Ciência e Natura, Santa Maria v.38 n.2, 2016, Mai.- Ago. p. 577 – 587
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA

Avaliação espaço-temporal da qualidade da água utilizando técnicas estatísticas multivariadas

Spatial and temporal assessment of water quality data using multivariate statistical techniques

Nícolas Reinaldo Finkler, Taison Anderson Bortolin , Jardel Cocconi , Ludmilson Abritta Mendes e Vania Elisabete Schneider

[1]Instituto de Saneamento Ambiental (ISAM). Universidade de Caxias do Sul (UCS)
nicolas.finkler@gmail.com; tabortol@ucs.br; jardelcocconi@gmail.com; veschnei@ucs.br

[2]Universidade Federal de Sergipe (UFS)
lamendes@ufs.br

Resumo

Os fatores naturais e, em especial os antrópicos, que contribuem para a variação espacial e temporal da qualidade da água superficial nas bacias hidrográficas do município de Caxias do Sul foram determinados com uso de técnicas multivariadas de análise de dados. Foi utilizada a técnica de Análise do Componente Principal (ACP) e Análise de Agrupamentos (AA) como bases para o estudo. O monitoramento foi realizado em 12 pontos de monitoramento no período compreendido entre janeiro de 2009 a janeiro de 2010, totalizando 13 campanhas. Ao total, foram analisados no total, 20 parâmetros físicos, químicos e biológicos. Os resultados obtidos demonstram que, com o emprego da ACP, foi possível explicar uma variância total de 70,94% para os dados de qualidade de água. Ainda, constatou-se que dentre os principais fatores contribuintes para a variação da qualidade da água na região estão a poluição doméstica e industrial, sobretudo do setor galvonotécnico. Foi verificada, por fim, uma tendência à atenuação dos poluentes nos corpos hídricos à jusante das áreas urbanas e de grande influência antrópica. Especialmente na medida em que há menor pressão das áreas urbanizadas sobre as bacias que drenam a região.

***Palavras-chave**: Análise de Componentes Principais (ACP), Análise de Agrupamento (AA), Bacia Hidrográficas Urbanas.*

Abstract

The natural factors and anthropogenic activities that contribute to spatial and temporal variation in superficial waters in Caxias do Sul's urban hydrographic basins were determined applying multivariate analysis of data. The techniques used in this study were Principal Component Analysis (PCA) and Cluster Analysis (CA). Monitoring was conducted in 12 sampling stations, from January, 2009 to January, 2010 with monthly periodicity and for a total of 13 campaigns. Between chemical, biological and physical, 20 parameters were analyzed. The results state that with the use of PCA, a data variance of 70.94% was observed. Therefore, it testifies that the major pollutants that contribute to a water quality variation in the county are classified as domestic and industrial pollutants, mainly from galvanic industry. Two clusters were found to be differentiated regarding their location and distance from areas with a high human density, leading to further identification of impacts due to human activities in urban rivers. Furthermore, it was determined a tendency to self-attenuation of pollutants in the water bodies downstream of urban areas with great anthropic influences. Especially as far as there is less pressure of urban areas in the basins draining the region.

***Keywords**: Principal Component Analysis (PCA), Cluster Analysis (CA), Urban Hydrographic Basin.*

Recebido: 28/05/2015 Aceito: 03/08/2015

1 Introduction

Surface water bodies are characterized as the most vulnerable to pollution due to the easy accessibility towards launching of industrial and domestic wastewaters, mainly in urban basins. The natural processes such as rainfall, erosion and sediment loading; as well as anthropogenic processes like urbanization, industrialization and agriculture, contribute to the loss of water resources and define the water quality of a region (Singh et al. 2009).

According to Vega et al. (1998), rivers are the main available water resource for human consumption supply, irrigation and industrial purposes. Thus, it becomes fundamental to assort reliable and consistent water quality data for an effective and stringent management of water resources.

Commonly, the assessment of special variations in river water quality is given by the periodically measurement of multiple parameters at different monitoring stations (Fan et al. 2010). The results obtained through this technique create a complex matrix of difficult interpretation. As a result, generating streamlined results that are representative to the reality of the studied basin are being sought (Noori et al. 2010; Ouyang 2005).

The multivariate techniques such as Cluster Analysis (CA), Principal Component Analysis (PCA), Factor Analysis (FA) and Discriminant Analysis (DA) have been broadly used for further interpretation of the water quality data and identifying possible factors or sources that affect water bodies; also, they form a trustworthy tool for proper water resource management (Vega et al. 1998; Singh et al. 2009; Rodrigues et al. 2009). These statistical tools allow reducing the number of variables to a small number of indexes ie, principal components or factors, seeking to preserve relationships in the original data (Ouyang 2005).

Studies related to the techniques above mentioned were conducted by Coletti et al. (2010), Fan et al. (2010), Hussain et al. (2008), Kazama and Shrestha (2007), Krishna et al. (2009), Mustonen et al.(2008), Noori et al. (2010), Pinto and Maheshwari (2011), Wang et al. (2012) and Singh et al. (2004); whom used multivariate analysis in several ecosystems.

The study analyses quality parameters and the main pollution sources responsible for temporal and spatial variation in the quality of rivers. Such water bodies originate in the urban region Caxias do Sul county, Rio Grande do Sul - Brazil.

2 Material And Methods

2.1 Study Area

The county of Caxias do Sul, situated in the northeast portion of Rio Grande do Sul, southern Brazilian state, holds the third highest Gross Domestic Product of the state (IBGE 2013). The economy of the municipality relies primarily on its diverse industrial park, with emphasis on the metal mechanical sector.

The city is located on the watershed splitter of Taquari-Antas and Caí river basins, both contributors of the Guaíba Hydrographic Region, as shown in Figure 1. The northern portion of the municipality falls under the Taquari-Antas Basin and represents 65% of the county's territory.

Caxias do Sul is drained by the microbasins Tega and São Marcos. The first one holds 90% of its area within the city's urban perimeter. As for the southern part, it is inserted in the Caí River Hydrographic Basin and represents 35% of the municipal territory. Under this system, there are three microbasins - Belo, Pinhal and D'Ouro - and the Piaí River Basin.

The sectioning of the sampling network, formed by 12 sampling points, sought to involve the main city urban basins. Criteria used for sampling point mapping were land use and occupation in the surroundings and the drainage area for each waterway.

The sampling network is portrayed in Figure 2; where the circled area represents the zones of greater population and industrial density. The sub basins to the monitoring points and the respective GPS coordinates are shown in Table 1.

2.2 Analytical procedures

To assure evaluation of the water resources quality, monthly samples were collected during the period January 2009 and January 2010, totalizing 13 campaigns. Twenty water quality parameters were analyzed; its methodologies are presented in Table 2.

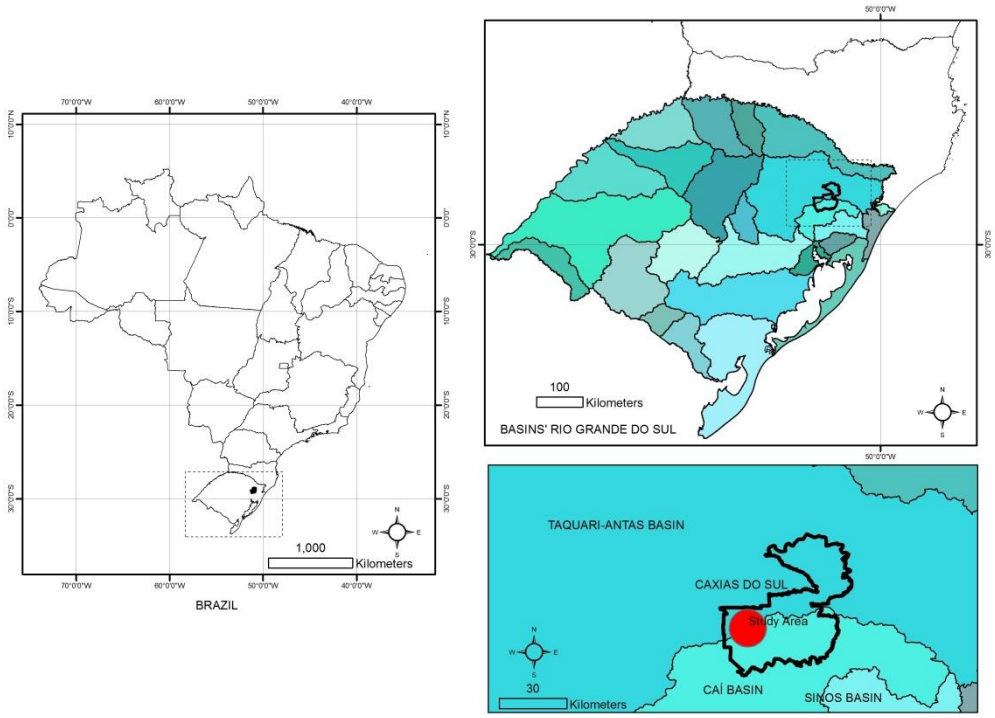

Figure 1: Localization of the Caxias do Sul county and the study area.

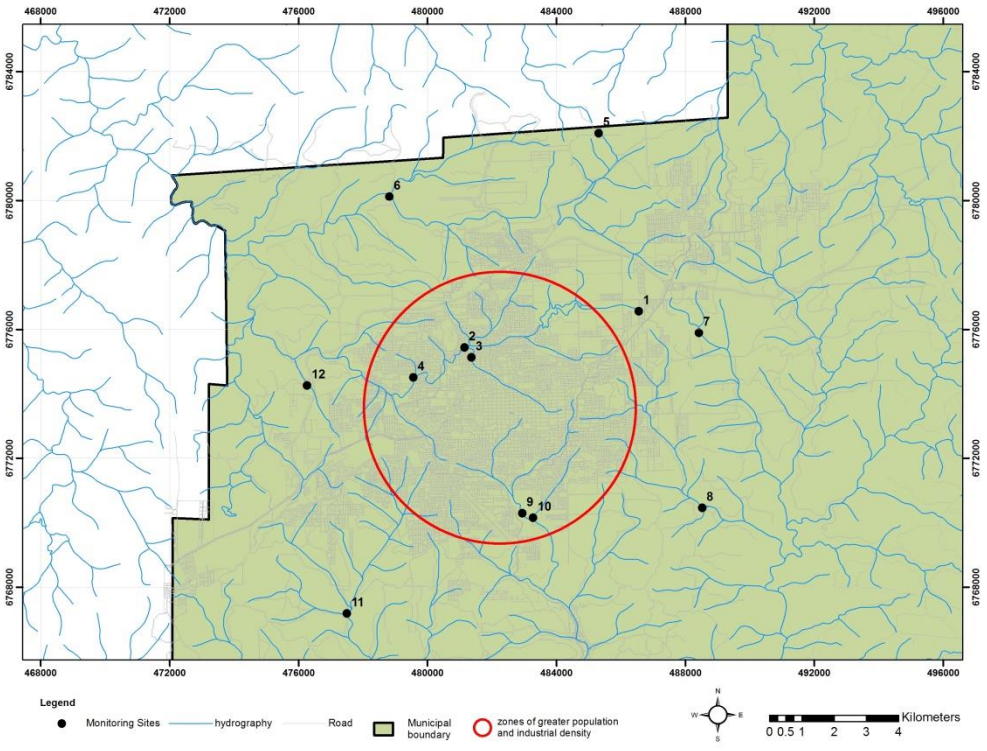

Figure 2: Monitoring points localization.

Table 1: Description of the monitoring points.

Point	Site	Sub basin/ River	UTM Coordinates (Datum: SA69 Area: 22s) N	E
1	Nascente Dal Bó	Dal Bó	486553	6776570
2	Ponte São José	Tega	481148	6775443
3	Santa Catarina	Tega	481361	6775140
4	Moinho	Tega	479557	6774509
5	Nascente Maestra	Maestra	485304	6782095
6	Ponte linha 40	Maestra	478817	6780127
7	Arroio Espelho	Espelho	488415	6775897
8	Ponte Pena branca	Pena Branca	488524	6770461
9	Arroio Pinhal	Pinhal	482940	6770296
10	Planalto	Pinhal	483276	6770161
11	Desvio Rizzo	Arroio Belo	477496	6767182
12	Dist. Industrial	Tega	476263	6774262

Table 2: Methodologies applied in determining parameters.

Parameter	Unit	Methodology/ Equipment	Detection Limit
QOD	mg $O_2.L^{-1}$	Open Reflux with K_2Cr_2O7 acidic	5
BOD	mg $O_2.L^{-1}$	Dilution an Incubation at 20°C for 5 days	1
TKN	mg $N.L^{-1}$	Titrimetric with com Nesslerization	5
NH_3	mg $NH_3.L^{-1}$	Titrimetric with com Nesslerization	2
P	mg $P.L^{-1}$	Colorimetric of Ascorbic Acid	1
AS	$mg.L^{-1}$	Methylene Blue – MBAS	25
OG	$mg.L^{-1}$	Sohxlet Extraction/Gravimetric	10
TS	$mg.L^{-1}$	Gravimetric at 103 – 105°C	10
CN	$mg.L^{-1}$	Spectrometry	1
Cr	$mg.L^{-1}$	Atomic Absorption	4
Zn	$mg.L^{-1}$	Atomic Absorption	1
Fe	$mg.L^{-1}$	Atomic Absorption	4
Al	$mg.L^{-1}$	Atomic Absorption	1
Ni	$mg.L^{-1}$	Inductively Coupled Plasma	1
Pb	$mg.L^{-1}$	Inductively Coupled Plasma	4
TC	$NMP.100mL^{-1}$	Multiple Tubes	18
pH	-	Potentiometric Method	-
EC	$\mu s.cm^{-1}$	Electrometry	-
DO	mg $O_2.L^{-1}$	Membrane Electrode	-
TH_2O	°C	Thermometer	-

The parameters analyzed in laboratory were: Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), Total Kjeldahl Nitrogen (TKN), Ammonia (NH3), Total Phosphorus (P), Anionic Surfactants (AS), Oils and Greases (OG), Total Solids (TS), Cyanide (CN), Chromium (Cr), Zinc (Zn), Iron (Fe), Aluminum (Al), Nickel (Ni), Lead (Pb), Thermotolerant Coliforms (TC).

As for the field parameters, one can list: pH, Water Temperature (TH2O), Dissolved Oxygen (DO) and Electrical Conductivity (EC).

During the sampling period, over 5,000 data points were obtained, which justifies the importance of using multivariate techniques in data analysis.

2.3. Statistical Analyzes

The statistical techniques used in the study were the Principal Component Analysis, as a plea for the discovery of a group of parameters that significantly interfere with the quality of water; and Cluster Analysis, in order to verify the formation of sampling point groups with similar water quality. With the purpose of testing whether the data set is sufficiently connected to proceed with the Principal Component Analysis, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) was used. To verify the possibility of rejection of the null-non correlation between variables hypothesis - which would preclude the application of the PCA - it was used the Barlett sphericity test.

2.3.1. Descriptive Statistics

Prior to the application of multivariate techniques, some analyzes were conducted in order to characterize the samples. Once the comparison between standard deviations of different magnitudes and variables proves to be complex, it was decided to calculate the variation coefficient which is equal to the standard deviation divided by the average. Thus, it is possible to compare the variation of the sets of observation that differ in average or that are expressed in different measuring units. And then, sort the variables extent of dispersion (França 2009). The criteria used during the evaluation of the variation coefficient were: values lower than 0.700 present low level of dispersion; values between 0.700 and 1.750 have average degree of dispersion; values greater than 1.750 feature high level of dispersion.

2.3.2. Principal Component Analysis (PCA)

The PCA can be defined as a multivariate mathematical technique that transforms the data to a new coordinate system. This new set of variables - the principal components (PCs), are linear functions of the original variables, uncorrelated, and the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on (Jolliffe 2002).

Mathematically, the PCA involves five main steps: (i) use of standard values for mean zero and variance 1, to ensure that they possess equal significance in the analysis, according to Equation 1; (ii) calculations of the correlation matrix R and covariance matrix Σ; (iii) determination of the eigenvectors $\lambda 1$, $\lambda 2$,…, λp and their corresponding eigenvalues a1, a2,..., ap through Equation 2; (iv) discarding of components that are part of a small proportion in the data variance; and (v) making of the factor loadings matrix with or without application of rotational methods of variables to the matrix, in order to reduce the number of PCs.

$$z_{ij} = \frac{x_{iv} - \overline{x_v}}{s(x_v)}, as\ i = 1,2 \dots, n;\ \text{and}\ v = 1,2, \dots, p \quad (1)$$

$$|R - I\lambda|\ or\ det[R - I\lambda] = 0 \quad (2)$$

Where: x_{iv}= characteristic of the element i; $\overline{x_v}$ = average value of the element v; λ = eigenvectors; ν = level of freedom; $|R|$ = correlation matrix determinant.

The PCs selection criterion is to include only those components whose eigenvalues are greater than 1. Such criteria, suggested by Kaiser (1960) apud Mardia et al. (1979), also tends to include fewer components when the number of original parameters is less than twenty. In general, components that hold a cumulative variance of around 70% of the total variance are used.

Furthermore the rotation of the variables in order to present results with more consistency was perfomed. For such, the VARIMAX rotation method, developed by Kaiser (1958), was used. It is a method commonly used in multivariate statistical studies since it makes interpretation simpler by reducing the number of correlations between variables (Abdi, 2003).

The application of PCA depends on two tests that must be performed in advance, in order to confirm if the technique fits to the available data: KMO test and Bartlett's test.

KMO tests ensure the adequacy of the data to the use of PCA by checking the correlation matrix as a whole, or simply the correlation between the independent variables. Ferreira Jr. et al. (2004) reported that the KMO test is an identifier that compares the magnitude of the observed correlation coefficient to the magnitude of the partial correlation coefficient. The KMO test values range between 0 and 1. It is considered that values bellow 0.5 indicate that data are suitable for PCA application.

Some authors such as Rencher (2002) suggest that for a PCA model to be properly fit to the data, it is necessary that the inverse correlation matrix Rpxp-1 remains next to the diagonal matrix. The measure of KMO sample adequacy is represented by the MSA index, calculated by Equation 3.

$$\text{MAS} = \frac{\sum_{j \neq k} \sum_{j \neq k} {r_{jk}}^2}{\sum_{j \neq k} \sum_{j \neq k} {r_{jk}}^2 + \sum_{j \neq k} \sum_{j \neq k} {q_{jk}}^2} \quad (3)$$

Where: rjk^2 is the square of the original correlation matrix elements (off-diagonal); qjk^2 is the square of the off-diagonal elements of the anti-image matrix (where qjk is the partial correlation coefficient between the variables Xj and Xk).

The Bartlett's test of sphericity also checks the adequacy of the data for the PCA application; testing whether the correlation matrix is an identity matrix, which would indicate no correlation between the data. Thus, when

seeking for a significance level of 5%, the null hypothesis of identity correlation matrix is rejected. The basic hypothesis says that the population correlation matrix is an identity matrix, which indicates that the factorial model is inappropriate. Test statistics is given from Equation 4:

$$x^2 = -\left[(n-1) - \frac{2p+5}{6}\right] * \ln|R| \quad (4)$$

The statistic has chi-square distribution (x^2), with level of freedom (ν) given by Equation 5:

$$\nu = \frac{p(p-1)}{2} \quad (5)$$

Where: n expresses the size of the sample, p is the number of variables; and $|R|$ indicates the correlation matrix determinant.

2.4.2 Cluster Analysis (CA)

The Cluster Analysis (CA) aims to grouping or segmenting a collection of elements into "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. An element can be described by a set of measurements, or by its relation to other ones. Sometimes, the CA aims to arrange the clusters into a natural hierarchy or groups of heterogeneous characteristics (Hastie *et al* 2011).

As for cluster elements, it is necessary to define a distance measure for similarity or dissimilarity. Albuquerque (2005) cites that the most commonly used types of distances in cluster analysis are: Euclidean Distance; Squared Euclidean Distance; Weighted Euclidean Distance; and Mahalanobis Distance.

The Euclidean Distance was used as the similarity measure in the present study. The premise of which that the distance between two cases i and j is the square root of the sum of squares of differences between values i and j for all variables v (v = 1, 2,..., p). Equation 6 expresses the calculation used to find the distance.

$$d_{ij} = \sqrt{\sum_{v=1}^{p}(x_{iv} - x_{jv})^2} \quad (6)$$

Where: x_{iv} represents the characteristic of the element i; p is the number of plots in the sample; and v is the element number in the sample.

The cluster analysis was applied to the averages of the water quality parameters, in order to identify similarities between sampling points. A similar procedure was performed by Kazama and Shrestha (2007) in the Fuji River Basin, in Japan; which identified the importance of the obtained results at various points for the local water resource management.

The statistical and mathematical calculations were performed using the software EXCEL 2010. The multivariate analyzes were performed with the SPSS STATISTICS 18 software and the dendogram was produced from the software STATISTICA 8.0.

3 Results and discussion

3.1 Descriptive Statistics

The descriptive statistics of the sample are shown in Table 3.

Table 3: Descriptive statistics from sample.

Parameter	Max.	Average	Min.	SD[1]	VC[2]
Al	24.04	0.826	0.10	2.546	3.084
Pb	0.067	0.008	0.004	0.006	0.746
CN	0.033	0.022	0.002	0.053	2.409
TC	7,000,000	358,434	0.01	902,159	2.517
EC	767	276.87	22	167.515	0.605
Cr	3.43	0.157	0.004	0.395	2.515
BOD	268	21.07	1	33.230	1.577
QOD	628	53.69	5	74.068	1.379
Fe	20	1.33	0.06	2.107	1.585
Ni	1.04	0.079	0.001	0.185	2.347
NH3	44.79	7.969	0.02	7.961	0.999
NTK	45.01	10.012	0.5	9.212	0.920
OD	13.95	7.176	1	2.926	0.408
OGT	61.60	10.437	10	5.955	0.571
pH	8.73	7.226	4.71	0.604	0.084
ST	4089	239.552	37	392.845	1.640
AS	13.8	0.777	0.025	1.354	1.743
TH20	25.05	17.436	10.5	2.980	0.171
Zn	3.38	0.231	0.01	0.457	1.976
P	6.83	1.06	0.011	1.069	1.014

[1]Standart Deviation; [2]Variation Coefficient.

It is possible to identify that the parameters Aluminum, Cyanide, Total Coliforms, Chromium, Nickel and Zinc showed a high degree of dispersion arising from the anthropic activity in the region. Such dispersion is permitted, given that the collection points are

situated in urban areas with different types of land use and occupation.

As presented, the analysis of PCA is based on the diagonalization of the correlation matrix. Further analisis of the matrix can indicate associations between parameters, consistency of the data set and evidences the participation of the individual parameter in several influence factors. For such reason, significant correlations (at a p=0.05 level) are usually searched for. However, due to the high number of degrees of freedom (df=n-2; df=174), in this study, $r_{critical}$ is low (<0.195 at p = 0.05), so the number of statistically significant correlations (with r > $r_{critical}$) is very high. The real usefulness of this test is questionable since it simply proves that r is significantly different from zero. As uniquely really stronger correlations will be useful, only those with r values higher than |0.500| have been considered (Guedes et al 2012; Toledo and Niconella 2002; França 2009; Helena et al 2000), which are highlighted in the Table 4.

Table 4: Matrix of Correlations between the Parameters (n=176).

	Al	Pb	CN	TC	EC	Cr	BOD	QOD	Fe	P	Ni	NH3	TKN	DO	OG	pH	TS	AS	TH20	Zn
Al	1.000																			
Pb	0.253	1.000																		
CN	0.011	0.081	1.000																	
TC	0.107	0.089	0.184	1.000																
EC	0.053	0.193	0.160	0.467	1.000															
Cr	-0.007	0.157	0.253	0.105	0.239	1.000														
BOD	0.168	0.127	0.108	**0.510**	**0.617**	0.115	1.000													
QOD	0.459	0.178	0.084	0.438	**0.571**	0.138	**0.836**	1.000												
Fe	**0.854**	0.216	0.033	0.098	0.052	0.072	0.164	0.452	1.000											
P	0.199	0.151	0.119	**0.563**	**0.765**	0.164	**0.766**	**0.736**	0.187	1.000										
Ni	-0.023	0.279	0.294	0.000	0.244	0.349	0.158	0.120	-0.028	0.144	1.000									
NH3	0.034	0.129	0.099	0.430	**0.860**	0.123	**0.580**	**0.544**	0.028	**0.775**	0.127	1.000								
TKN	0.082	0.150	0.125	0.488	**0.891**	0.159	**0.655**	**0.625**	0.072	**0.805**	0.149	**0.973**	1.000							
OD	-0.022	-0.188	-0.177	-0.307	-0.333	-0.227	-0.330	-0.306	-0.185	-0.268	-0.164	-0.188	-0.231	1.000						
OG	0.198	0.380	0.001	0.069	0.199	0.035	0.317	0.404	0.196	0.258	0.081	0.221	0.237	0.028	1.000					
pH	0.064	-0.020	-0.036	0.130	0.444	-0.021	0.213	0.247	-0.060	0.358	0.036	0.431	0.432	0.145	0.034	1.000				
TS	0.157	0.055	0.003	0.082	0.230	0.014	0.166	0.210	0.131	0.132	0.000	0.198	0.237	-0.072	0.047	0.166	1.000			
AS	**0.506**	0.219	0.149	0.257	0.423	0.153	0.459	0.389	0.425	0.480	0.132	0.314	0.382	-0.296	0.122	0.101	0.124	1.000		
TH20	-0.028	-0.014	-0.023	0.025	0.316	0.080	0.143	0.147	-0.123	0.193	0.184	0.230	0.232	-0.204	-0.245	0.238	0.070	0.083	1.000	
Zn	0.069	0.183	0.105	0.078	0.276	**0.692**	0.133	0.208	0.249	0.176	0.243	0.164	0.194	-0.227	0.096	-0.036	0.060	0.159	0.109	1.000

The parameters that show a greater number of possible correlations are Phosphorus and BOD, with six at total. Conductivity, QOD, Ammonia and Total Kjeldahl present five correlations. And finally, Aluminum and Thermotolerant Coliforms present two correlations; Chromium, Iron, Anionic Surfactants and Zinc relate to one correlation only. Ammonia and Total Nitrogen showed high correlation with Phosphorus, 0.775 and 0.805 respectively, and also with indicators of organic matter. Conductivity showed high correlations with Nitrogen, Phosphorus and Organic Matter, indicating that it is possible to use conductivity to infer the presence of nutrients and organic matter in water.

3.2. Principal Component Analysis

The PCA has revealed 6 Principal Components (PC), and the total variance clarified reached 70.94% for the water quality data.

Liu et al. (2003) classifies the factor loadings corresponding to the absolute composition of PCs in values. The author considers values greater than 0.75 as relevant; in between 0.50 and 0.75 as of medium relevance; and irrelevant for values ranging between 0.30 and 0.50. Distribution of the factor loadings and correlations of PCs with the rotated variables are given in Table 5. The highlighted values in bold denote strong and medium correlation with the respective PC.

The PC1 load components analysis - formed by the parameters TC, EC, BOD, QOD, P, NH3 and TKN - suggests that the water quality variation in the county arises mostly from wastewater discharges lacking proper treatment. Electrical Conductivity has high correlation with the aforementioned parameters.

Table 5: Factor loadings with variable rotation for the PCs found.

Parameters	Principal Components (PC)					
	1	2	3	4	5	6
TC	**0.648**	0.056	-0.013	-0.125	-0.040	-0.404
EC	**0.868**	0.014	0.174	0.205	0.172	0.094
BOD	**0.816**	0.194	0.007	0.050	-0.090	-0.169
QOD	**0.733**	0.452	0.081	0.035	-0.135	-0.032
P	**0.896**	0.165	0.056	0.061	-0.001	-0.033
NH_3	**0.890**	-0.049	0.074	0.095	0.066	0.183
TKN	**0.924**	0.014	0.095	0.104	0.066	0.139
Al	0.046	**0.939**	-0.045	0.050	-0.065	0.085
Fe	0.031	**0.911**	0.140	-0.033	-0.131	-0.050
AS	0.387	**0.586**	0.018	0.183	0.131	-0.176
Cr	0.093	-0.016	**0.865**	0.205	0.039	-0.127
Zn	0.114	0.126	**0.916**	0.075	-0.005	-0.011
Pb	0.084	0.254	0.069	**0.661**	-0.297	0.037
Ni	0.073	-0.062	0.219	**0.779**	0.153	-0.068
OG	0.286	0.142	0.025	0.290	**-0.740**	0.189
TH_2O	0.199	-0.003	0.054	0.174	**0.743**	0.142
DO	-0.296	-0.173	-0.187	-0.124	-0.239	**0.601**
pH	0.439	-0.024	-0.094	0.067	0.259	**0.593**
CN	0.109	-0.037	0.044	0.455	0.081	-0.458
TS	0.201	0.251	0.066	-0.025	0.172	0.302
Value	6.187	2.36	2.023	1.327	1.251	1.04
%variance	30.936	11.798	10.114	6.634	6.256	5.201
%cumulative	30.936	42.734	52.849	59.483	65.739	70.94

Rotation method: VARIMAX with Kaiser Normaliz**ation**.

PC2 shows the major local soil constituents Fe and Al, with significant variation, indicating that natural processes also contribute to water quality variations. Simeonov et al. (2003) found a similar component to the one observed in this study, it can be considered an indicative of physicochemical variation of the monitored waters. Such PC suggests that weathering of rocks that gave rise to the soil and subsequent surface runoff appear as a common source for these components.

Components PC3 and PC4 are given with greater values for metals used in the galvanic industry (Cr, Zn, Pb, Ni). Such activities are widely applied by local industries, that add metallic compounds in its production process, especially in the coating of metallic structures.

The parameters forming the components PC5 and PC6 show the effect of anthropic activities over the fundamental aquatic environment characteristics to maintain the biota.

3.2.2 Cluster Analysis

The cluster analysis sought to identify which points had similar water quality characteristics during the monitored period. These groups were formed from the average values of the parameters. Figure 3 shows the dendogram for groups formed by cluster analysis, and the graph of amalgamation schedule, that suggests a cutoff for the tree diagram.

According to the dendogram, two distinct clusters can be identified among the sampling points. Group 1 is formed by points 1, 5, 6, 7, 11 and 12; while Group 2 clearly identifies points 2, 3, 4, 9 and 10 as a different cluster.

When comparing the CA results with the mapping presented on Figure 2, it is possible to infer that the clusters were produced according to the geographic location of the sampling points. Points from Group 1 are situated on peripheral areas and are less influenced by anthropic contribution.

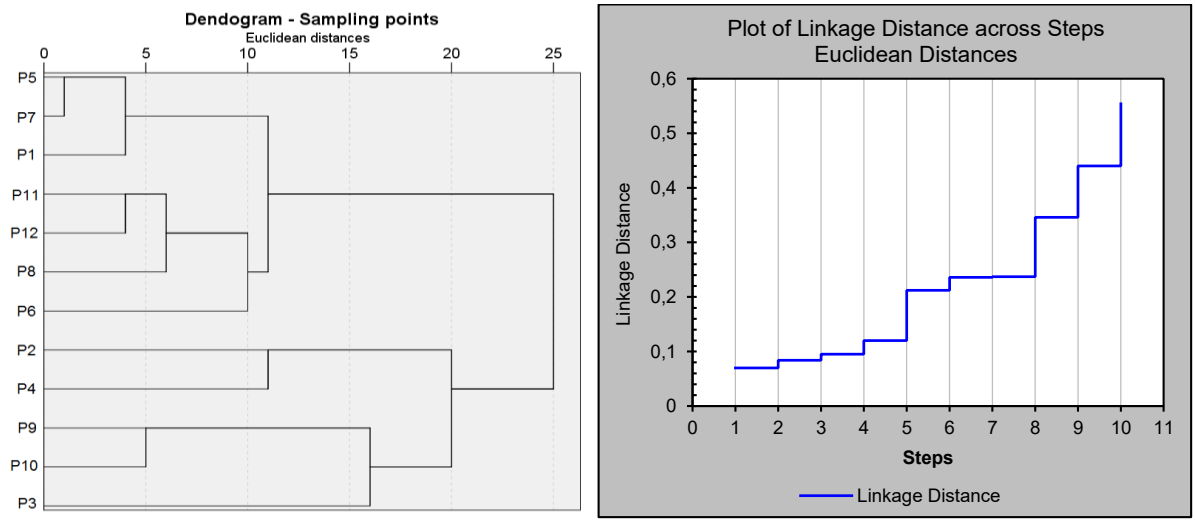

Figure 3: Cluster Analysis results and Amalgamation Schedule.

Points from Group 1 present lower concentrations of metallic compounds, nutrients and organic matter. As for points from Group 2, they are located on strictly urban areas, greater populated and with industrial contribution on the surroundings. Due to the before mentioned, these points are more vulnerable to contamination by wastewater discharge.

As for the amalgamation schedule inferences, it is possible to conclude that as we move further to the right (increase on the linkage distances), larger and larger clusters are formed of greater and greater within-cluster diversity, among Group 2 points. It means that many clusters were formed at essentially the same linkage distance. That distance may be the optimal cut-off when deciding how many clusters to retain (and interpret).

4 Conclusions

The multivariate techniques used on data analysis proved to be a relevant tool in the adjustment of water quality data. These techniques provide a distinguished investigation for decision making in the municipal water resources management.

The PCA allowed to identify the main parameters that interfere on water quality variation for the municipality and proof that such are related to anthropic activities in the region. The organic matter and nutrient indicators figure as the components with higher variance for the water quality data. Even though parameters associated with natural soil components for the region were identified, their interference on water quality variation is less significant than from those related to non-treated effluent discharges.

While grouping the monitoring points based on their similarity to water quality levels, two clusters were found to be different due to their localization and distance from the area with higher anthropic presence.

The results obtained showed to be consistent to the county reality. Although some wastewater treatment plants were implanted on microbasins, only a small fraction of the generated wastewater is reaches the treatment plants. At the same time, the discharge of industrial wastewater without proper treatment contributes to the water quality levels observed in the study.

River water quality monitoring is an extremely useful instrument to the understanding of water dynamics and its relation to the urban environment. Data analysis, as presented in this study, allows the public administrations to identify the necessity for structural and nonstructural measurements on environmental sanitation. Such, need to lead to the improvement of the urban rivers quality, mainly when it comes to irregular domestic effluent discharges and to better inspection of wastewater treatment systems.

Acknowledgements

The authors thank to Prefeitura Municipal de Caxias do Sul for funding the research.

References

Abdi, H. (2003). Factor Rotations in Factor Analyses. Encyclopedia of Social Sciences, In: Lewis-Beck M., Bryman, A., Futing T. (Eds.), Research Methods. The University of Texas at Dallas. Thousand Oaks.

Albuquerque, M.A. (2005). Estabilidade em análise de agrupamento (cluster analysis). Biometry Masters Thesis. Recife: Universidade Federal Rural de Pernambuco.

Coletti, C.; Testezlafl, R.; Ribeiro, T.A.P.; Souza, R.T.G. de; Pereira, D. de A. (2009). Water quality index using multivariate factorial analysis. Revista Brasileira de Engenharia Agrícola e Ambiental, 14, 517-522.

Fan, X.; Cui, B.; Zhao, H.; Zhang, Z.; Zhang, H. (2010). Assessment of river water quality in Pearl River Delta using multivariate statistical techniques. Procedia Environmental Sciences. 2, 1220-1234.

Ferreira Jr., S.; Baptista, A.J.M.S.; Lima, J.E. (2004). A Modernização Agropecuária nas Microrregiões do Estado de Minas Gerais. RER. 42, 3-89.

França, M. S. (2009). Análise estatística multivariada dos dados de monitoramento de qualidade de água da Bacia do Alto Iguaçu:

uma ferramenta para a gestão de recursos hídricos. Water Resources and Environmental Engineering Masters Thesis. Curitiba: Universidade Federal do Paraná.

Guedes, H. A. S.; Silva, D. D. da; Elesbon, A.; Ribeiro, C. B. M.; Matos, A. T. de; Soares, J. H. P. (2012). Aplicação da análise estatística multivariada no estudo da qualidade da água do Rio Pomba, MG. Revista Brasileira de Engenharia Agrícola e Ambiental. 16, 558-563.

Hastie, T.; Tibshirani, R.; Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag New York. 739 p.

Helena, B.; Pardo, R.; Vega, M.; Barrado, E.; Fernandez, J. M.; Fernandez, L. (2000). Temporal evolution of groundwater composition in na aluvial aquifer (Pisuerga River, Spain) by Principal Component Analysis. Water Research. 34, 807-816.

Hussain, M.; Ahmedb, S. M.; Abderrahmanc W. (2008). Cluster analysis and quality assessment of logged water at an irrigation project, eastern Saudi Arabia. Journal of Environmental Management. 86, 297-307.

IBGE. (2013). Instituto Brasileiro de Geografia e Estatística. Produto Interno Bruto dos Municípios 2010. Cidades@.<http://www.ibge.gov.br/cidadesat/comparamun/compara.php?coduf=43&idtema=103&codv=v05&order=dado&dir=desc&lista=uf&custom=>. Accessed on 6 sep. 2014.

Jolliffe, I. T. (2002). Principal Component Analysis. Second ed. Springer Series in Statistics. New York: Springer-Verlag New York. 487 p.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187-200.

Kazama, F.; Shrestha, S. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environmental Modelling & Software. 22, 464-475.

Krishna, A.K.; Satyanarayanan, M.; Govil, P.K. (2009). Assessment of heavy metal pollution in water using multivariate statistical techniques in an industrial area: A case study from Patancheru, Medak District, Andhra Pradesh, India. Journal of Hazardous Materials. 167, 366-373.

Liu, S.; Manson, J.E.; Stampfer, M.J.; Hu, F.B.; Giovannucci, E.; Colditz, G.A.; Hennekens, C,H.; Willett, W.C. (2003). A prospective study of whole-grain intake and risk of type 2 diabetes mellitus in US women. Am J Public Health. 90,
1409-1415.

Mardia, K. V.; Kent, J. T.; Bibby, J. (1979). Multivariate analysis, ed. Academic, London.

Mustonen, S.M.; Tissari, S.; Huikko, L.; Kolehmainen, M.; Lehtolab, M.J.; Hirvonen, A. (2008). Evaluating online data of water quality changes in a pilot drinking water distribution system with multivariate data exploration methods. Water Research. 42, 2421-2430.

Noori, R.; Sabahi, M.S.; Karbassi, A.R.; Baghvand, A.; Zadeh, H.T. (2010). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. Desalination. 260, 129-136.

Ouyang, Y. (2005). Evaluation of river quality monitoring stations by principal component analysis. Water Research. 39, 2621-2635.

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.

Rencher, A. (2002). Methods of multivariate analysis, second ed John Wiley & Son, New York.

Rodrigues, P.M.S.M.; Rodrigues, R.M.M.; Costa, B.H.F.; Martins, A.L.T.; Silva, J.C.G.E. (2010). Multivariate analysis of the water quality variation in the Serra da Estrela (Portugal) Natural Park as a consequence of road deicing with salt. Chemometrics and Intelligent Laboratory Systems. 102, 130-135.

Simeonov, V.; Stratis, J.A.; Samara, C.; Zachariadis, G.; Voutsa, D.; Anthemidis, A.; Sofoniou, M.; Kouimtzis, T. (2003). Assessment of the surface water quality in northen Greece. Water Research. 37, 4119-4124.

Singh, K. P.; Basant, A.; Malik, A.; Jain, G. (2009). Artificial neural network modeling of the river water quality – a case study. Ecological Modelling. 220, 888-895.

Singh, K.P.; Malik, A.; Mohan, D.; Sinha, S. (2004). Multivariate statistical techniques for

Toledo, L.G. de; Nicolella, G. (2002). Índice de qualidade de água em microbacia sob uso agrícola e urbano. Scientia Agricola. 59, 181-186.

Varol, M., Gökot, B., Bekleyen, A., Sen, B. (2011). Water quality assessment and apportionment of pollution sources of Tigris River (Tukey) using multivariate statistical technique – A case study. River Research and Applications. 27, 1553-1564.

Vega, M.; Pardo, R.; Barrado, E.; Deban, L. (1998). Assessement of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Research. 32, 3581-3592.

Wang, X.; Cai, Q.; Ye, L.; Qu, X. (2012). Evaluation of spatial and temporal variation in stream water quality by multivariate