SELECTION OF THE DETERMINATES TRUSSU RIVER WATER QUALITY FACTORS USING MULTIVARIABLE ANALYSIS

H. de Araújo Queiroz Palácio[1], Eunice Maia de Andrade[2], L. Araújo Crisostomo[3], A. dos Santos Teixeira[3]
Ivam Holanda de Souza[4]

ABSTRACT
Principal component analysis (PCA) was applied to evaluate and interpret a large water quality data set and apportionment of pollution sources factors with a view to get better information about water quality of the Trussu River valley. The investigation was carried out in the part of the valley where several farms with livestock activities and some villages are located. Water quality parameters were sampled from September/2002 to March/2004 at nine stations located along 24 km of the Trussu River for thirteen physical-chemical (2,223 observations). The PCA application resulted in three significant components, explaining 73.78% of the total variance of the data set. The first one, PC_1 (accounting for 48.36% of the variance) was mainly associated with sodium, Electric Conductivity, chloride, magnesium, sulphate and hydrogen-carbonate. It, basically, reflects ionic group of salts (mineralization processes). PC_2 (15.91% of the variance), was dominated by organic contaminations in water (NO_3-N and NH_4-N), suggesting anthropic activities. PC_3 was mainly contributed by pH e PO_4 and than, may be related to the effects caused in the water by non-point sources pollutants, such as agricultural runoff. This study suggests that PCA technique is a useful tool for identification of important surface water quality monitoring parameters.
Key Words: *Water quality. Data reduction. Multivariate analysis*

*

1. INTRODUCTION

World population growth, increased water use by the world population associated with improvements in life quality and absence of policies for a more conservative use associated with the increased man made contamination of water resources has decreased water availability (VEGA et al., 1998). Water plays a vital role in determining life quality of humankind and other living creatures.

In the last decades there has been innumerous studies on agriculture and its role as a nonpoint source pollutant, in water contamination in many countries (LAKE et al., 2003; ELMI et al., 2004). The large number of variables involved in the characterization of the quality of water bodies has prompted to the application of multivariate statistical methods of analysis such as principal components analysis (PCA).

[1] *Escola Agrotécnica Federal de Iguatu-Ce, Brazil*
[2] *Departamento de Eng. Agrícola, CCA/UFC, Fortaleza-Ce, Brazil*
[3] *Pesquisador da EMBRAPA/CNPAT, Brazil*
[4] *Escola Agrotécnica Federal de Iguatu-Ce, Brazil*

This technique has been applied to investigate a variety of hydrological, meteorological and chemical processes (HELENA et al., 2000; GANGOPADHYAY et al., 2001; TOLEDO; NICOLELLA, 2002; MENDIGUCHÍA et al., 2004).

PCA main goal is to apply linear techniques to transform a large set of variables into an uncorrelated smaller one containing most of the information stored in the original set (HOTTELING, 1933). Applications of PCA in water quality studies have been reported by several authors, such as Simeonov et al. (2003); Brodnjak-Voncina et al. (2002); Bordalo et al. (2001). The main goal of this paper is to apply PCA to identify the most important variables describing the quality of water at the Trussu River, an originally intermittent stream now regularized by the Trussu dam in Iguatu County, Ceará State, Brazil.

2. MATERIALS AND METHODS

The Trussu River is a tributary of the Jaguaribe River system in Ceará State, Brazil. It originates at an altitude of 580 m and flows eastwards for approximately 42 km before discharging into Jaguaribe River. Weather is identified by hot dry summer days and warm winter nights and classified according to Köopen as type BSh'w'. The average maximum daily temperature in October is around 34 ºC while the colder nights of July can reach 20 ºC. Rainfall distribution is unimodal and strongly concentrated in the autumn months (March to May). As a general rule, more than 80 percent of the annual rain falls during this period. Average annual rainfall is 750 mm year^{-1}. The annual potential evapotranspiration is 2,940 mm, which is about four times the average annual rainfall. In average, there are 2,945 hours of sunshine during a year.

The investigated river run is located approximately between the parallels 6°20'59" S and 6°16'48" S, and meridians 39°27' W and 39°16' W (Figure 1), with mean elevation of 300 m over the sea level. It runs 25 km from the Trussu dam to the mouth of Trussu River. Some small villages are located in the studied area.

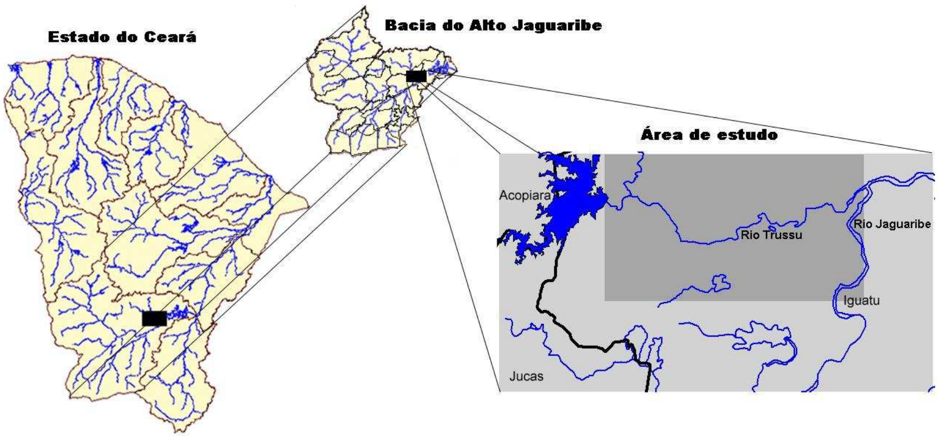

Fig.1 *Location of studied area*

Nine points spread out over the studied area were selected as monitoring stations. They were designated as: EA1, EA2, EA3, EA4, EA5, EA6, EA7, EA8, and EA9. Stations EA1, EA2, EA5, EA6, EA7 were distributed along the river course to monitor surface water, while stations EA3, EA4, EA8, and EA9 are shallow wells designed to monitor ground water (Figure 2). Station EA1 has never received any municipal or agricultural wastewater, and the water quality in this station reflects pollution from overland flow. AE2 (Pedreira farm) receives water from agricultural fields; while AE5 (Santa Clara village) and AE6 (Varjota village) receive water from urban areas. Water quality in station EA7 (Barreira dos Pinheiros village) reflects the effect of contaminants upstream from it. AE3 (Pedreira farm) and AE4 (Varzea da Lama) receives water from agricultural sites; while AE8 (Suassurana village) and AE9 (Varjota village) receive water from human settlements.

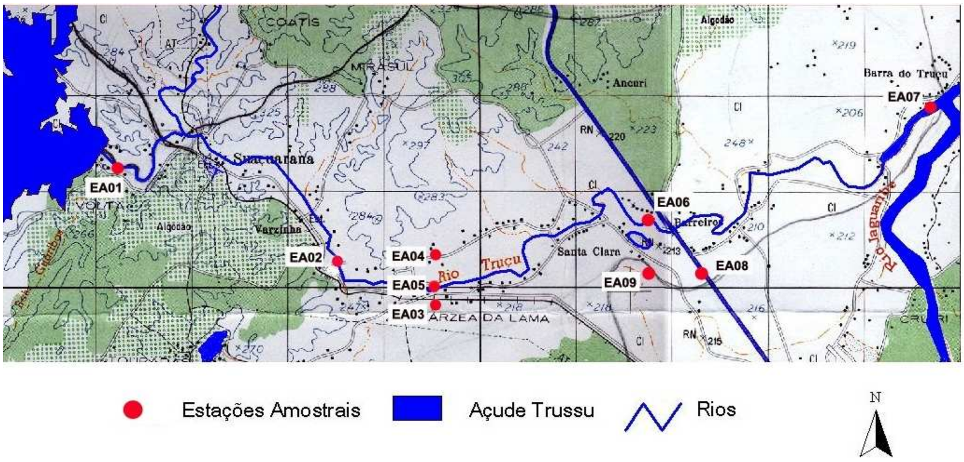

Fig. 2 ***Location of sampling sites***

Samples were collected monthly from September/2002 to March/2004 for thirteen physical-chemical parameters (2,223 observations). Samples were stored in polyethylene bottles provided with an hermetic-locking cap, and immediately transported to the laboratory and stored at 4 ºC before being analyzed to determine water quality variables. Units and methods of analysis of these variables are presented in Table 1.

Multivariate statistical methods, such as Factor Analysis/Principal Component Analysis (FA/PCA) allow the identification of important components or factors that explain most of the variance of a system. In this study, factors were estimated from principal component methods. The number of factors, called principal component (PCs), were defined according to the criterion that only factors that account for variances greater than 1 (eigenvalue-one criterion) should be included. The rationale for this criterion is that any component should account for more variance than any single variable in the standardized test score space. The dataset was standardized through z-scale transformation in order to avoid misclassification due to wide differences in data dimensionality. Standardization reduces the influence of variables variance and eliminates the effect of different units of measurement.

Water quality variables measured during 2002-2004 in Trussu valley, Brazil

Table 1

Variable	Abbreviation	Unit	Analytical method
Ammonical nitrogen	NH_4–N	mg L^{-1}	Spectrophotometric
Calcium	Ca	$mmol_c\ L^{-1}$	Flame AAS
Chloride	Cl	$mmol_c\ L^{-1}$	Spectrophotometric
Electrical Conductivity	EC	dS m^{-1}	Electrometric
Hydrogen-carbonate	HCO_3	$mmol_c\ L^{-1}$	**Richard (1954)**
Magnesium	Mg	$mmol_c\ L^{-1}$	Flame AAS
Nitrate nitrogen	NO_3–N	mg L^{-1}	Spectrophotometric
pH	pH	dimensionless	pH-meter
Phosphate	PO_4	$mmol_c\ L^{-1}$	Spectrophotometric
Potassium	K	$mmol_c\ L^{-1}$	Flame photometer
Sodium adsorption ratio	SAR	dimensionless	Richard (1954)
Sodium	Na	$mmol_c\ L^{-1}$	Flame photometer
Sulphate	SO_4	$mmol_c\ L^{-1}$	Spectrophotometric

After the correlation matrix definition, the appropriateness of the factor model was evaluated. A measure of sampling adequacy was computed using the Kayser Mayer Olkim (KMO) index, which compares the magnitude of the observed correlation coefficients to the magnitude of the partial correlation coefficients. If variables share common factors, partial correlation coefficients between pairs of variables should be small when the linear effects of the other variables are eliminated. Factor analysis model is acceptable when KMO > 0.5 (Monteiro and Pinheiro, 2004).

Although the factor matrix obtained in the extraction phase indicates the relationship between the factors and the individual variables, it is usually difficult to identify meaningful factors based on this matrix. Interpretation of the matrix may be easier using the rotation procedure. Rotation does not affect goodness of fit of a factor solution. That is, although the factor matrix changes, the communalities and the percentage of total variance explained, does not change. The rotation process in factor analysis allows flexibility by presenting a multiplicity of views of the same data set (Dillon and Goldstein, 1984).

3. RESULTS AND DISCUTIONS

The Pearson correlation matrix for the complete set of variables analyzed (Table 2) shows the most auto-correlated variables are: EC, Ca, Mg, Na, K, HCO_3, Cl, SO_4, SAR, with correlation coefficients (r) larger than 0.5. Several authors dispute the threshold for r. Silveira and Andrade (2003) are less conservative as they adopted a value of 0.3, while Helena et al. (2000) adopted a value of 0.5.

The high correlation found between calcium (Ca), magnesium (Mg), hydrogen-carbonate (HCO_3) and sulphate (SO_4) can be explained by the presence of limestone sediments associated with the geology of the catchment. On the other hand, the correlation between chloride (Cl) and sodium (Na) cannot be attributed to soil origin, since these minerals are not found in the local geology (COTEC, 1989). It is believed that higher concentration of these elements is due to anthropic influences, such as irrigation and sewage conveyed from urban areas. The strong correlation between electrical conductivity

(EC), chloride (Cl), sodium (Na), magnesium (Mg) and sulphate (SO_4) was previously expected because the electrical conductivity expresses salt concentration in water.

Pearson Correlation Matrix for water quality variables for Trussu Rive

Table 2

Variable	pH	EC	Ca	Mg	Na	K	HCO_3	PO_4	Cl	NH_4	NO_3	SO_4	SAR
pH	1.000												
EC	0.069	1.000											
Ca	-0.091	0.710	1.000										
Mg	0.017	0.850	0.647	1.000									
Na	0.099	0.960	0.670	0.814	1.000								
K	0.100	0.562	0.370	0.449	0.543	1.000							
HCO_3	-0.088	0.725	0.595	0.536	0.772	-0.476	1.000						
PO_4	0.257	0.006	0.080	-0.014	0.080	-0.051	0.048	1.000					
Cl	0.097	0.971	0.687	0.888	0.935	-0.511	0.637	0.032	1.000				
NH_4	0.119	-0.025	-0.265	-0.077	-0.016	0.046	0.030	-0.024	0.001	1.000			
NO_3	0.165	-0.024	-0.291	-0.081	-0.013	0.048	-0.002	0.083	0.002	0.877	1.000		
SO_4	0.070	0.841	0.496	0.735	0.816	-0.508	0.641	0.087	0.846	0.029	0.075	1.000	
SAR	0.117	0.932	0.603	0.720	0.982	-0.562	0.797	0.096	0.885	0.010	0.012	0.790	1.000

Total variance (eigenvalue) explained by each factor

Table 3

Component	Eigenvalue	Percentage of variance (%)	Cumulative percentage (%)
1	6,287	48,36	48,360
2	2,068	15,91	64,271
3	1,237	9,51	73,785
4	0,871	6,70	80,489
5	0,641	4,93	85,417
6	0,609	4,69	90,103
7	0,420	3,23	93,334
8	0,246	1,90	95,230
9	0,180	1,39	96,617
10	0,151	1,17	97,782
11	0,136	1,05	98,828
12	0,104	0,80	99,630
13	0,048	0,37	100,000

Eigenvalues for the set of variables are presented on Table 3. One can conclude that PC_1, PC_2 and $PC_{3,}$ with values of eigenvalues large than one, explain most of the variance of the data set. They explain, respectively, 48.36, 15.91 and 9.51% of the total variance of the data, i.e., three dimensions concentrate 73.78% of the information that is contained in thirteen variables. The results are consistent with reports by Mendiguchía et al. (2004), when, in a study of a seven-parameter model for the river Guadalquivir in Spain, have found 79.1% of the total variance concentrated into the first two components.

The values of communalities of the variables presented on Table 4 suggest that the common factors (principal components) explain more than 61% of the variance, exception to potassium. Projections of the original variables on the subspace of the principal components are called loadings and coincide with the correlation between the factors and variables. Loads of PC1, PC2 and PC3 are presented in Table 4. They can be used to identify variables that detain high correlation with each factor.

The first principal component is highly affected by sodium, EC, chloride, magnesium, sulphate and hydrogen-carbonate with loads higher than 0.8. It, basically, reflects ionic group of salts. The second component has strong positive loadings (> 0.92) on NO_3-N and NH_4-N and low positive loadings on Na, EC, Cl, Mg, HCO_3, K and SO_4. It represents the nutrients group of pollutants which points to some source of wastewater and runoff. The third one is mainly contributed by pH and PO_4 and therefore, may be related to agriculture activities.

One can also observe that the variables explained by the first and the second component present values of communalities above 0.9, while the variables explained by the third component presented communalities above 0.61, that is 61% of the variance of the variables PO_4 4 and pH is explained by the three first components of the model. This suggests the model applied is a better indicator of salt and organic naturally available contents than agricultural fertilizers effects on water quality.

Communalities and factor loadings for PC_1, PC_2 and PC_3

Table 4

Variable	Communalitie	Factor loading (λ)		
		1	2	3
Na	0,941	0,968	0,047	0,039
EC	0,877	0,936	0,008	-0,017
Cl	0,817	0,901	0,073	0,001
SO_4	0,763	0,865	0,123	0,011
Mg	0,732	0,852	-0,074	-0,040
HCO_3	0,668	0,808	0,037	-0,116
SAR	0,648	0,781	0,184	0,061
Ca	0,663	0,740	-0,339	0,033
K	0,398	-0,615	0,051	0,132
NO_3-N	0,917	-0,043	0,945	-0,148
NH_4-N	0,917	-0,042	0,922	-0,255
PO_4	0,618	0,077	0,149	0,768
pH	0,644	0,040	0,349	0,721

Varimax rotated factors for the first three principal components

Table 5

Variables	Communalitie	Factor loading (λ)		
		1	2	3
Na	0,941	0,965	-0,009	0,096
EC	0,877	0,936	-0,030	0,030
Cl	0,817	0,901	0,030	0,063
SO_4	0,763	0,866	0,077	0,085
Mg	0,732	0,850	-0,099	-0,017
HCO_3	0,668	0,815	0,029	-0,064
SAR	0,648	0,782	0,126	0,145
Ca	0,663	0,726	-0,369	-0,021
K	0,398	-0,614	0,042	0,112
NO_3-N	0,917	0,002	0,957	-0,007
NH_4-N	0,917	-0,005	0,952	0,103
PH	0,644	0,009	0,148	0,789
PO_4	0,618	0,036	-0,059	0,783

Dillon e Goldstein (1984) showed that, although the load factor matrix obtained in the extraction phase indicates the relationship between each variable, orthogonal rotation of the factors could produce more meaningful interpretations. This produces a set of orthogonal factors, which are independent and uncorrelated. In a study of multivariate structure of evapotranspiration, Andrade et al. (2003) applied the varimax rotation algorithms and obtained a matrix that was easier to interpret. However, as demonstrated on Table 5, this procedure resulted in little improvements for the water quality model of Trussu River for the first three components is still correlated with the same variables as before the varimax was applied.

4. CONCLUSIONS

The results of the present study point to the existence of three main independent types of contaminants determining water quality of the Trussu valley. PCA loadings indicate that variables responsible for water quality variations are mainly related to salts/mineralization processes (soil erosion/leaching followed by overland flow process), organic pollution (representing influences from municipal effluents) and nutrient variables (representing influences from non-point sources such as agricultural runoff) Orthogonal rotation on the component axis resulted in no aid in determining other meaningful variables related to water quality in Trussu valley.

Acknowledgement. The authors gratefully acknowledge the financial support by Comissão de Aperfeiçoamento dos Professores de Ensino Superior - CAPES and the Fundação Cearense de Apoio a Pesquisa – FUNCAP.

REFERENCES

Andrade, E. M.; Silveira, S. S., Azevedo B. M., (2003), *Investigação da estrutura multivariada da evapotranspiração na região centro sul do Ceará pela análise de componentes principais*. Revista Brasileira de Recursos Hídricos**,** v. 08, n. 01, p. 39-44.

Bordalo, A. A., Nilsumrachit, W.; Chalermwat, K., (2001), *Water quality and uses of the Bangpalong river (Eastern Thailand)*. Water Research, v. 35, n. 15, p. 3635–3642.

Brodnjak-Vocina, et al., (2002), *Chemometrics characterization of the quality of river water*. Analytica Chimica Acta, v. 462, p. 87–100.

Cotec - Consultoria Técnica Ltda. Barragem Trussu -*Estudo e avaliação dos impactos ambientais*. Tomo I:diagnóstico ambiental, Fortaleza, PRONI-DNOCS. 1989. 118p.

Dillon, W. R.; Goldstein, M., (1984), *Multivariate analysis methods and applications*. New York, John Wiley e Sons. 587p.

Elmi, A. A. et al. , (2004), *Water and fertilizer nitrogen management to minimize nitrate pollution from a cropped soil in southwestern Quebec*, Canada. Water, Air, and Soil Pollution. v. 151, p. 117-134.

Gangopadhyay, S.; Gupta, A.; Nachabe, M. H., (2001), *Evaluation of ground water monitoring network by principal component analysis*. Ground Water, v. 39, n. 2, p. 181-191.

Helena, B. et al. , (2000), *Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis*. Water Research v. 34, n. 03, p. 807-816.

Hotteling, H. , (1933), *Analysis of a complex of statistical variables into principal components*. Journal Education Psychology, v. 24, p. 498-520.

Lake, I.R. et al., (2003), *Evaluating factors influencing groundwater vulnerability to nitrate pollution: developing of G.I.S*. Journal of Environmental Management, v. 68, p. 315-328.

Mendiguchia, C. et al., (2004), *Using chemometric tools to assess anthropogenic effects in river water a case study: Guadalquivir river (Spain)*. Analytic Chemical Acta, v. 515, p. 143–149.

Monteiro, V. P.; Pinheiro, J. C. V. , (2004), *Critério para Implantação de Tecnologias de Suprimentos de Água Potável em Municípios Cearenses Afetados pelo Alto Teor de Sal*. Revista de Economia Rural**,** v. 42, n. 02, p. 365-387.

Silveira, S. S; Andrade, E. M., (2002), *Análise de componentes principais na investigação da estrutura multivariada da evapotranspiração*. Engenharia Agrícola, v. 22, n. 02, p. 171-177.

Simeonov, V. et al., (2003), Assessement of the surface water quality in northen Greece. Water Research, v. 37, p. 4119-4124.

Toledo, L. G.; Nicolella, G., (2002), *Índice de qualidade de água em microbacia sob uso agrícola e urbano*. Scientia Agrícola, v. 59, n. 01, p. 181-186.

Vega, M. et al., (1998), *Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis*. Water Research, v. 32, n.12, p. 3581-3592.