OTIMIZAÇÃO DE PONTOS AMOSTRAIS DE QUALIDADE DE ÁGUA EM RESERVATÓRIOS

Alan Pinheiro de Souza[1]; Francisco Jácome Sarmento[2]

Universidade Federal da Paraíba - alanpinheirolp@gmail.com[1]; jacomesarmento@hotmail.com[2].

Resumo: O presente artigo propõe e aplica uma metodologia de determinação do número mínimo de pontos amostrais de coleta de água em um reservatório artificial (açude), de maneira a garantir representatividade adequada dos parâmetros associados aos pontos de coleta em relação ao corpo hídrico como um todo. A análise estatística multivariada, que tem por objetivo a maximização da variância explicada, é utilizada para determinar a componente principal que mais influencia na carga total de fósforo aportada ao reservatório. O estudo foi realizado no reservatório Araras, no município de Varjota - CE, onde dispõe-se de dados em sete pontos de amostragem, a partir dos quais foi gerada uma matriz de correlação. A matriz de correlação foi normatizada e dela foram extraídos seus autovalores (matriz R) e autovetores (matriz L), de modo a se construir a matriz de carregamento (matriz L). Foram obtidos resultados que demonstram que, com apenas quatro pontos amostrais pode-se obter uma representatividade de 98% para dados de fósforo, variável essa que é de primordial importância para a determinação do índice de estado trófico de reservatórios. Com tal racionalização da amostragem, os custos para análise e para a própria coleta podem ser reduzidos.

Palavras-chave: Análise Multivariada; Análise de Componentes Principais; Otimização; Qualidade de água.

Introdução

A identificação dos fatores principais (componentes principais) que contribuem para a carga de fósforo aportada em um reservatório é de primordial importância para a identificação do nível de eutrofização do mesmo. Observando as fontes geradoras de poluição, é fácil confirmar que a piscicultura, agricultura entres outras atividades humanas possuem uma grande influência na contaminação dos corpos hídricos, de modo a alterar a qualidade das águas e tornando-as inadequada para uso humano, por exemplo.

O monitoramento dos fatores que influenciam essa qualidade é de extrema importância para o planejamento da gestão desses recursos, porém esse monitoramento tende a ser oneroso, pois em algumas situações são feitas muitas amostras e análises, para obter resultados que poderiam ser da mesma forma obtidos com uma menor quantidade de pontos amostrais, otimizando assim o monitoramento e minorando os custos.

Este estudo tem como objetivo principal a identificação e também a quantificação do número mínimo de amostras que devem ser coletadas e analisadas em um determinado reservatório para que os resultados sejam significativos para o corpo hídrico como um todo. Para este estudo, o reservatório selecionado foi o Araras, no município de Varjota -CE. A metodologia utilizada é

baseada em Análise Multivariada e tem como ideia central a maximização da variância explicada, variância essa expressa na matriz de correlações afetas ao conjunto de dados originais. O uso dessa técnica na área de Recursos Hídricos é bastante consolidada e relativamente antiga. Miríades de aplicações podem ser encontradas. Alguns exemplos de referências são MATALAS e REIHER (1967), HARMAN (1976), DILLON e GOLDSTEIN (1984).

Metodologia

A representatividade dos dados geralmente disponíveis nas instituições responsáveis pelo monitoramento qualitativo das águas é aqui pressuposta. Via de regra, são acervos que as próprias instituições utilizam na sua lida cotidiana com os problemas qualitativos que enfrentam nos reservatórios pelos quais são responsáveis. É evidente que, se no futuro houver um adensamento dos pontos de coleta nos lagos monitorados por essas instituições, os resultados produzidos pela metodologia que se valem desses dados ganharão em precisão, na mesma proporção em que o adensamento da amostragem agrega em termos de representatividade.

Fato é que, nenhuma metodologia, em nenhuma parte do mundo, é capaz de contornar eventuais deficiências presentes nos dados dos quais ela se alimenta para fornecer os resultados. Isso é particularmente verdadeiro quando a questão envolve representatividade. Assim, no que diz respeito à suficiência das informações, não se fará aqui qualquer juízo de valor sobre a qualidade dos dados utilizados, muito menos sobre a representatividade dos mesmos. A metodologia proposta simplesmente pressupõe o uso dos dados de que dispõe as instituições responsáveis, sob a hipótese de representatividade acima referida.

O desenvolvimento metodológico a ser apresentado também procurou preservar as práticas corriqueiras das instituições gerenciadoras de águas, de maneira a não implicar em uma mudança radical na forma como admitem e processam as informações que geram ou obtém por contratação, em particular, no que concerne à amostragem que realizam para caracterizar a qualidade da água nos reservatórios que gerenciam. Conforme evidencia o raciocínio lógico, nenhuma metodologia tem o condão de criar informações novas, mas tão somente extrair do que existe, ou seja, dos dados disponíveis, o essencial para os fins que se deseja atingir.

No caso em tela, a pergunta a ser respondida é: Qual o número mínimo de amostras que precisam ser coletadas e analisadas em um dado reservatório para que os resultados encontrados possam ser considerados suficientemente representativos do corpo hídrico como um todo?

Sendo m o número de locais onde são coletadas n amostras de água ao longo tempo na bacia hidráulica de um determinado reservatório para o qual se deseja determinar a representatividade do conjunto de dados afetos a um certo parâmetro P, a matriz de covariâncias entre os valores medidos é:

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & \dots & s_{1,m} \\ s_{2,1} & s_{2,2} & \dots & s_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n,1} & s_{n,2} & \dots & s_{n,3} \end{bmatrix}$$

Para um conjunto de dados amostrais de um par de vaiáveis A e B, contando com n valores observados, com médias $\bar{a}$ e $\bar{b}$ respectivamente, a covariância será:

$$Cov(A,B) = \frac{1}{(n-1)} \sum_{i=1}^{n} (a_i - \bar{a})\left(b_i - \bar{b}\right)$$

Sendo as variáveis normatizadas, a matriz S é equivalente à matriz de correções R, a qual tem em sua diagonal principal a unidade. Logo, a matriz R será:

$$R = \begin{bmatrix} 1 & r_{1,2} & \dots & r_{1,m} \\ r_{2,1} & 1 & \dots & r_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n,1} & r_{n,2} & \dots & 1 \end{bmatrix}$$

No conjunto original de dados, cada ponto no espaço $m-$dimensional corresponde a uma linha da matriz que abriga os valores originais padronizados ou não. Portanto, um ponto no espaço $m-$dimensional é uma informação sobre o parâmetro P na i-ézima amostra que pode ser referida como um vetor linha X_i com m componentes. Como as m dimensões são correlacionadas entre si, cada um dos n pontos do gráfico exibe um conteúdo que, em parte e subjacentemente, é comum às m variáveis.

Uma vez que as m variáveis são correlacionadas positivamente, sendo m inicialmente um valor relativamente grande, certamente não serão necessárias m dimensões para comunicar um conteúdo parcialmente redundante. Em face dessa redundância, é possível sintetizar tal conteúdo em um número menor de dimensões, de forma a expressar, se não toda, mas a maior parte do conteúdo das m variáveis.

Boa parte da Análise Multivariada lida exatamente com a busca da redução de dimensões, de forma a permitir considerar apenas aquelas que explicam suficientemente a quantidade de informações contidas nas variáveis originais. A quantidade de informação se traduz pela medida de

dispersão dos dados, ou seja, pela variância. Assim, para explicar o máximo possível de informação contida nas variáveis, basta determinar uma dimensão para a qual a variância seja máxima.

Faz-se então necessário determinar a combinação linear das variáveis X_j (vetores colunas) que exiba máxima variância, para que, com isso, cada componente principal explique o máximo de informação contida nas variáveis correlacionadas originais. Geometricamente, estas combinações lineares representam a adoção de um novo sistema de coordenadas a partir de uma rotação aplicada ao sistema original.

Os novos eixos apresentam a vantagem de representarem variáveis não correlacionadas (o que simplifica a análise), orientados segundo as direções com variabilidade máxima. A variância das novas variáveis decresce da primeira dimensão principal para a última.

A determinação matemática das orientações das componentes principais implica em formular o problema considerando que as dimensões principais orientam-se na direção da máxima variância, de maneira que a j-ésimo dimensão explique o máximo da variância remanescente, não explicada pelo seu predecessor. A expressão matricial da j-ésima dimensão principal expressa uma função linear dada por:

$$\vec{z}_j = X\vec{c}_j$$

Onde $\vec{z}_j$ é o vetor coluna $(n \times 1)$ com os n valores transformados da variável analisada; $\vec{c}_j$ é o vetor coluna $(p \times 1)$ que orienta a transformação na direção do j-ésimo componente.

Para se encontrar a variância de $\vec{z}_j$ basta se calcular a variância da função linear acima, ou seja:

$$Var(\vec{z}_j) = Var(X\vec{c}_j) = \vec{c}_j{}^T Var(X)\vec{c}_j = \vec{c}_j{}^T S\vec{c}_j$$

Onde S é a matriz de variância e covariância que estima $Var(X)$ que será igual a matriz de correlações R se as variáveis (colunas da matriz X) forem padronizadas. Para a primeira dimensão principal, orientado na direção dada pelo vetor $\vec{c}_1$, a função objetivo a ser maximizada é:

$$Max\left[Var(\vec{z}_1) = \vec{c}_1{}^T S\vec{c}_1\right]$$

O ponto de máximo da função V pode ser encontrado igualando-se a zero sua derivada em relação ao vetor direcional $\vec{c}_1$:

$$\frac{\partial V}{\partial \vec{c}_1} = 2S\vec{c}_1 - 2\lambda_1\vec{c}_1 = 2(S - \lambda_1 I)\vec{c}_1 = 0$$

A equação a ser resolvida será $(S - \lambda_1 I)\vec{c}_1 = 0$ que reproduz a forma clássica de um problema de autovalores e autovetores, No caso presente, λ_1 e $\vec{c}_1$ representam a raiz característica (um escalar) e o vetor característico da matriz de covariância S (ou de correlação R, se a variável estiver padronizada), respectivamente.

Com essa formulação, a variância total das dimensões principais é expressa pelo traço da matriz de autovalores λ que é igual ao número de variáveis envolvidas na análise, ou seja:

$$\sum_{i=1}^{m} \lambda_{i,i} = m$$

E a proporção da variância total explicada pela dimensão principal de ordem j será igual ao quociente entre o respectivo autovalor e o número de variáveis envolvidas:

$$\frac{\lambda_j}{m}, com\ j = 1,2,3, \dots, m$$

O processo de otimização do número de locais amostrados passa então a ser determinado com base nas percentagens de variância explicada acumulada, observando-se que:

(i) Para um conjunto inicial de dados com m locais amostrados na bacia hidráulica, uma concentração de percentual explicado de variância nas primeiras k dimensões é indicativo de excesso de informação e possibilidade de reduzir o valor de m;

(ii) A presença de autovalores nulos associados à matriz de covariância ou de correlação indica a possibilidade de eliminação de locais de coleta e análise por apresentarem redundância de informação em relação ao conjunto;

(iii) Ao fazer variar m de maneira sistemática pode-se determinar o número mínimo de locais amostrados de forma a preservar a explicação da parcela preponderante da variância do sistema, concentrada no número mínimo de dimensões principais.

Dentre outras, a metodologia descrita tem a vantagem de contornar o óbice analiticamente inabordável do excesso de variáveis influentes no problema da representatividade amostral relatado. Sua aplicação, portanto, é individualizada por reservatório, dada a singularidade desses corpos d'água em termos de valores assumidos pelas variáveis influentes no grau de representatividade que certa

malha de pontos amostrais pode ter em um lago. Outras referências sobre o assunto podem ser encontradas em Dillon & Goldstein (1984), Rencher (2002) e Hair et al. (2005).

Resultados e discussão

O Açude Araras, objeto da aplicação realizada, está localizado no município de Varjota, no semiárido cearense, pertencendo a bacia hidrográfica do Rio Acaraú, e as águas deste reservatório são destinadas para o abastecimento dos municípios de Ipueiras, Nova Russas, Pires Ferreira, Hidrolândia, Tamboril e Varjota, beneficiando uma população de 145 mil habitantes. Por esse motivo é de grande valia o monitoramento desses recursos.

Utilizando-se do banco de dados públicos disponibilizado pela COGERH-CE (Companhia de gestão de Recursos Hídricos do Estado do Ceará), contendo os resultados das amostras coletadas no reservatório mencionado, foi observada a existência de 07 (sete) locais de amostragens denominados respectivamente, ARA-01, ARA-02, ARA-03, ARA-04, ARA-05, ARA-07.

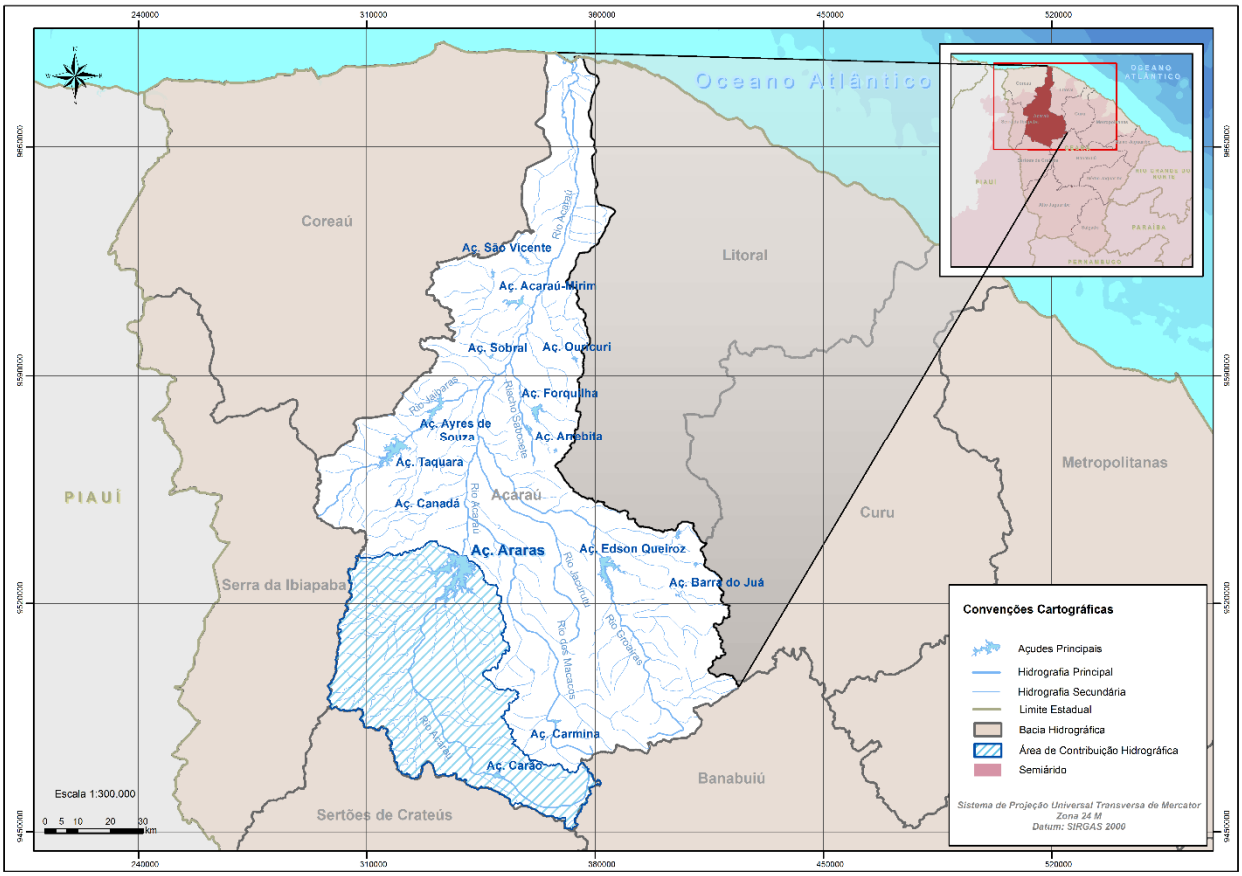

Figura 1 - Localização do Reservatório Araras

Os quadros 1, 2 e 3 a seguir apresentam os dados coletados em cada um dos pontos de amostragem do reservatório. Conforme se observa, existem algumas lacunas nas informações, como por exemplo, referentes às campanhas de 2007,2008 e 2009.

ARA - 01			ARA - 02			ARA - 03		
Data	Mg/l P	Volume (%)	Data	Mg/l P	Volume (%)	Data	Mg/l P	Volume (%)
28/11/2005	0.34	61.56	28/11/2005	0.46	61.56	28/11/2005	0.64	61.56
22/11/2006	0.014	63.53	22/11/2006	0.024	63.53	22/11/2006	0.022	63.53
05/04/2005	0.061	64.15	26/10/2005	0.05	64.92	01/08/2012	0.064	64.15
26/10/2005	0.06	64.92	08/05/2006	0.74	71.01	26/10/2005	0.06	64.92
08/05/2006	0.74	71.01	30/03/2005	0.16	75.34	08/05/2006	0.92	71.01
05/05/2010	0.09	71.11	02/08/2006	0.018	75.44	05/05/2010	0.12	71.11
30/03/2010	0.16	72.49	01/06/2005	0.17	79.74	01/08/2005	0.75	74.46
30/03/2010	0.26	72.49	-	-	-	30/03/2005	0.17	75.34
30/03/2010	0.14	72.49	-	-	-	02/08/2006	0.031	75.44
31/03/2010	0.18	72.49	-	-	-	01/06/2005	0.12	79.74
30/03/2005	0.23	75.34	-	-	-	-	-	-
02/08/2006	0.027	75.44	-	-	-	-	-	-
23/02/2010	0.1	75.64	-	-	-	-	-	-
23/02/2010	0.1	75.64	-	-	-	-	-	-
01/06/2005	0.1	79.74	-	-	-	-	-	-

Quadro 1 - Dados dos pontos de coleta ARA01, ARA02, ARA03

ARA - 04			ARA - 05			ARA - 06		
Data	Mg/l P	Volume (%)	Data	Mg/l P	Volume (%)	Data	Mg/l P	Volume (%)
28/11/2005	0.93	61.56	28/11/2005	0.25	61.56	28/11/2005	0.81	61.56
22/11/2006	0.025	63.53	22/11/2006	0.027	63.53	22/11/2006	0.024	63.53
26/10/2005	0.12	64.92	01/08/2012	0.059	64.15	01/08/2012	0.066	64.15
08/05/2006	0.68	71.01	26/10/2005	0.06	64.92	26/10/2005	0.19	64.92
01/08/2005	0.48	74.46	08/05/2006	0.68	71.01	08/05/2006	0.56	71.01
30/03/2005	0.15	75.34	05/05/2010	0.125	71.11	05/05/2010	0.14	71.11
02/08/2006	0.021	75.44	30/03/2005	0.19	75.34	01/08/2005	0.95	74.46
01/06/2005	0.36	79.74	02/08/2006	0.022	75.44	30/03/2005	0.16	75.34
-	-	-	01/06/2005	0.32	79.74	02/08/2006	0.028	75.44
-	-	-	-	-	-	01/06/2005	0.34	79.74

Quadro 2 - Dados dos pontos de coleta ARA04, ARA05, ARA06

ARA - 07		
Data	Mg/l P	Volume (%)
22/11/2006	0.028	63.53
26/10/2005	0.05	64.92
08/05/2006	0.68	71.01
01/08/2005	0.62	74.46
30/03/2005	0.22	75.34
02/08/2006	0.028	75.44
01/06/2005	0.41	79.74

Quadro 3 - Dados do ponto de coleta ARA07

De modo a tentar quantificar o número mínimo de amostras que sejam suficientemente representativas para obter as cargas de Fósforo do corpo hídrico como um todo, foi confeccionada a matriz S que é a matriz de covariância entre os valores medidos nos 7 (sete) pontos de coletas presentes no reservatório. A matriz S leva em conta os dados coletados de Fósforo Total nos períodos em que o reservatório se encontrava na faixa de 60% a 80% do seu volume máximo, e é mostrada a seguir:

$$S = \begin{bmatrix} -0.4788 & -0.1066 & -0.0309 & -0.4000 & -0.6617 & -0.3999 & -0.0257 \\ -0.1173 & 0.7186 & -0.0582 & 0.0576 & 0.3035 & -0.6084 & 0.0309 \\ -0.5116 & -0.1345 & 0.2545 & 0.1808 & 0.1987 & 0.0703 & 0.7604 \\ -0.2911 & 0.5465 & -0.3985 & -0.1073 & -0.1830 & 0.6407 & 0.0483 \\ -0.4637 & -0.1836 & 0.0481 & -0.4589 & 0.5957 & 0.1006 & -0.4164 \\ -0.4139 & 0.0376 & 0.2765 & 0.6921 & -0.1553 & 0.0606 & -0.4940 \\ 0.1553 & 0.3471 & 0.8327 & -0.3207 & -0.1345 & 0.2018 & -0.0200 \end{bmatrix}$$

Normatizando as variáveis encontradas na matriz S, obtém-se a matriz R, onde S e R são equivalentes. Logo a matriz R é:

$$R = \begin{bmatrix} 1 & 0.0285 & 0.7900 & 0.4217 & 0.8688 & 0.4827 & -0.2445 \\ 0.0285 & 1 & 0.0358 & 0.7953 & -0.0433 & 0.2204 & 0.3069 \\ 0.7900 & 0.0358 & 1 & 0.2740 & 0.8424 & 0.8840 & -0.2091 \\ 0.4217 & 0.7953 & 0.2740 & 1 & 0.2945 & 0.3061 & -0.0842 \\ 0.8688 & -0.0433 & 0.8424 & 0.2945 & 1 & 0.4122 & -0.2298 \\ 0.4827 & 0.2204 & 0.8840 & 0.3061 & 0.4122 & 1 & -0.1467 \\ -0.2445 & 0.3069 & -0.2091 & -0.0842 & -0.2298 & -0.1467 & 1 \end{bmatrix}$$

Cada linha e cada coluna da matriz R representa um ponto de coleta, e os valores de cada um dos elementos existentes na referida matriz são obtidos a partir da correlação direta dos resultados das amostras coletadas em cada um desses pontos.

Observa-se uma forte correlação do ponto de coleta ARA – 01 com o ARA – 03 e ARA – 05 (0.7900 e 0.8688 respectivamente), fator esse que indica a redundância de informações em cada um desses pontos, conforme ilustra a Figura 2 a seguir.

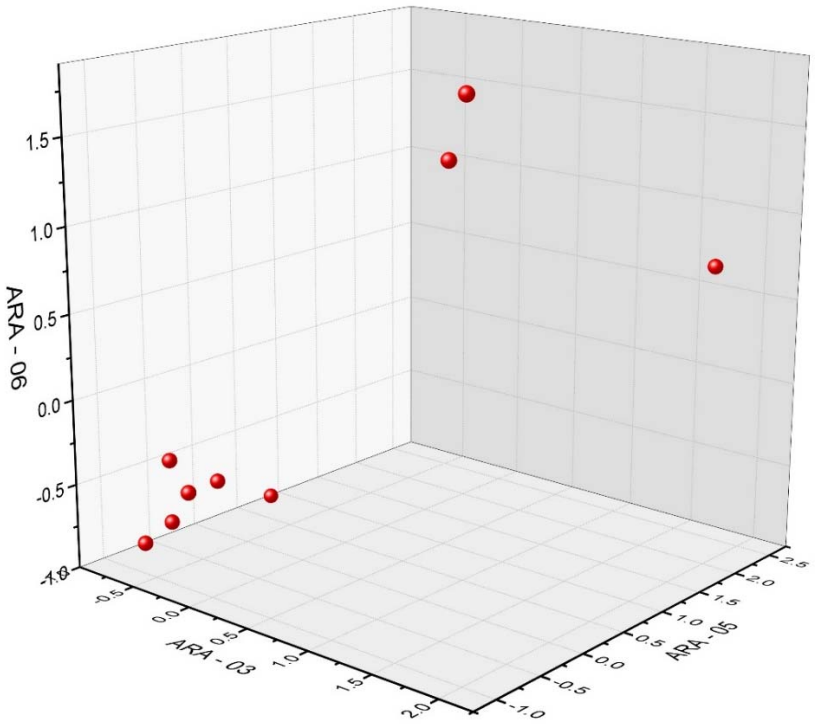

Figura 2 - Correlação dos pontos ARA - 06, ARA - 03 e ARA - 05

Da mesma forma, a Figura 3 que segue, demonstra a total falta de correlação entre as amostras coletadas nos pontos ARA – 01, ARA – 02 e ARA – 07

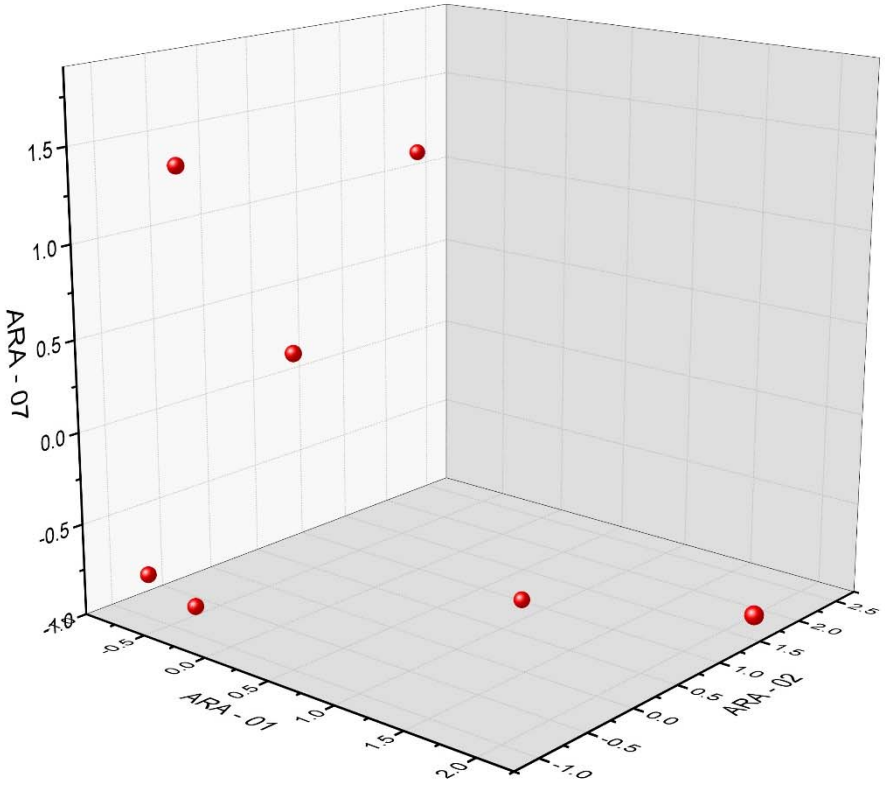

Figura 3 - Correlação dos pontos ARA - 07, ARA - 01 e ARA - 02

A partir da utilização dos autovalores da matriz R, representados pela matriz D, calcula-se os fatores de carga da matriz L, que se dá pela equação:

$$L = S\sqrt{D}$$

$$D = \begin{bmatrix} 3.4455 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.7646 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.902 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7174 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1479 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0604 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0379 \end{bmatrix}$$

Desse modo, temos L:

$$L = \begin{array}{c} \begin{matrix} CP1 & CP2 & CP3 & CP4 & CP5 & CP6 & CP7 \end{matrix} \\ \begin{bmatrix} \mathbf{-0.889} & -0.142 & -0.029 & -0.339 & -0.254 & -0.098 & -0.005 \\ -0.218 & \mathbf{0.955} & -0.055 & 0.049 & 0.117 & -0.150 & 0.006 \\ \mathbf{-0.950} & -0.179 & 0.242 & 0.153 & 0.076 & 0.017 & 0.148 \\ -0.540 & \mathbf{0.726} & -0.378 & -0.091 & -0.070 & 0.157 & 0.009 \\ \mathbf{-0.861} & -0.244 & 0.046 & -0.389 & 0.229 & 0.025 & -0.081 \\ \mathbf{-0.768} & 0.050 & 0.263 & 0.586 & -0.060 & 0.015 & -0.096 \\ 0.288 & 0.461 & \mathbf{0.791} & -0.272 & -0.052 & 0.050 & -0.004 \end{bmatrix} \end{array}$$

É possível visualizar a significância física da matriz acima ao se observar a correlação entre os seus elementos com as componentes principais. Segundo Helena et al. (2000), coeficientes de correlação superiores a 0,5 expressam uma forte relação entre as variáveis de qualidade de água. A componente principal 1 (CP1) é a dimensão com a qual a maioria dos pontos amostrais exibem maior correlação. A Tabela 1 a seguir mostra os autovalores calculados, bem como o percentual de variância explicado individual e acumulado. Nela se vê que a consideração da quarta componente já implica em um percentual de 98% da variância explicada

.

Tabela 1 - Autovalores encontrados

Autovalor	%	% Acumulada
3.4455	49%	49%
1.7646	25%	74%
0.902	13%	87%
0.7174	10%	98%
0.1479	2%	100%
0.0604	1%	101%

Conclusões

A utilização da técnica estatística de análise Multivariada cumpre seu papel de reduzir a quantidade de dimensões necessárias à caracterização dos corpos d´água, minimizando-se a redundância de informações. A racionalização do número de dimensões envolvidas preserva

satisfatoriamente o real panorama de uma das variáveis que influenciam diretamente no nível de eutrofização do reservatório, no caso em questão, o fósforo.

A metodologia proposta, para ser efetivada na prática, demanda uma exploração inicial sistemática do lago, em termos de uma grade de pontos amostrais, a partir dos quais se possa, para diferentes níveis de armazenamento do reservatório, obter as matrizes de correlação, sobre as quais procede-se os passos descritos no presente artigo. A partir dos resultados encontrados, pode-se então eliminar da grade amostral de pontos de coleta aqueles que não agregam significantemente informações à caracterização qualitativa das águas do corpo hídrico em questão.

Referências

DILLON, W. R.; GOLDSTEIN, M., 1984, **Multivariate analysis methods and application**. New York: John Wiley. 1984. 587p.

HARMAN, H. H., 1976: **Modern Factor Analysis**. Chicago: University of Chicago Press.

MATALAS, N. C.; REIHER, B. J.,1967: **Some Comments on the Use of Factor Analysis.** In: Water Resources Research 3(1), 213-224.