Water quality investigation in the Hawkesbury-Nepean River in Sydney using Principal Component Analysis

U. Kuruppu [a], **A. Rahman** [a], **M. Haque** [a] **and A.Sathasivan** [a]

[a] *School of Computing, Engineering and Mathematics, University of Western Sydney, Australia*
E-mail: u.kuruppu@uws.edu.au

Abstract: The Hawkesbury-Nepean River System (HNRS) is an icon of Australia's largest city Sydney. Although there are a number of dams and in-stream structures throughout the river system, the HNRS is considered to be an unregulated river. Since European settlement, the HNRS has been used as a primary water source to meet the drinking water needs for over 80% of Sydney's population. The fact that it is located in the peri-urban areas dictates that it receive pollutants from a number of sewage treatment plants, located within the catchment, and the storm water runoff from agricultural and urban lands. Sydney Catchment Authority regularly monitors the water quality of the HNRS which generates a large three dimensional (different sampling stations and different parameters over time) data set containing useful information on pollutant build up and washoff on/from this river system.

In this study, factor analysis was used to identify the most significant water quality monitoring station(s). It was found that the three principal components explained more than 90% of the total variance in the data set. Moreover, this study showed that the dimensionality of the water quality parameters can be reduced to eight principal components which explained more than 70% of the total variance. Information obtained in this study can be used to design an optimal sampling strategy, which could reduce the number of sampling stations in the river system. However, further study is needed to confirm this initial finding.

Keywords: *Hawkesbury-Nepean River, principal component analysis, factor analysis, water quality*

1. INTRODUCTION

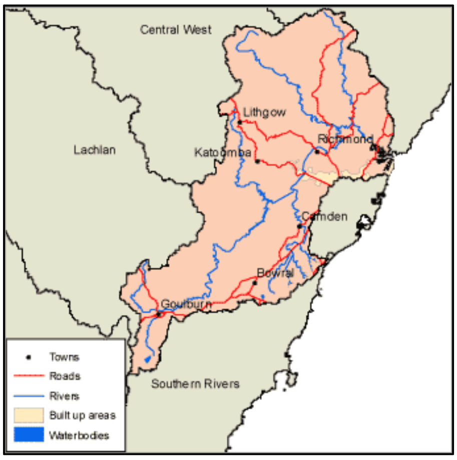

Figure 1. Hawkesbury-Nepean River System (CFOC, 2013)

Hawkesbury-Nepean River System (HNRS, shown in Figure 1) is the main source of fresh drinking water supply to more than 4.8 million people living in and around Sydney. The HNRS system is a combination of two major rivers (Figure 1), the Nepean River (155 km) and the Hawkesbury River (145 km) (Markich and Brown, 1998). The river system is complex in nature, the upper part contains poorly accessible gorges, the middle part is running through irrigated farm lands and the lower part has tidal slopes with deposited soil pockets (Diamond, 2004). The middle part of the river is being continuously influenced by increasing population growth, urbanization, industrialization and other human activities which cause contamination of the quality of the river water from different sources (e.g. sewage, stormwater, runoff from disused mines, toxic forms of blue-green algae, and waste from domestic and native animals). Pinto and Maheshwari (2011) have shown that river health in peri-urban landscapes is prone to higher degrees of degradation. Within the HNR catchment, vegetation clearance has been continuously practised over the last 200 years causing increased subsurface and agricultural runoff and sediment loads into the river system (Thomas et al. 2000). Land use in the HNR catchment includes regions that are heavily peri-urbanised and industrialised and which are important for recreational and agricultural activities and tourism (Baginska et al. 2003). Agricultural runoff contributes approximately 40 % to 50 % of phosphorus loads and 25 % of nitrate loads into the HNRS which are believed to have originated from agricultural and animal farms (Markich and Brown 1998).

A regular water quality monitoring program generates reliable data which reflect the state of the water quality of a river. However, generating good data is not enough to meet the objectives of a water quality monitoring program. Data must be processed and presented in a manner that provides the understanding of the spatial and temporal patterns in water quality parameters. The intent is to use the collected set of data to explain the current state of the water more widely and make the necessary controls to overcome future water quality issues. One of the problems with many sets of multivariate data generated from a monitoring program is that there are too many variables to analyse and then to draw meaningful conclusion from this.

The multidimensionality (i.e. different sampling stations and different parameters over time) of data make analysis more complicated. Principal component analysis (PCA) and factor analysis (FA) are the two multivariate techniques with the central aim of reducing as much as possible of the laity of a multivariate data set while retaining their variation/useful information as much as possible. This objective is achieved by transforming the original variables to a new set of hypothetical variables called principal components or factors (PC/F) that are uncorrelated. They are obtained as a linear combination of the original variables. Principal components or factors explain the original variance in a monotonically decreasing way (Kovans et al., 2012). Factor analysis is similar to principal component analysis, but the two are not identical. In FA, components extracted from PCA are rotated according to a mathematically established rule (i.e., varimax, equamax and quarimax) yielding easily interpretable new variables, called varifactors (VFs) (Pinto et al 2013). FA use regression modeling techniques to test hypotheses producing error terms, while PCA is a descriptive statistical technique (Bartholomew et al., 2008). The difference between PCs obtained in PCA and VFs obtained in FA is that PCs are linear combinations of observable water quality parameters but VF are unobservable, hypothetical and latent variables (Alberto et al. 2001).

The differences between PCA and FA are further illustrated by Suhr (2009) as follows:

- PCA results in principal components that account for a maximal amount of variance for observed variables. FA account for common variance in the data. PCA inserts ones on the diagonals of the correlation matrix. FA adjusts the diagonals of the correlation matrix with the unique factors.

- PCA minimizes the sum of squared perpendicular distance to the component axis. FA estimates factors which influence responses on observed variables.
- The component scores in PCA represent a linear combination of the observed variables weighted by eigenvectors. The observed variables in FA are linear combinations of the underlying and unique factors.

In this study, PCA and FA both were applied to find out the most significant water quality monitoring stations in the HNRS. In addition, PCA was applied to find out the most significant water quality parameters in the HNRS. The StatistiXL software package was employed for the data analysis.

2. STUDY AREA AND DATA

In this study, a total 28 physical, chemical and biological water quality parameters were considered in PCA and FA. These parameters were measured fortnightly by Sydney Catchment Authority at 15 different water quality monitoring stations of the HNRS. Descriptions and locations of water quality monitoring stations are presented in Table 1 and Figure 2, respectively. The descriptive statistics of the water quality parameters are presented in Table 2.

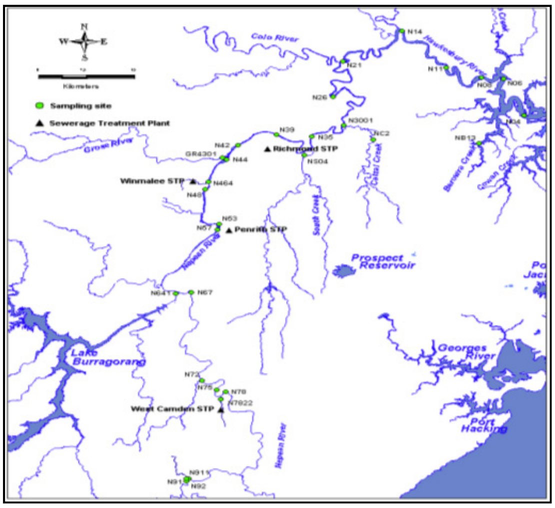

Figure 2. Map of water quality monitoring stations along Hawkesbury-Nepean River System (SCA, 2001)

Table 1. Water quality monitoring stations

Site code	Site
E851	Shoalhaven River downstream of Tallowa Dam
N14	Hawkesbury River at Wisemans Ferry downstream of Car Ferry
N21	Hawkesbury River at Lower Portland upstream of Colo River
N35	Hawkesbury River at Wilberforce upstream of Cattai Creek
N42	Hawkesbury River at North Richmond upstream of North Richmond Water Treatments Works
N44	Nepean River at Yarramundi Bridge upstream of Grose River
N57	Nepean River at Penrith Weir upstream of Boundary Creek and Penrith Sewage Treatment Plant
N64	Nepean River 500m downstream of confluence of Warra river
N641	Warragamba River (North Basin) downstream of Warragamba STP
N67	Nepean River at Wallacia Bridge upstream of Warragamba River
N75	Nepean River at Sharpes Weir downstream of Matahil Creek and Camden Sewage Treatment Plant
N85	Nepean River at Menangle Bridge
N86	Nepean River at Pheasants Nest
N881	Nepean River at downstream of Broughtons pass Weir
N92	Nepean River at Maldon Weir upstream of Stonequarry Creek and Picton Sewage Treatment Plant

Table 2. Descriptive statistics and abbreviations of the data set

Water quality Parameter	Abbreviation	Units	Min	Max	Median	No. of data values
pH	PH		5.78	9.94	7.63	1666
Lorenzen	LOR	ug/L	0.10	539.90	4.40	1666
Iron Total	TI	mg/L	0.04	5.62	0.29	1666
Phaeophytin	PHA	ug/L	0.10	25.20	0.80	1666
Nitrogen TKN	TKN	mg/L	0.02	5.40	0.27	1666
Temperature	TEMP	Deg C	8.10	30.60	19.50	1666
Chlorophyll-a	CHLA	ug/L	0.20	253.10	5.10	1666
E. coli	ECOL	orgs/100mL	0.00	6100.00	13.00	1666
Iron Filtered	FI	mg/L	0.01	3.43	0.09	1666
True Colour	TCOL		1.00	93.00	11.00	1666
Nitrogen Total	TN	mg/L	0.08	5.90	0.45	1666
Turbidity	TUR	NTU	-0.60	380.00	3.85	1666
Alkalinity	ALK	mgCaCO3/L	1.00	298.00	40.00	1666
Aluminium Total	TA	mg/L	0.01	3.97	0.08	1666
Manganese Total	TM	mg/L	0.00	0.48	0.03	1666
Dissolved Oxygen	DO	mg/L	1.50	16.20	9.10	1666
Enterococci	ECOCC	cfu/100mL	0.00	8400.00	20.00	1666
Phosphorus Total	TP	mg/L	0.01	0.18	0.01	1666
Suspended Solids	SS	mg/L	1.00	105.00	3.00	1666
Nitrogen Oxidised	NO	mg/L	0.00	5.00	0.17	1666
Aluminium Filtered	FA	mg/L	0.00	0.45	0.01	1666
Manganese Filtered	FM	mg/L	0.00	0.35	0.01	1666
Conductivity Field	EC	mS/cm	0.01	48.40	0.30	1666
Nitrogen Ammonical	NH-N	mg/L	0.01	0.41	0.01	1666
Phosphorus Filterable	FP	mg/L	0.00	0.11	0.01	1666
Silicate Reactive	RS	SiO2 mg/L	0.01	14.90	1.71	1666
Dissolved Organic Carbon	DOC	mg/L	0.20	350.00	4.60	1666
UV Absorbing constituents	UV		0.01	0.93	0.12	1666

3. METHODOLOGY

In this study, PCA was performed first to identify the most important water quality monitoring station(s) in the HNRS. For the purpose of this analysis, the median value of each parameter was used as the median is better suited for a skewed distribution to describe the central tendency of the data. In this analysis, stations with correlation coefficient greater than 0.9, were taken as principal water quality monitoring stations. Equations for principal components were derived by considering the loadings of the variables (water quality monitoring stations).

To further identify the monitoring stations that are important in revealing surface water quality variations, a FA was employed. Varimax rotation was selected as the data rotation method, as it makes an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor on all the variables in a factor matrix which has the effect of differentiating the original variables by extracted factor. Each factor has either large or small loadings of any particular variable. A varimax solution was used to identify each variable with a single factor. This is the most common rotation option used in the PCA and PF analysis. However, the orthogonality (i.e., independence) of factors is often an unrealistic assumption (Russell, 2002). In the second step, PCA was performed on water quality data to identify the principal components that explain the most of the variance in the water quality data set.

4. RESULTS AND DISCUSSION

When 15 monitoring stations were reduced to three principal components, it explained 95.2% of the total variance and the rest of the 12 components only accounted for 4.8% (Table 3). Further, the first, second and third components accounted for about 79.6%, 8.8% and 6.6% of the total variance in the data set, respectively. Therefore the discussion is focused only on the first three principal components.

Table 3. Principal components with eigenvalues > 1

Value	PC 1	PC 2	PC 3
Eigenvalue	11.96	1.328	0.993
Percentage of Variance	79.731	8.855	6.62
Cumulative Percentage	79.731	88.586	95.206

Table 4. Component score coefficients for first three PCs (for monitoring stations)

Variable	PC 1	PC 2	PC 3
E851	0.218	0.316	-0.386
N14	0.265	-0.221	-0.176
N21	0.243	-0.116	-0.338
N35	0.272	-0.081	0.098
N42	0.287	0.09	0.03
N44	0.249	0.094	0.489
N57	0.229	0.145	0.572
N64	0.279	-0.106	-0.197
N641	0.284	-0.031	-0.061
N67	0.278	-0.17	0.15
N75	0.274	-0.261	0.022
N85	0.284	-0.073	0.076
N86	0.234	0.489	0.006
N881	0.21	0.547	-0.207
N92	0.251	-0.374	-0.119

The first component had almost equal loadings on all variables (Table 4) and therefore was a measure of overall performance of the stations and also it showed an extremely high correlation with the measured data. It accounted for 79.7% of the data variance (Table 3). Similarly, the second and third components had different loadings on different variables. Hence, PC2 and PC3 represented a difference among the stations. Loading reflected only the relative importance of a variable within a component, and did not reflect the importance of the component itself (Davis, 1986).

The results of the first PCA identified three important components that accounted for 95.2% of the variance in the dataset.

Table 5 demonstrates the rotated factor correlation coefficient (obtained from FA) for 15 water quality monitoring stations. In this study, the factor correlation coefficient was considered to be significant if the value was greater than 0.7. This conservative criterion was selected because the study area was large and the river system was deemed to be highly non-linear and dynamic. From Table 5 water quality monitoring stations N14, N64, N641, N67, N75, N85, N86, N881, N92, N57 and N21 have coefficient values greater than 0.70, and hence these are considered to be the most important water quality monitoring stations.

Table 5. Varimax rotated factor loadings for first 5 factors

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
E851	0.378	0.672	0.125	-0.19	0.594
N14	**0.766**	0.265	0.27	-0.452	0.081
N21	0.582	0.339	0.147	**-0.717**	0.098
N35	0.555	0.267	0.566	-0.524	0.118
N42	0.621	0.536	0.498	-0.266	0.07
N44	0.404	0.303	0.85	-0.128	0.058
N57	0.293	0.288	**0.904**	-0.099	0.033
N64	**0.818**	0.432	0.248	-0.27	0.094
N641	**0.768**	0.473	0.373	-0.185	0.07
N67	**0.776**	0.244	0.537	-0.174	0.082
N75	**0.842**	0.189	0.424	-0.244	0.116
N85	**0.749**	0.368	0.498	-0.175	0.11
N86	0.261	**0.846**	0.43	-0.12	0.037
N881	0.195	**0.926**	0.238	-0.184	0.074
N92	**0.946**	0.119	0.227	-0.16	0.111

PCA for the water quality parameters dataset developed eight principal components with eigenvalues > 1 explaining about 72.7% of the total variance in the data set. The first PC accounts for 24.1% of the total variance which was highly correlated (loading > 0.7) with total iron (TI), true color (TCOL), turbidity, aluminum total, and UV absorbent. Whereas, the other seven PCs, although accounted for 12.7%, 8.3%, 7.3%, 6.6%, 5.2%, 4.4% and 3.8%, respectively, correlated (loading > 0.7) with none of the parameters (Table 6 and Table 7).

Principal components extracted for water quality parameters did not have a strong correlation when comparing with principal components extracted for water quality monitoring stations. Monitoring stations are primarily controlled by hydrological conditions, while water quality parameters are controlled by a combination of hydrological, chemical, physical and biological conditions, so it is expected that the monitoring stations would have higher correlation than the water quality parameters.

Table 6. Explained variance and eigenvalues (for water parameters)

Value	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
Eigenvalue	6.756	3.559	2.343	2.067	1.851	1.460	1.243	1.073
Percentage of variance	24.130	12.712	8.366	7.383	6.612	5.215	4.441	3.832
Cumulative percentage	24.130	36.842	45.209	52.591	59.203	64.419	68.859	72.691

Table 7. Component loadings for first eight PCs (water quality parameters)

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
PH	-0.404	0.450	-0.092	0.065	-0.148	-0.036	0.351	-0.464
LOR	0.052	0.402	0.080	-0.599	0.517	0.092	0.207	-0.112
TI	**0.907**	-0.133	0.080	-0.078	-0.007	0.179	-0.016	-0.030
PHA	0.124	0.378	0.032	-0.295	0.045	0.081	-0.186	0.202
TKN	0.322	0.515	-0.239	-0.102	-0.038	-0.218	0.081	0.042
TEMP	0.059	0.168	0.487	-0.315	-0.260	-0.629	0.089	0.019
CHLA	0.085	0.507	0.080	-0.630	0.502	0.089	0.141	-0.059
ECOL	0.459	0.169	0.250	0.324	-0.167	0.170	0.426	0.183
FI	0.504	-0.554	-0.220	-0.236	-0.011	0.050	0.059	-0.161
TCOL	**0.754**	-0.342	-0.207	-0.016	0.195	-0.296	0.061	-0.134
TN	0.172	0.618	-0.665	-0.017	-0.110	-0.063	-0.073	0.250
TUR	**0.748**	0.294	0.288	0.203	0.013	0.205	0.101	0.055
ALK	-0.236	0.482	-0.157	-0.003	-0.424	-0.122	0.143	-0.448
TA	**0.802**	0.220	0.170	0.272	0.138	0.052	-0.080	-0.064
TM	0.553	-0.207	0.038	-0.535	-0.383	0.264	0.017	0.017
DO	-0.243	-0.042	-0.491	0.258	0.380	0.530	0.134	-0.213
ECOCC	0.450	0.201	0.225	0.328	-0.188	0.156	0.504	0.180
TP	**0.700**	0.428	0.063	0.107	0.066	-0.025	-0.166	-0.087
SS	0.605	0.409	0.361	0.033	0.076	0.220	-0.295	-0.069
NO	0.061	0.510	-0.694	0.026	-0.117	0.022	-0.123	0.283
FA	0.487	-0.288	-0.198	0.208	0.332	-0.232	0.008	-0.254
FM	0.486	-0.430	-0.171	-0.415	-0.405	0.287	0.096	-0.027
EC	-0.046	0.116	0.305	0.046	-0.127	0.193	-0.554	-0.219
NH-N	0.498	-0.038	-0.345	-0.222	-0.477	0.044	-0.059	-0.144
FP	0.527	0.395	-0.082	0.275	-0.110	-0.132	-0.213	-0.259
RS	0.570	-0.352	-0.272	0.106	0.165	-0.202	-0.033	0.079
DOC	0.117	0.039	-0.041	-0.014	0.092	-0.170	0.004	0.266
UV	**0.742**	-0.155	-0.122	0.010	0.211	-0.303	0.096	-0.036

4. CONCLUSION

Water-quality monitoring programs generate complex multidimensional data. Multivariate statistical techniques can be used to extract useful information from this data. In this case study, factor analysis was performed to identify the most significant water quality monitoring stations in the Hawkesbury-Nepean River System. The stations N14, N64, N641, N67, N75, N85, N86, N881, N92, N57 and N21were found to be the most significant sampling sites explaining the most variation in the water quality data in the Hawkesbury-Nepean River System. This result might be used to reduce the number of sampling sites in the river system. Principal component analysis allowed deriving three principal components which explained more than 90% of the total variance in data set. The findings of this preliminary data analysis need further confirmation by a more in-depth analysis, which is being undertaken.

ACKNOWLEDGEMENTS

The authors would like to thank Sydney Catchment Authority, New South Wales, Australia for providing water quality data used in this study.

REFERENCES

Alberto,W., Diaz,M.P., Ame,M.V., Pesce,S.B., Hued,A.C., and Bistoni,M.A. (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A Case Study: Suqu a River Basin (Córdoba–Argentina). *Water Research,* 35, 2881–2894.

Baginska,B., Pritchard,T., and Krogh,M. (2003). Roles of land use resolution and unit-area load rates in assessment of diffuse nutrient emissions. *Journal of Environmental Management,* 69, 39–46.

Bartholomew, D.J., Steele, F., Galbraith, J., and Moustaki, I. (2008). Analysis of Multivariate Social Science Data. *Statistics in the Social and Behavioral Sciences Series (2nd ed.)*

CFOC (2013). Caring for our Country, accessed on http://www.nrm.gov.au/about/nrm/regions/nsw-hnep.html

Davis, J.C. (1986). *Statistical and data analysis in geology, second ed. John Wiley & Sons, New York.*

Diamond, R. (2004).Water andSydney'sFutureeBalancing theValue of Our Rivers and Economy. Final Report of the Hawkesbury- Nepean River Management Forum. NSW, *Department of Infrastructure, Planning and Natural Resources, Sydney.*

Jackson, J.E. (1991). A User's Guide to Principal Components. *John Wiley & Sons, New York.*

Kovacs, J., Tanos, P., Korponai, J., Szekely, I.K., Gonda, K., Soregi, K.J., Hatvani, I.G. (2012). Analysis of Water Quality Data for Scientists. *Water Quality Monitoring and Assessment (Chapter3)*

Markich, S.J., and Brown, P.L. (1998). Relative importance of natural and anthropogenic influences on the fresh surface water chemistry of the Hawkesbury–Nepean River, southeastern Australia. *Science of the Total Environment,* 217, 201–230.

Pinto, U., Maheshwari, B.L., Ollerton, R.L. (2013). Analysis of long-term water quality for effective river health monitoring in peri-urban landscapes—a case study of the Hawkesbury–Nepean river system in NSW, Australia. *Enviromental Monitoring and Assessment,* June 2013.

Pinto, U., Maheshwari, B.L. (2011), River health assessment in peri-urban landscapes: An application of multivariate analysis to identify the key variables, *Water Research,* 45, 3915 -3924

Russell, D.W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin,* 28 (12): 1629–46.

SCA (2001). Annual water quality monitoring report. *Sydney Catchment Authority,* 2000-2001.

Suhr, D. (2009). Principal component analysis vs. exploratory factor analysis. *SUGI 30 Proceedings.* Retrieved 5 April 2012.

Thomas, M., Parker, C., and Simons, M. (2000). The dispersal and storage of trace metals in the Hawkesbury River valley. *S. Brizga and B. Finlayson (Eds.), River management: the Australasian experience* (pp. 197–219). Sydney: John Wiley and Sons.