5

Data Reduction for Water Quality Modelling, Vaal Basin

Bloodless Dzwairo[1], George M. Ochieng'[1], Maupi E. Letsoalo[1] and Fredrick A.O. Otieno[2]
[1]Tshwane University of Technology
[2]Durban University of Technology
South Africa

1. Introduction

Constructing models, comparing their predictions with observations, and trying to improve them, constitutes the core of the scientific approach to understanding complex systems like large river basins (Even et al., 2007). These processes require manipulation of huge historical data sets, which might be available in different formats and from various stakeholders. The challenge is then to first pre-process the data to similar lengths, with minimal loss of integrity, before manipulating it as per initial objectives. In the Upper and Middle Vaal Water Management Areas (WMAs) of the Vaal River, bounded by Vaal dam outlet and Bloemhof dam inlet, the overall objective of on-going research is to model surface raw water quality variability in order to predict cost of treatment to potable water standard. This paper reports on part of the overall research. Its objective was to show how a huge and non-consistent water quality data set could be downsized to manageable aspects with minimal loss of integrity. Within that scope, challenges were also highlighted.

One of the more important forms of knowledge extraction is the identification of the more relevant inputs. When identified, they may be treated as a reduced input for further manipulation. In water quality data analysis, data collection, cleaning and pre-processing are often the most time-consuming phases. All inputs and targets have to be transferred directly from instrumentation or from other media, tagged and arranged in a matrix of vectors with the same lengths (Alfassi et al., 2005). If vectors have outliers and/or missing values these have to be identified for correction or to be discarded. More complex mathematical correlations are sometimes employed to identify redundant, co-linear inputs, or inputs with little information content (Alfassi et al., 2005).

Sources and sinks of variables in hydrodynamics, also known as forcing functions, are the cause of change in water quality (Martin et al., 1998). To capture intermediate scale processes that are spotty in spatial extent, extensive sampling and averaging of the calibration data over sufficient spatial scales is done to capture that condition over time. Although many water constituents are non-conservative in nature, a few conservative ones that approach ideal behaviour under limited conditions, could be used for modelling and calibration.

The study area is a major focus of modelling and pollution tracing in the Vaal basin, South Africa, (Dzwairo et al., 2010b, Cloot and Roux, 1997, DWAF, 2007, Gouws and Coetzee, 1997, Naicker et al., 2003, Pieterse et al., 1987, Stevn and Toerien, 1976, Dzwairo et al., 2010a, Dzwairo and Otieno, 2010, Herold et al., 2006).
Data sets spanning many years have been collected by various stakeholders including the Department of Water Affairs (DWA) and Water Boards which treat bulk water for potable use. For management of the basin as a whole these data sets come handy but the major challenge is collating them into uniform and useable data, while noting that the different stakeholders monitor selected parts of the basin for their own specific purposes. Some sampling points might be dropped off or new points picked up as emerging pollution threats require tracing and monitoring in order to mitigate effects. Still a useable data set has to be constructed to monitor pollution and other threats, in addition to informing and alerting decision makers regarding environmental and human health issues. This paper shows how inconsistent and scattered data sets from 13 monitoring points were pre-treated and downsized to SO_4^{2-} inter-relationships. SO_4^{2-} is a very important parameter in surface water quality variability in this region because of the existence of gold and coal mining activities. Threats from acid mine drainage are real.

2. Study area

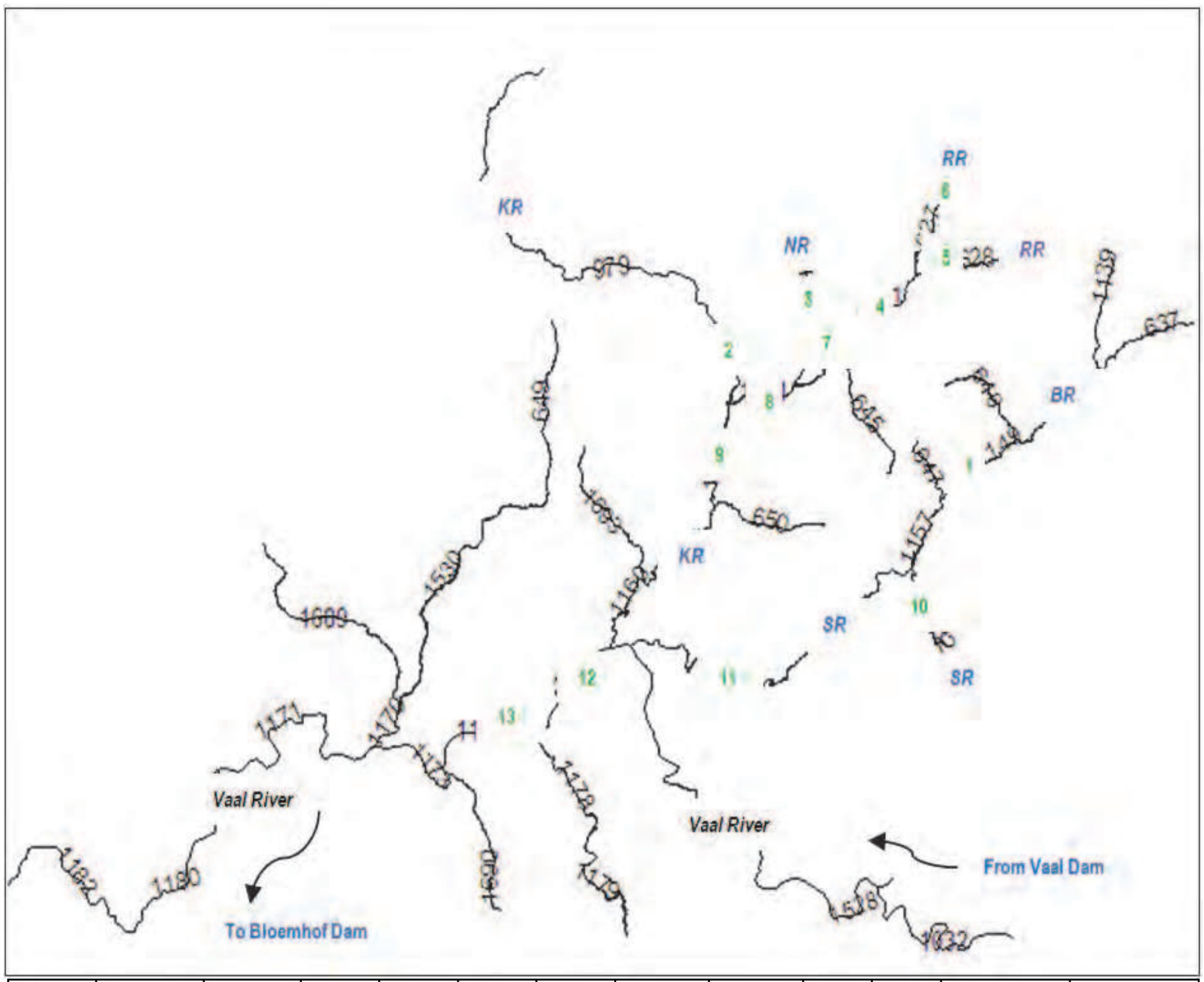
The study area as indicated in Fig. 1 shows spatial relationships of the sampling points located on VR and its tributaries as follows: B1-B10 on Blesbokspruit River (BR); K10-K10, K6-K25 and K9-K19 on Klip River (KR); K12-N8 on Natalspruit River (NR); K1-R2 on Withokspruit River, which is a tributary of Rietspruit River (RR); K3-R3 on another tributary of RR; K2-R1 and K4-R4 on RR; S1-S1 and S4-S2 on Suikerbosrant River (SR); and V7-VRB37 and V9-VRB24 on Vaal River (VR).

3. Methods and materials

Water quality data from 13 surface raw water quality monitoring points covering the period 1 January 2003 to 30 November 2009 was manipulated to remove limits of detection as well as gaps in sampling periods. An example of raw data is presented in Table 1 for sampling points Y and Z and for only Chl-α, COD, EC and DOC. The extracted data sample covered 5 July 2004 to 26 July 2004.
Using the list of variables in Table 2, comparisons among points entailed obtaining or converting the raw data to match sampling periods among the points. Although there are several interpolation techniques, cubic interpolation was chosen for the time-series data set because the method is shape-preserving. Interpolation created date-interpolated daily data using Matlab R2009b.

3.1 Manipulating data falling below or above detectable limits

Data that was above limit (e.g. $500 < x$) was assumed to be one magnitude higher than the given value, whereas that which was reported as below detectable limit (e.g. $x < 1.1$) was multiplied by 0.75 to give absolute values that could be manipulated as normal data (Ochse, 2007).



B1-B10	K10-K10	K12-N8	K1-R2	K2-R1	K3-R3	K4-R4	K6-K25	K9-K19	S1-S1	S4-S2	V7-VRB37	V9-VRB24
1	2	3	4	5	6	7	8	9	10	11	12	13

Fig. 1. Monitoring points in study area bounded, by the two dams.

Date	Chl-α	COD	EC	DOC	Chl-α	COD	EC	DOC
Sampling point	Y				Z			
5-Jul-04					17.00	19.00	105.00	4.90
7-Jul-04	8.10	20.00	80.00	8.30				
12-Jul-04					5.60	19.00	99.00	6.10
19-Jul-04					8.30	21.00	96.00	
21-Jul-04	74.00	27.00	88.00	8.70				
26-Jul-04					6.90	24.00	97.00	5.50

Table 1. Raw data for monitoring points Y and Z.

Parameter	Unit	Description	Abbreviation
so42_	mg/L	sulphate	SO_4^{2-}
cn_	mg/L	cyanide	CN^-
ec	mS/m	conductivity	EC
do	mg/L	dissolved oxygen	DO
fc	CFU/100mL	faecal coliforms	Fc
Hg	µg/L	mercury	Hg
Cl_	mg/L	chloride	Cl^-
f_	mg/L	fluoride	F^-
no2_	mg/L	nitrite	NO_2^-
no3_	mg/L	nitrate	NO_3^-
Low_Hg	µg/L	low mercury	Hg
Mn	mg/L	manganese	Mn
pH	-	-	-
po43_	mg/L	phosphate	PO_4^{3-}
s	mg/L	sulphur	S
ss	mg/L	suspended solids	SS
Temp	°C	temperature	-
T_Silica	mg/L	total silica	-
Turb	NTU	turbidity	-
nh4_	mg/L	ammonium	NH_4^+
Chla	µg/L	chlorophyll -α	Chl-α
cod	mg/L	chemical oxygen demand	COD
doc	mg/L	dissolved organic carbon	DOC
Mo	mg/L	molybdenum	Mo
Si	mg/L	silicone	Si
p	mg/L	phosphorus	P
Fe	mg/L	iron	Fe

Table 2. Parameters under consideration.

3.2 Matlab codes for cubic interpolation

3.2.1 Cubic interpolation

Data interpolation is an application based on underlying geometric algorithms. Data may be uniform, that is, sampling occurs over uniform intervals or it may be scattered, that is, sampling occurs over irregular intervals. When the sample data is scattered, the interpolation techniques use a triangulation-based approach as a basis for computing interpolated values. Table 3 provides a Matlab code for date-interpolating a single column. To interpolate many columns, the single-column code was adjusted as in Table 4.

3.2.2 Challenges during interpolation

An empty cell at any position of the matrix, for example a missing date or value, returned an error similar to the one in Table 5.

```
% Load the data with lots of missing dates. Note that in this example
% missing dates are not represented by NaN but are left out completely

>>[data,textdata] = xlsread('book.xls');

% Convert the text date to date numbers (you may have to change the date
% format depending on how your dates appear in Excel)

>>dates = datenum(textdata,'mm/dd/yyyy');

% Plot the data

>>plot(dates,data,'LineStyle','none','Marker','o')

% Show the x axis as a date

>>datetick('x')

% Create a new date series starting at the first date in dates and
% ending at the last but with every date in-between

>>newDates = dates(1):dates(end);

% Interpolate to find the missing data

>>newData = interp1(dates,data,newDates,'cubic');

% Convert the date numbers to strings and then to cell arrays

>>stringDates = cellstr(datestr(newDates));

% Combine the dates and the data

>>outputData = [stringDates, num2cell(newData')];

% Write the data to Excel
>>xlswrite('outbook.xls',outputData);
```

Table 3. Coding for interpolating a single column.

```
>>newDates = dates(1):dates(end);

%Run the tic toc (3 instructions below at once by copying and pasting, it should
give elapsed time as eg 0.305720 seconds)

>>tic
newColumnData = interp1(dates,columnData,newDates,'cubic');
toc

Elapsed time is 0.305720 seconds.

%In a new figure, plot both the new data and the existing data

figure

>>plot(newDates,newColumnData,dates,columnData,'LineStyle','none','Marker','o')

%Change date format to years

>>datetick('x')

%Convert the date numbers to strings and then to cell arrays

>> stringDates = cellstr(datestr(newDates));

%Combine the dates and the data

>>outputData = [stringDates, num2cell(newColumnData)];

Write the data back to Excel
```

Table 4. Code for interpolating many columns.

```
>tic
newColumnData = interp1(dates,columnData,newDates,'cubic');
toc

Warning: NaN found in Y, interpolation at undefined values will result in undefined values.
In interp1 at 178

Warning: All data points with NaN in their value will be ignored.
In polyfun\private\chckxy at 103
In pchip at 59
In interp1 at 283

Elapsed time is 0.042557 seconds.
```

Table 5. NaN.

Another common error was that of a misplaced decimal point or full stop during data capture (Table 6). Matlab would not be able to manipulate this entry for interpolation because it was not a value. A duplicated or non-formatted date would also present an error that would require debugging before a complete interpolated data set could be obtained. These, among other similar errors, required manual debugging through a whole data set, each a 2526 x28 matrix. With a perfect matrix, an interpolation took a fraction of a second.

Measured parameter	Measured parameter
72.00	0.29
3.75.0	0.31
70.00	0.29

Table 6. A highlighted error arising from data capture.

The 13 sampling points' data was interpolated to the same lengths from 1 January 2003 to 30 November 2009, for the 27 parameters, and then combined into one file for processing using Stata, in order to reduce the matrix. Analysis used case-wise correlation, factor analysis, multivariate linear regression and one-way ANOVA.

4. Results

Initial inspection indicated that the data exhibited gross temporal inconsistency. Sampling dates did not match, in addition to missing values. Table 7 shows the interpolated data for points Z and Y for 5 to 21 July 2004.

Date	Chl-α	COD	EC	DOC	*Chl*-α	*COD*	*EC*	*DOC*
Sampling point		Y				Z		
5-Jul-04					*17.00*	*19.00*	*105.00*	*4.90*
6-Jul-04					*16.26*	*19.00*	*104.74*	*4.97*
7-Jul-04	8.10	20.00	80.00	8.30	*14.58*	*19.00*	*104.04*	*5.14*
8-Jul-04	8.80	20.13	80.12	8.32	*12.36*	*19.00*	*103.06*	*5.37*
9-Jul-04	10.80	20.35	80.44	8.35	*9.97*	*19.00*	*101.92*	*5.63*
10-Jul-04	13.93	20.66	80.94	8.37	*7.80*	*19.00*	*100.77*	*5.86*
11-Jul-04	18.01	21.04	81.59	8.39	*6.21*	*19.00*	*99.75*	*6.03*
12-Jul-04	22.87	21.50	82.33	8.41	*5.60*	*19.00*	*99.00*	*6.10*
13-Jul-04	28.35	22.01	83.15	8.44	*5.75*	*19.07*	*98.41*	*6.09*
14-Jul-04	34.28	22.56	84.00	8.46	*6.14*	*19.26*	*97.82*	*6.06*
15-Jul-04	40.48	23.16	84.85	8.48	*6.66*	*19.54*	*97.26*	*6.01*
16-Jul-04	46.79	23.78	85.67	8.51	*7.24*	*19.88*	*96.76*	*5.96*
17-Jul-04	53.04	24.43	86.41	8.54	*7.76*	*20.25*	*96.36*	*5.90*
18-Jul-04	59.05	25.08	87.06	8.58	*8.15*	*20.64*	*96.10*	*5.85*
19-Jul-04	64.66	25.73	87.56	8.61	*8.30*	*21.00*	*96.00*	*5.80*
20-Jul-04	69.70	26.38	87.88	8.65	*8.22*	*21.39*	*96.03*	*5.75*
21-Jul-04	74.00	27.00	88.00	8.70	*8.02*	*21.86*	*96.12*	*5.70*

Table 7. Date-interpolated data for monitoring point Y and Z.

A full length raw data set for Z (2003 to 2009), shown in Fig. 2, was interpolated and graphed in Fig. 3, for only 4 out of the 27 variables, that is, Chl-α, COD, EC and DOC, to reduce congestion and enhance clarity to the cubic interpolation concept.

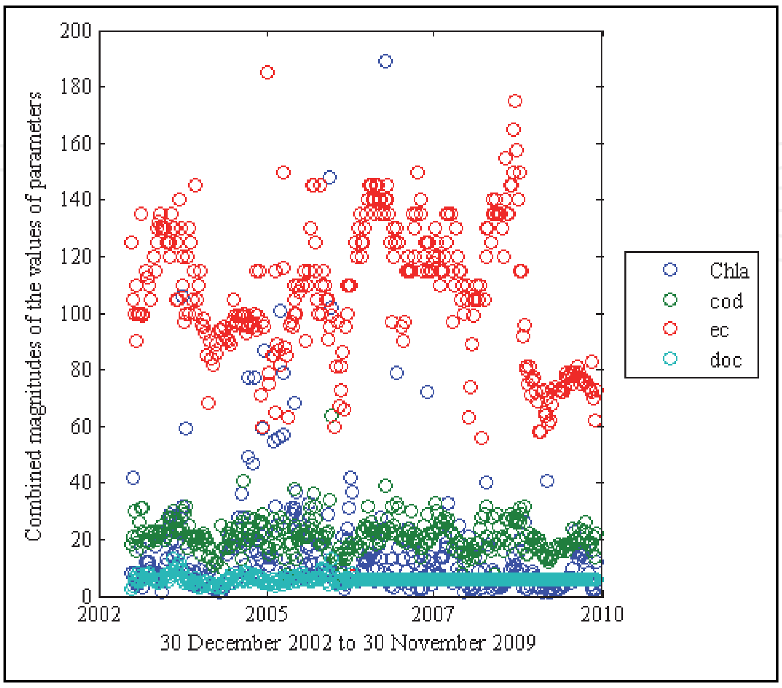


Fig. 2. Monitoring point (Z)'s raw input data.

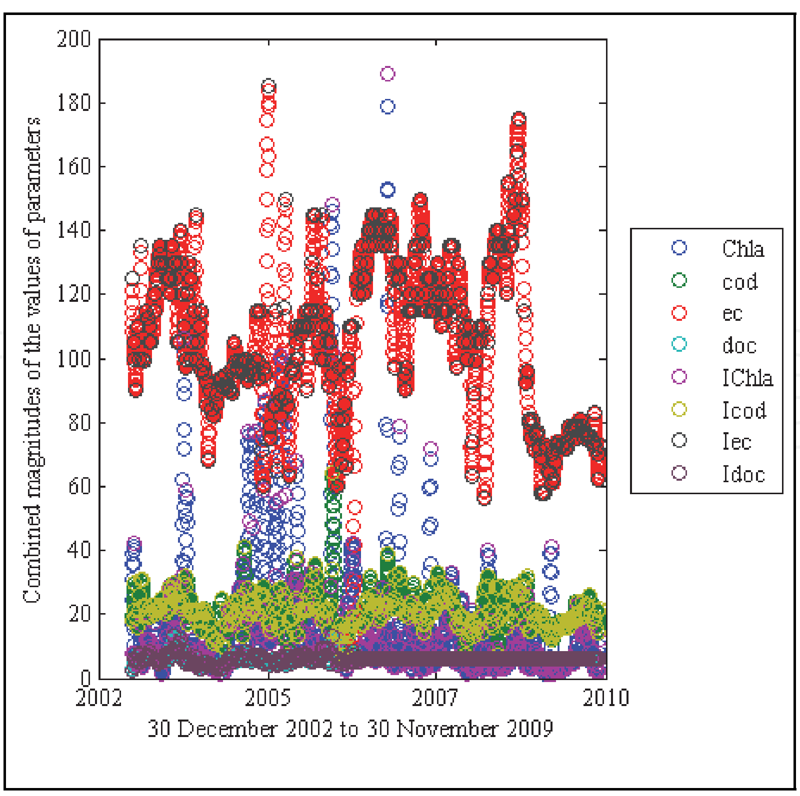


Fig. 3. Monitoring point (Z)'s cubic-interpolated data.

Whereas Fig. 2 showed a legend with 4 data sets, Fig. 3's legend included the interpolated data, colour-coded for clarity. IChla, Icod, Iec and Idoc (IChl-α, ICOD, IEC and IDOC) represented the interpolations of the 4 variables used. Daily interpolation was chosen for this study because after interpolation, any other data interval, for example monthly or yearly variation, could be computed without repeating the time-consuming interpolation process.

4.1 Case-wise correlation analysis

Although case-wise correlation analysis indicated that SO_4^{2-} had a significant linear relationship with all variables except DO, it was strongly positively correlated with EC (0.8720), Cl^- (0.7273), S (0.9053) and Mn (0.4779). It was strongly negatively correlated with pH (-0.5380). Table 8 provides detailed output.

4.2 Factor analysis

The major aim of factor analysis is to orderly simplify a large number of interrelated measures to a few representative constructs or factors (Ho, 2006). The 27 variables were subjected to this technique for that reason, to reduce the data set. The data was collapsed into 3 latent constructs (Table 9 and Table 10).

Their Eigen values were noted to be 5.82041, 2.62148 and 2.12070. Factors 1 and 3 were cross-loaded thus Table 11 was constructed because DOC appeared to be conceptually relevant to Factor 3 (physical parameters) while cod remained relevant to Factor 1 (conductivity related). Factor 2 incorporated unique variables which were not cross-loaded into any of the other factors but for which no good common description could readily be assigned. Variables which could not be placed into any of the 3 factors were also deleted from Table 11, effectively reducing the variables, (see Ho, 2006).

	cn_	ec	do	fc	Hg	Cl_	f_
cn_	1.0000						
ec	0.0908*	1.0000					
do	-0.0106	0.0112*	1.0000				
fc	0.0014	0.0217*	0.0141*	1.0000			
Hg	-0.0523*	-0.1087*	0.0110	-0.0594*	1.0000		
Cl_	0.0783*	0.8699*	0.0039	0.0062	-0.0192*	1.0000	
f_	-0.0053	0.1819*	-0.0404*	0.0239*	-0.1666*	0.0259*	1.0000
no2_	-0.0708*	-0.1365*	0.1629*	0.0809*	0.1839*	-0.0458*	-0.0787*
no3_	-0.0628*	0.1223*	0.1033*	0.0658*	0.1916*	0.0876*	0.0115*
so42_	0.0961*	0.8720*	-0.0064	0.0288*	-0.2013*	0.7273*	0.2798*
Low_Hg	-0.0009	0.2998*	0.0450*	-0.0260*	-0.2516*	0.1762*	0.3496*
Mn	0.0147*	0.3936*	-0.0102	0.0668*	-0.1783*	0.1815*	0.2316*
pH	0.0290*	-0.4242*	0.0481*	-0.0856*	0.1456*	-0.1382*	-0.3480*
po43_	-0.0367*	-0.0858*	0.0283*	0.0418*	0.1250*	-0.0193*	-0.0683*
s	0.0807*	0.8861*	-0.0176*	0.0226*	-0.1974*	0.7435*	0.2593*
ss	-0.0302*	-0.2024*	-0.0336*	0.0138	0.0350*	-0.1852*	-0.0387*
Temp	-0.0120*	-0.0369*	-0.0424*	0.0201*	-0.0948*	-0.0544*	0.0481*
T_Silica	-0.0343*	0.1377*	-0.0693*	0.0422*	-0.1797*	-0.0889*	0.2674*
Turb	-0.0434*	-0.2525*	-0.0862*	0.0284*	-0.0893*	-0.2899*	0.0213*
nh4_	0.0267*	0.3493*	-0.0444*	0.2118*	-0.0952*	0.2378*	0.1670*
Chla	0.0039	0.0918*	0.1341*	-0.0320*	0.0218	0.1432*	0.0204*
cod	-0.0546*	-0.2345*	-0.0950*	0.0367*	-0.2205*	-0.1833*	-0.1091*
doc	-0.0661*	-0.4022*	-0.0080	-0.0702*	0.0607*	-0.2446*	-0.1826*
Mo	-0.0172*	-0.0089	0.0123*	0.0099	-0.0743*	0.0042	0.1316*
Si	-0.0335*	0.1380*	-0.0697*	0.0420*	-0.1789*	-0.0880*	0.2640*
p	-0.0621*	-0.1345*	0.0126*	0.0885*	0.1870*	-0.0679*	-0.0701*

Fe	-0.0026	0.2262*	-0.0275*	-0.0253*	-0.1989*	0.0694*	0.1825*

	no2_	no3_	so42_	Low_Hg	Mn	pH	po43_
no2_	1.0000						
no3_	0.2349*	1.0000					
so42_	-0.1744*	0.0673*	1.0000				
Low_Hg	0.0043	-0.0671*	0.3492*	1.0000			
Mn	-0.1449*	0.1893*	0.4779*	0.3674*	1.0000		
pH	0.2318*	-0.3675*	-0.5380*	-0.2211*	-0.6252*	1.0000	
po43_	0.1689*	0.1384*	-0.1203*	-0.0227*	-0.0982*	0.1494*	1.0000
s	-0.1950*	0.1345*	0.9053*	0.3696*	0.4557*	-0.5663*	-0.1342*
ss	0.1240*	-0.0633*	-0.1845*	-0.0333*	-0.1029*	0.1072*	0.0077
Temp	0.0630*	-0.0771*	-0.0238*	0.0534*	0.0040	-0.0540*	-0.0178*
T_Silica	-0.0896*	0.2473*	0.3091*	0.0611*	0.4608*	-0.5813*	-0.0378*
Turb	-0.0204*	-0.1152*	-0.1688*	0.0356*	-0.0306*	-0.0228*	-0.0251*
nh4_	-0.0580*	0.2917*	0.4024*	0.1017*	0.4185*	-0.5250*	-0.0108
Chla	-0.0342*	-0.1310*	0.0877*	0.1332*	-0.1281*	0.2824*	-0.0399*
cod	0.0019	-0.0659*	-0.2149*	-0.0550*	-0.1509*	0.1585*	0.0490*
doc	0.1798*	-0.1293*	-0.4339*	-0.0791*	-0.3741*	0.5086*	0.1084*
Mo	0.3506*	0.0616*	-0.0121*	0.2235*	-0.0400*	0.0553*	0.0226*
Si	-0.0888*	0.2485*	0.3090*	0.0569*	0.4613*	-0.5798*	-0.0380*
p	0.2196*	0.2139*	-0.1467*	-0.0735*	-0.1026*	0.1271*	0.3997*
Fe	-0.0672*	0.0155*	0.3688*	0.2579*	0.3347*	-0.3531*	-0.0490*

	s	ss	Temp	T_Silica	Turb	nh4_	Chla
s	1.0000						
ss	-0.1908*	1.0000					
Temp	-0.0181*	0.1191*	1.0000				
T_Silica	0.2816*	-0.0421*	0.0921*	1.0000			
Turb	-0.1748*	0.4495*	0.1172*	0.1098*	1.0000		
nh4_	0.3914*	-0.0889*	-0.0171*	0.4106*	-0.0744*	1.0000	
Chla	0.0871*	-0.0764*	0.1166*	-0.2724*	-0.0942*	-0.0613*	1.0000
cod	-0.2205*	0.0726*	0.0453*	-0.0157*	0.1842*	-0.1168*	0.2257*
doc	-0.4562*	0.2118*	0.0307*	-0.2426*	0.2224*	-0.3000*	0.1317*
Mo	-0.0146*	0.1181*	0.0840*	-0.0464*	-0.0400*	-0.0398*	-0.0106
Si	0.2797*	-0.0429*	0.0911*	0.9992*	0.1082*	0.4096*	-0.2750*
p	-0.1633*	0.0182*	0.0381*	0.0554*	-0.0311*	-0.0118*	-0.0532*
Fe	0.2761*	-0.0276*	0.0350*	0.3531*	0.1083*	0.3579*	-0.0873*

	cod	doc	Mo	Si	p	Fe
cod	1.0000					
doc	0.5436*	1.0000				
Mo	0.0334*	0.0810*	1.0000			
Si	-0.0168*	-0.2441*	-0.0451*	1.0000		
p	0.0381*	0.1008*	0.0430*	0.0570*	1.0000	
Fe	-0.0369*	-0.1302*	-0.0176*	0.3519*	-0.0767*	1.0000

Table 8. Case-wise correlation analysis from CN to Fe.

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	5.82041	3.19894	0.5510	0.5510
Factor2	2.62148	0.50078	0.2482	0.7992
Factor3	2.12070	1.29933	0.2008	1.0000

Table 9. Factor analysis/correlation.

Variable	Factor1	Factor2	Factor3	Uniqueness
cn_				0.9977
ec	0.6603			0.4260
do				0.9881
fc				0.9666
Hg	-0.4816			0.7544
Cl_	0.7176			0.1997
f_				0.9921
no2_		0.5019		0.7768
no3_		0.8243		0.3693
so42_	0.8206			0.2361
Low_Hg	0.6888			0.6217
Mn		0.7274		0.5483
pH		-0.4832		0.6090
po43_				0.9908
s	0.8318			0.2598
ss			0.8475	0.3456
Temp			0.3315	0.8679
T_Silica		0.6666		0.2333
Turb			0.8739	0.2462
nh4_		0.7095		0.5037
Chla				0.8587
cod	0.6745		0.4000	0.5787
doc	0.7211		0.3964	0.4579
Mo	0.4133			0.8677
Si		0.6684		0.2326
p				0.9023
Fe			0.6249	0.6065

(blanks represent abs(loading)<.33)

Table 10. Rotated factor loadings (pattern matrix) and unique variances.

EC and Cl^-, together with FC, Hg, F^-, NO_3^-, Low_Hg, Mn, pH, S, SS, Temp, T_Silica, Turb, NH_4^+, COD, Si, P and Fe, were good predictors for SO_4^{2-} concentration, and the fitted model explains 82% of the total variation (Table 12).

4.3 One-way ANOVA

Table 13 gives the means and standard deviations for each of the sampling points over the entire sampling period.

Comparison of SO_4^{2-} by sample_ID (Table 14) showed that K6-K25, K9-K19, V7-VRB37 and V9-VRB24; K10-K10 and K3-R3; and K2-R1 and K4-R4, were statistically similar. The mean values of $SO_4^{2-}$of the remaining sampling points were significantly different.

Variable	Factor1	Factor2	Factor3	Uniqueness
ec	0.6603			0.4260
Hg	-0.4816			0.7544
Cl_	0.7176			0.1997
no2_		0.5019		0.7768
no3_		0.8243		0.3693
so42_	0.8206			0.2361
Low_Hg	0.6888			0.6217
Mn		0.7274		0.5483
pH		-0.4832		0.6090
s	0.8318			0.2598
ss			0.8475	0.3456
Temp			0.3315	0.8679
T_Silica		0.6666		0.2333
Turb			0.8739	0.2462
nh4_		0.7095		0.5037
cod	0.6745			0.5787
doc			0.3964	0.4579
Mo	0.4133			0.8677
Si		0.6684		0.2326
Fe			0.6249	0.6065

(blanks represent abs(loading)<.33)

Table 11. "Clean" factors.

Source	SS	df	MS
Model	122818707	26	4723796.43
Residual	26802038.4	7551	3549.46873
Total	149620746	7577	19746.7

Number of obs	= 7578
F(26, 7551)	= 1330.85
Prob > F	= 0.0000
R-squared	= 0.8209
Adj R-squared	= 0.8203
Root MSE	= 59.577

so42_	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cn_	-22.32404	18.52691	-1.20	0.228	-58.64195	13.99386
ec	.3736444	.0227941	16.39	0.000	.3289616	.4183271
do	.0131522	.0926716	0.14	0.887	-.1685098	.1948143
fc	.0000566	.0000189	2.99	0.003	.0000195	.0000938
Hg	-89.09861	11.70687	-7.61	0.000	-112.0473	-66.14989
Cl_	.7573463	.042237	17.93	0.000	.67455	.8401425
f_	32.3612	8.280861	3.91	0.000	16.12841	48.59399
no2_	-10.90126	13.10631	-0.83	0.406	-36.59327	14.79075
no3_	3.180277	1.003154	3.17	0.002	1.213816	5.146738
Low_Hg	-4.527516	.8473181	-5.34	0.000	-6.188495	-2.866536
Mn	51.43273	4.405735	11.67	0.000	42.79626	60.06919
pH	-7.478322	2.569807	-2.91	0.004	-12.51586	-2.440786
po43_	.8106866	.7992836	1.01	0.310	-.7561315	2.377505
s	1.743953	.0246683	70.70	0.000	1.695596	1.79231

ss	.072502	.0324992	2.23	0.026	.0087946	.1362095
Temp	2.217133	.3666414	6.05	0.000	1.498414	2.935852
T_Silica	9.155261	3.393863	2.70	0.007	2.502346	15.80818
Turb	-.3478313	.0465679	-7.47	0.000	-.4391174	-.2565452
nh4_	-4.445574	.9591881	-4.63	0.000	-6.32585	-2.565299
Chla	.0047781	.0346057	0.14	0.890	-.0630587	.0726149
cod	.326694	.0819311	3.99	0.000	.1660862	.4873018
doc	.0588864	.4554843	0.13	0.897	-.8339896	.9517625
Mo	302.1217	183.4853	1.65	0.100	-57.56057	661.804
Si	-25.85465	7.243482	-3.57	0.000	-40.05389	-11.65541
p	8.823756	2.506464	3.52	0.000	3.910389	13.73712
Fe	40.61979	13.49268	3.01	0.003	14.17039	67.0692
_cons	104.0456	25.89705	4.02	0.000	53.28019	154.811

Table 12. Regression.

	Summary of so42_		
Sample_ID	Mean	Std. Dev.	Freq.
B1-B10	405.26118	140.67122	2526
K1-R2	66.18701	115.52301	2526
K10-K10	120.27818	58.483346	2526
K12-N8	303.80768	116.03529	2526
K2-R1	1128.8242	815.12126	2526
K3-R3	121.64965	170.8744	2526
K4-R4	1123.08	607.58752	2526
K6-K25	172.05588	44.633777	2526
K9-K19	163.85514	45.159634	2526
S1-S1	21.228942	11.581847	2526
S4-S2	346.77498	144.27252	2526
V7-VRB37	159.3354	44.584895	2526
V9-VRB24	154.30907	45.776534	2526
Total	329.7421	462.44325	32838

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	4.1391e+09	12	344925487	3926.94	0.0000
Within groups	2.8832e+09	32825	87835.795		
Total	7.0223e+09	32837	213853.757		

Bartlett's test for equal variances: chi2(12) = 7.4e+04 Prob>chi2 = 0.000

Table 13. One way ANOVA.

(Sidak)

Row Mean-Col Mean	B1-B10	K1-R2	K10-K10	K12-N8	K2-R1	K3-R3
K1-R2	-339.074					
	0.000					
K10-K10	-284.983	54.0912				
	0.000	0.000				
K12-N8	-101.453	237.621	183.529			
	0.000	0.000	0.000			
K2-R1	723.563	1062.64	1008.55	825.017		
	0.000	0.000	0.000	0.000		
K3-R3	-283.612	55.4626	1.37148	-182.158	-1007.17	
	0.000	0.000	1.000	0.000	0.000	
K4-R4	717.819	1056.89	1002.8	819.272	-5.7442	1001.43
	0.000	0.000	0.000	0.000	1.000	0.000
K6-K25	-233.205	105.869	51.7777	-131.752	-956.768	50.4062
	0.000	0.000	0.000	0.000	0.000	0.000
K9-K19	-241.406	97.6681	43.577	-139.953	-964.969	42.2055
	0.000	0.000	0.000	0.000	0.000	0.000
S1-S1	-384.032	-44.9581	-99.0492	-282.579	-1107.6	-100.421
	0.000	0.000	0.000	0.000	0.000	0.000
S4-S2	-58.4862	280.588	226.497	42.9673	-782.049	225.125
	0.000	0.000	0.000	0.000	0.000	0.000
V7-VRB37	-245.926	93.1484	39.0572	-144.472	-969.489	37.6857
	0.000	0.000	0.000	0.000	0.000	0.000
V9-VRB24	-250.952	88.1221	34.0309	-149.499	-974.515	32.6594
	0.000	0.000	0.004	0.000	0.000	0.007

Row Mean-Col Mean	K4-R4	K6-K25	K9-K19	S1-S1	S4-S2	V7-VRB37
K6-K25	-951.024					
	0.000					
K9-K19	-959.225	-8.20074				
	0.000	1.000				
S1-S1	-1101.85	-150.827	-142.626			
	0.000	0.000	0.000			
S4-S2	-776.305	174.719	182.92	325.546		
	0.000	0.000	0.000	0.000		
V7-VRB37	-963.745	-12.7205	-4.51974	138.106	-187.44	
	0.000	1.000	1.000	0.000	0.000	
V9-VRB24	-968.771	-17.7468	-9.54607	133.08	-192.466	-5.02633
	0.000	0.929	1.000	0.000	0.000	1.000

Table 14. Comparison of SO_4^{2-} by Sample_ID.

5. Discussions and conclusions

Case-wise correlation, focussing on SO_4^{2-} , indicated that the variable 'DO' was not significant. Among the other significant variables, it was noted that SO_4^{2-} was highly significantly correlated to EC, Cl^- and S.

Factor analysis yielded some underlying correlations to support the case-wise correlation analysis. In addition to grouping the variables into 3 factors, the variables which were highly correlated to SO_4^{2-} from case-wise correlation, were loaded together with SO_4^{2-} in Factor 1. This was expected because factor analysis is also based on the assumption that all variables are correlated to some degree. Factor 3 was made up of largely physical parameters while Factor 1 contained variables that had something to do with conductivity of a water sample. Factor 2 did not exhibit any cross-loading with the other 2 factors, yet it was still very difficult to assign a common description to it. Variables CN, DO, FC, F^-, PO_4^{3-}, Chl-α and P could be safely deleted as they were not loaded into any of the 3 factors.

Multivariate linear regression indicated that out of the 26 variables that could predict SO_4^{2-}, only 20 were significant, accounting for 82% of the total variation of SO_4^{2-}.
While correlation and regression provided linear relationships, factor analysis, on the other hand, could be used for data reduction. Even though sometimes it is difficult to find a common name to assign to a factor, still, based on these statistical approaches, individual factors or elements within a factor could be further analysed as necessary, with minimal loss of data integrity.
From one-way ANOVA, SO_4^{2-} mean concentration values indicated that monitoring point K2-R1 (1128.82±815 mg/L) was within the vicinity of the source of SO_4^{2-}. Attenuation of the variable was noted as its mean value decreased along the Rietspruit River at K4-R4 and then Klip River at K6-K25 and K9-K19, before Klip River discharged into the Vaal River. From monitoring point B1-B10 (also close to a source of SO_4^{2-}), another established route was through S4-S2, before Suikerbosrant River discharged into the Vaal River upstream of the Klip River. Surface raw water containing high levels of SO_4^{2-} was not draining via K1-R2 and S1-S. Based on SO_4^{2-} mean concentration values only and for management purposes, K1-R2 and S1-S could be left out of the monitoring programme, saving on financial resources. Comparison of SO_4^{2-} by sample_ID showed that K6-K25, K9-K19, V7-VRB37 and V9-VRB24; K10-K10 and K3-R3; and K2-R1 and K4-R4, were significantly similar.
The major challenge was pre-processing of the non-consistent water quality data over the 7 years. Non-consistent data was as a result of missing data, largely where some of the stakeholders dropped or established some water quality variables and monitoring points over the years as monitoring prioritizations changed because of new and emerging pollution threats. The challenge of insufficient and inconsistent data for water quality modelling remains a limitation in the formulation of good and practically useable models. However, interpolations and correlations, including factor analysis and regression, could help build better data sets, especially for pollution trending in river basin management. This could be used to support large-scale public decisions.

6. Acknowledgement

The financial assistance of the South African Department of Science Technology (DST) is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the DST. The authors would also like to thank Tshwane University of Technology for hosting and co-funding this research. DWA, the Water Research Commission, Rand Water Board (co-funding), Midvaal Water Company and Sedibeng Water, are also sincerely acknowledged, especially for providing very valuable and vital data.

7. References

Alfassi, Z. B., Boger, Z. & Ronen, Y. (2005). *Statistical treatment of analytical data,* Oxford, Blackwell Science Ltd, 0-632-05367-4, CRC Press, Australia.

Cloot, A. & Roux, G. L. (1997). Modelling algal blooms in the middle Vaal River: a site specific approach. *Water Research,* 31, 2, 271-279, 0043-1354.

DWAF (2007). Integrated water quality management plan for the Vaal River system. Pretoria, South Africa.

Dzwairo, B. & Otieno, F. A. O. (2010). Integrating quality and cost of surface raw water: Upper and Middle Vaal Water Management Areas South Africa. *Water Science and Technology: Water Supply* 10, 2, 201–207, 1606-9749.

Dzwairo, B., Otieno, F. A. O. & Ochieng', G. M. (2010a). Making a case for systems thinking approach to integrated water resources management (IWRM). *International Journal of Water Resources and Environmental Engineering,* 1, 5, 107-113 2141-6613.

Dzwairo, B., Otieno, F. A. O., Ochieng', G. M. & Letsoalo, M. A. (2010b). Downsizing water quality data for river basin management – Focussing on Sulphate: Vaal River, South Africa. *Proceedings of the 11th WaterNet/WARFSA/GWP-SA Symposium: 'IWRM for National and Regional Integration: Where Science, Policy and Practice Meet. Elephant Hills Hotel, Victoria Falls, Zimbabwe. 27 October - 29 October, 2010.*

Even, S., Billen, G., Bacq, N., Théry, S., Ruelland, D., Garnier, J., Cugier, P., Poulin, M., Blanc, S., Lamy, F. & Paffoni, C. (2007). New tools for modelling water quality of hydrosystems: An application in the Seine River basin in the frame of the Water Framework Directive. *Science of The Total Environment,* 375, 1-3, 274-291.

Gouws, K. & Coetzee, P. P. (1997). Determination and partitioning of heavy metals in sediments of the Vaal Dam System by sequential extraction. *Water SA,* 23, 3, 217-226, 0378-4738.

Herold, C. E., Le Roux, P. J., Nyabeze, W. R. & Gerber, A. (2006). WQ2000 Salinity Model: enhancement, technology transfer and implementation of user support for the Vaal system. Umfula Wempilo Consulting. Pretoria, South Africa.

Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS.,* Florida, Chapman and Hall/CRC: Tailor and Francis Group, 1584886021.

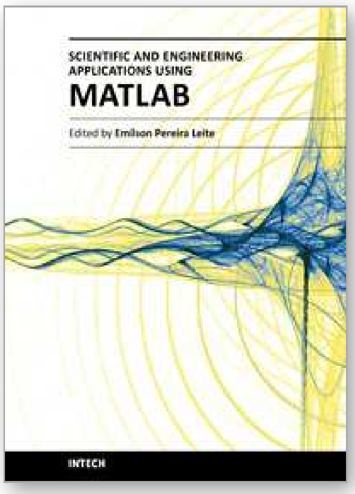
Martin, J. L., Mccutcheon, S. C. & Martin, M. L. (1998). *Hydrodynamics and Transport for Water Quality Modeling* Taylor & Francis, Inc, 978-0873716123.

Naicker, K., Cukrowska, E. & Mccarthy, T. S. (2003). Acid mine drainage arising from gold mining activity in Johannesburg, South Africa and environs. *Environmental Pollution,* 122, 1, 29-40, 0269-7491.

Ochse, E. (2007). *Seasonal rainfall influences on main pollutants in the Vaal River barrage reservoir: a temporal-spatial perspective.* Magister Artium MA, University of Johannesburg.

Pieterse, A., Roos, J., Roos, K. & Pienaar, C. (1987). Preliminary observations on cross-channel and vertical heterogeneity in environmental and algological parameters in the Vaal River at Balkfontein, South Africa. *Water SA,* 12, 4, 173-184, 0378-4738 .

Stevn, D. J. & Toerien, D. F. (1976). Eutrophication levels of some South African impoundments. IV. Vaal dam. *Water SA,* 2, 2, 53-57.



Scientific and Engineering Applications Using MATLAB

Edited by Prof. Emilson Pereira Leite

ISBN 978-953-307-659-1
Hard cover, 204 pages
Publisher InTech
Published online 01, August, 2011
Published in print edition August, 2011

The purpose of this book is to present 10 scientific and engineering works whose numerical and graphical analysis were all constructed using the power of MATLABÂ® tools. The first five chapters of this book show applications in seismology, meteorology and natural environment. Chapters 6 and 7 focus on modeling and simulation of Water Distribution Networks. Simulation was also applied to study wide area protection for interconnected power grids (Chapter 8) and performance of conical antennas (Chapter 9). The last chapter deals with depth positioning of underwater robot vehicles. Therefore, this book is a collection of interesting examples of where this computational package can be applied.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Bloodless Dzwairo, George M. Ochieng', Maupi E. Letsoalo and Fredrick A.O. Otieno (2011). Data Reduction for Water Quality Modelling, Vaal Basin, Scientific and Engineering Applications Using MATLAB, Prof. Emilson Pereira Leite (Ed.), ISBN: 978-953-307-659-1, InTech, Available from: http://www.intechopen.com/books/scientific-and-engineering-applications-using-matlab/data-reduction-for-water-quality-modelling-vaal-basin

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821