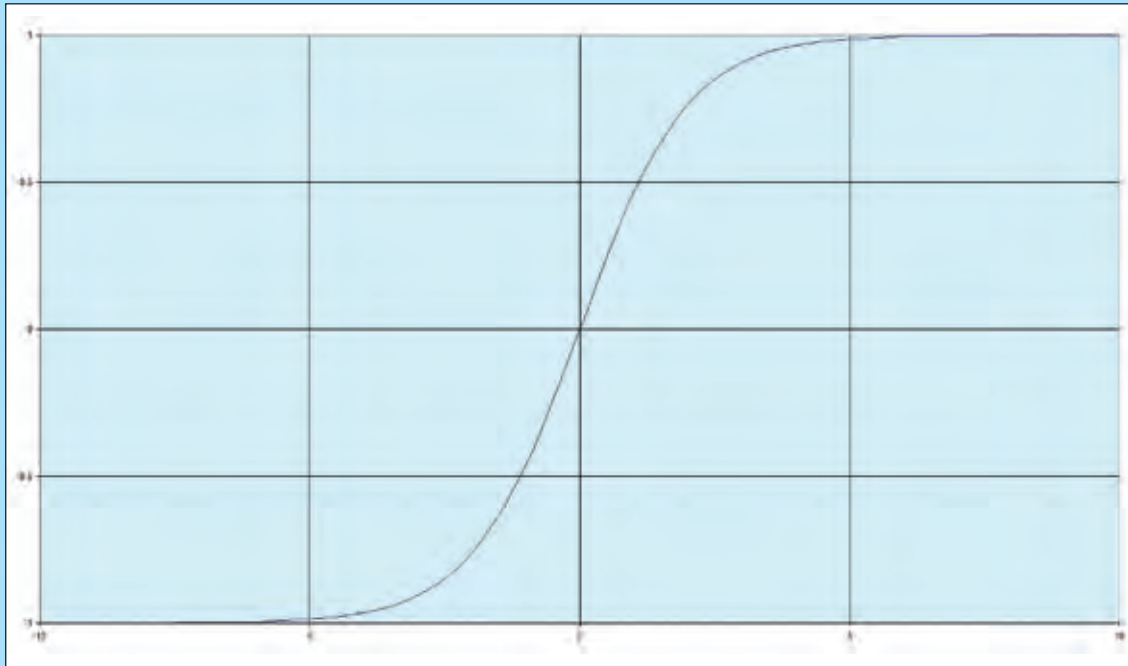
The above network configuration was used to estimate the missing values at Lobith. It is part of the feed-forward networks, as each successive layer receives data via the transfer function of its previous layer, in this case from left to right (one-way traffic).

Logistic function

For the transfer function a logistic function was chosen (fig 2).

Logistic functions are “S” shaped. They slowly rise in the beginning and then rise very sharp, to flatten out in the end.

Fig 2, logistic function



The logistic activation function is defined by:

$$f(x) = \frac{p2 - p3}{1 + e^{-p1x}} + p3$$

p1 determines the slope at x=0

p2 is the upper limit

p3 is the lower limit

Logistic functions are often used as a transfer function in neural networks to enable the description of nonlinear behavior by the model.

The “learning” of this network occurs by offering a training set of data to the input side and offering a known target value to the output side. This makes the network used into a supervised back propagation network (the deviation of the target value from the output proceeds in layers back into the network and the parameters of the transfer functions are adapted along the way so as to optimize the match between input and output). In this case the input and output neurons are also directly connected. These connections ensure that the linear dependencies are described well; the non-linear components are calculated by the intermediate layers.

During the learning, a part of the learning set was used to test if the results of the trained network agreed with the target values. This appeared to be the case. If this had not been so, then the network would be over dimensioned and it would have to be simplified.

Source: Het Waterlaboratorium



Results and conclusions

Missing values in the data series of X-ray contrast agents measured at Lobith, Nieuwegein and Andijk can be estimated with reasonable precision, using a neural network.

Where possible, missing values were, therefore, estimated and included in the RIWA base.

Once completing these data series of X-ray contrast agents, the standard RIWA trend analysis could be applied. The example below (fig 3) shows in RIWApicts the results of the trend analysis and the compliance with standards for 2009.

Fig 3, RIWAPICT X-ray contrast agents

	cas number		Lobith	Nieuwegein	Andijk
Amidotrizoic acid	117-96-4	µg/l			
Iodipamide	606-17-7	µg/l			
Iohexol	66108-95-0	µg/l			
Iomeprol	78649-41-9	µg/l			
Iopamidol	62883-00-5	µg/l			
Iopanoic acid	96-83-3	µg/l			
Iopromide	73334-07-3	µg/l			
Iotalamic acid	2276-90-6	µg/l			
Ioxaglic acid	59017-64-0	µg/l			
Ioxitalamic acid	28179-44-4	µg/l			

Excerpt from the annual report 2009

	dimension	dl	Jan	Feb	Mar	Apr	May	Jun
RIWApict								
Lobith								
Diatrizoic Acid	µg/l		0.28	0.47	0.18	0.19	0.38	0.15
Iodipamide	µg/l	0.01	<	<	<	<	<	<
Iohexol	µg/l		0.15	0.43	0.2	0.15	0.2	0.073
Iomeprol	µg/l		0.67	1.3	0.52	0.48	0.53	0.26
Iopamidol	µg/l		0.31	0.53	0.29	0.24	0.32	0.19
Iopromide	µg/l		0.26	0.45	0.21	0.18	0.24	0.19
Iothalamic acid	µg/l	0.01	<	<	<	<	<	<
Ioxaglic acid	µg/l	0.01	<	<	<	<	<	<
Ioxitalamic acid	µg/l		0.042	0.068	0.04	0.043	0.045	0.033
Nieuwegein								
Diatrizoic Acid	µg/l		0.35	0.22	0.25	0.14	0.34	0.19
Iodipamide	µg/l	0.01	<	<	<	<	<	<
Iohexol	µg/l		0.11	0.14	0.14	0.02	0.08	0.04
Iomeprol	µg/l		0.32	0.52	0.853	0.331	0.422	0.366
Iopamidol	µg/l		0.18	0.3	0.248	0.142	0.258	0.17
Iopanoic acid	µg/l	0.01	<	<	<	<	<	<
Iopromide	µg/l		0.16	0.42	0.298	0.0121	0.213	0.176
Iothalamic acid	µg/l	0.01	<	<	<	<	<	<
Ioxaglic acid	µg/l	0.01	<	<	<	<	<	<
Ioxitalamic acid	µg/l	0.01	0.35	0.054	0.0253	<	0.0281	0.0241
Andijk								
Diatrizoic Acid	µg/l		0.32	0.26	0.25	0.14	0.34	0.19
Iodipamide	µg/l	0.01	0.01	<	<	<	<	<
Iohexol	µg/l		0.13	0.13	0.14	0.02	0.08	0.04
Iomeprol	µg/l		0.4	0.5	0.425	0.242	0.23	0.182
Iopamidol	µg/l		0.27	0.28	0.135	0.142	0.132	0.11
Iopanoic acid	µg/l	0.01	<	<	<	<	<	<
Iopromide	µg/l		0.17	0.2	0.138	0.0838	0.0827	0.08
Iothalamic acid	µg/l	0.01	<	<	<	<	<	<
Ioxaglic acid	µg/l	0.01	<	<	<	<	<	<
Ioxitalamic acid	µg/l	0.01	0.036	0.041	0.0132	<	0.0137	0.0109
Nieuwersluis								
Diatrizoic Acid	µg/l	0.01	0.25	0.25	0.065	0.04	0.04	0.0115
Iodipamide	µg/l	0.01	<	<				
Iohexol	µg/l	0.01	0.11	0.13	0.022	<	<	<
Iomeprol	µg/l	0.01	0.53	0.75	0.096	<	<	0.0125
Iopamidol	µg/l	0.01	0.19	0.24	0.016	<	<	0.0165
Iopanoic acid	µg/l	0.01			<	<	<	<
Iopromide	µg/l	0.01	0.45	0.49	0.111	<	<	<
Iothalamic acid	µg/l	0.01	<	<	<	<		<
Ioxaglic acid	µg/l	0.1	<	<		<	<	<
Ioxitalamic acid	µg/l	0.01	0.081	0.1	0.014	<	<	<
Unreliable results								

Jul	Aug	Sep	Oct	Nov	Dec	n	mmin	mP10	mP50	mgem	mP90	mmax	
0.2	0.29	0.33	0.39	0.21	0.13	13	0.13	0.138	0.21	0.262	0.438	0.47	
<	<	<	<	<	<	13	<	<	<	<	<	<	
0.0755	0.054	0.088	0.088	0.16	0.11	13	0.054	0.0616	0.11	0.143	0.338	0.43	
0.35	0.29	0.38	0.39	0.58	0.37	13	0.26	0.272	0.39	0.498	1.05	1.3	
0.305	0.33	0.41	0.38	0.44	0.23	13	0.19	0.206	0.31	0.329	0.494	0.53	
0.39	0.46	0.13	0.13	0.17	0.15	13	0.13	0.13	0.21	0.258	0.456	0.46	
<	<	<	<	<	<	13	<	<	<	<	<	<	
<	<	<	<	<	<	13	<	<	<	<	<	<	
0.036	0.031	0.036	0.056	0.049	0.034	13	0.031	0.031	0.041	0.0422	0.0632	0.068	
0.19	0.26	0.32	0.38	0.47	0.26	12	0.14	0.155	0.26	0.281	0.443	0.47	
<	<	<	<	<	<	12	<	<	<	<	<	<	
0.04	0.01	0.02	0.03	0.063	0.11	12	0.01	0.013	0.0515	0.0669	0.14	0.14	
0.268	0.27	0.281	0.337	0.29	0.43	12	0.268	0.269	0.334	0.391	0.753	0.853	
0.26	0.25	0.308	0.292	0.25	0.32	12	0.142	0.15	0.254	0.248	0.316	0.32	
<	<	<	<	<	<	12	<	<	<	<	<	<	
0.34	0.359	0.186	0.114	0.15	0.18	12	0.0121	0.0425	0.183	0.217	0.402	0.42	
<	<	<	<	<	<	12	<	<	<	<	<	<	
<	<	<	<	<	<	12	<	<	<	<	<	<	
0.0301	0.0221	0.0277	0.0322	0.027	0.056	12	<	0.0101	0.0279	0.0568	0.262	0.35	
0.19	0.26	0.32		0.073	0.13	11	0.073	0.0844	0.25	0.225	0.336	0.34	
<	<	<	<			10	<	<	<	<	<	0.01	
0.04	0.01	0.02	0.03	0.03	0.04	12	0.01	0.013	0.04	0.0592	0.137	0.14	
0.13	0.146	0.152	0.176	0.098	0.21	12	0.098	0.108	0.196	0.241	0.478	0.5	
0.125	0.138	0.148	0.142	0.088	0.15	12	0.088	0.0947	0.14	0.155	0.277	0.28	
<	<	<	<	<	<	12	<	<	<	<	<	<	
0.118	0.159	0.091	0.13	0.073	0.086	12	0.073	0.0751	0.104	0.118	0.191	0.2	
<	<	<	<	<	<	12	<	<	<	<	<	<	
<	<	<	<	<	<	12	<	<	<	<	<	<	
0.0154	<	0.0125	0.0152	<	0.029	12	<	<	0.0135	0.0172	0.0395	0.041	
0.04	<	0.02	<	0.62	0.19	13	<	<	0.04	0.119	0.472	0.62	
						2	*	*	*	*	*	*	
<	<	<	<	0.06	0.075	13	<	<	<	0.0336	0.122	0.13	
<	<	<	<	0.37	0.38	13	<	<	<	0.168	0.662	0.75	
<	<	<	0.013	0.25	0.14	13	<	<	0.013	0.0698	0.246	0.25	
<	<	<	<	<	<	11	<	<	<	<	<	<	
0.01	0.012	<	<	0.44	0.23	13	<	<	0.01	0.136	0.474	0.49	
<	<	<	<	<	<	12	<	<	<	<	<	<	
<	<	<	<	<	<	12	<	<	<	<	<	<	
<	<	<	<	0.16	0.091	13	<	<	<	0.0374	0.136	0.16	
Unreliable results													

■ det. lim. = detection limit

■ n = number of results per year

■ min = minimum

■ p10 p50 p90 = percentile values

■ gem = average

■ max = maximum

■ RIWA pict = for explanation see last page of this report


■ ! = series completely or partially calculated with the aid of the neural network


Visualisation of measurement results


RIWA developed so-called pictograms, as a means to visualise water quality measurement results. These pictograms show both trends and compliance (or non-compliance) with threshold values. In addition, from the pictograms it can be inferred whether or not there are sufficient data points for a statistically reliable statement.

The color indicates the concentration level in relation to the threshold value:


0 – 79 % of the value is blue 


80 – 99 % of the value is yellow 

100 and greater is red 


No color, only a symbol, means: no threshold value 


The symbol indicates the direction of a trend:


A horizontal line means that no trend can be demonstrated 

An arrow shows the direction of a (significant) trend (95% 2-sided conf.) 

The filling of the pict indicates how many data points were used for the statement:

0 – 19 data points: a colored symbol and a white square 

20 or more data points: a white symbol and a colored square 

An empty square means there are no, or not enough data points for a reliable statement. 

Routinely, the monitoring frequency is 13 as a minimum. For trend detection, the annual values over a five-year period are used to calculate quarterly averages which are then used to calculate the trend.

The costs over 5 years of measuring X-ray contrast agents in the Dutch section of the monitoring network are about 27,000 euros. Only after the estimation of missing values the standard RIWA analysis tools could be used (trend analyses and establishing compliance with standards). Estimating missing values, therefore, prevents a considerable loss of capital and information.

The estimated values can not just simply be mixed with the real data in the database. Therefore, they were marked. And they were also combined with the original series to a new, but also marked series. This resulted in three series for each X-ray contrast agent: the reconstructions (1st half of 2008), the original (2nd half of 2008) and the composite. In the annual reports and ad hoc print-out of this data this latter series, which does not contain a gap, is used. From the marks it can be inferred which statistical designations, such as minimum and maximum, average, percentages, compliance with standards and trends, are partly or wholly based on reconstructed data for the year concerned.

RIWA is currently investigating the further practical use of neural networks to establish procedures which will hopefully lead to an even more effective use of available data.

Colophon

Colophon

Authors:	A. Smits Drs. P.K. Baggelaar
Contribution by:	Dr. P.G.M. Stoks Ing. G. van de Haar
Publisher:	RIWA-Rhine, The Netherlands
Design:	Meyson Communicatie, Amsterdam
Print:	ATP Digitale Media, Hoofddorp
ISBN/EAN:	ISBN 9789066831421

Groenendaal 6
NL-3439 LV Nieuwegein
The Netherlands
T +31 (0)30 600 90 30
F +31 (0)30 600 90 39
E riwa@riwa.org
W www.riwa.org