CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN impresso 1413-9022 / ISSN on-line 2317-4536 - v.43, p.18 - 38, 2017
DOI: 10.12957/cadest.2017.31363

PADRÕES DE VARIABILIDADE EM VAZÕES AFLUENTES A USINAS HIDRELÉTRICAS E ASSOCIAÇÕES COM MASSAS DE AR

Erick Meira de Oliveira
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
erickmeira89@gmail.com

Fernando Luiz Cyrino Oliveira
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
cyrino@puc-rio.br

Resumo

Os padrões fluviais brasileiros são sensíveis a diversos fatores geográficos e atmosféricos locais, sofrendo variações consideráveis ao longo da extensão territorial do País. Apesar da vasta gama de informações hoje disponíveis sobre esses fatores, sabe-se que a vazão de rios brasileiros possui relações significativas com padrões climáticos bem característicos, podendo se substituir, por vezes, uma quantidade expressiva de elementos por amostras que concentrem a maior parte das informações estatísticas da base de dados. Nesse contexto, este trabalho buscou identificar, por meio da aplicação de diferentes técnicas de análise estatística multivariada, características comuns a diferentes rios brasileiros, utilizando como informação principal suas vazões. Os resultados sugerem que os comportamentos de grande parte dos rios brasileiros usados para aproveitamento hidrelétrico podem ser caracterizados por um conjunto pequeno de padrões bem definidos, que traduzem os diferentes regimes de massas de ar aos quais o território brasileiro está submetido.

***Palavras-chave**: Análise Multivariada; Vazões de Rios; Massas de ar.*

1. Introdução

A hidrografia do Brasil é caracterizada pela ampla extensão e diversidade de bacias. Estima-se que o país concentre entre 12% a 13% das reservas de água doce disponíveis no mundo, distribuídas entre mais de dois milhões de hectares de pântanos, reservatórios e estuários, bem como nos 25 mil rios ao longo do país (ANA, 2013).

Os padrões de vazão de rios brasileiros possuem relações significativas com padrões climáticos e de teleconexões (Cardoso e Cataldi, 2012). As influências remotas desses padrões no comportamento dos rios ocorrem, usualmente, via precipitação (Ferraz *et al*., 2013). Argumenta-se que as ocorrências de eventos cíclicos também são responsáveis, em parte, pela mudança nas vazões dos rios. A título de exemplo, o El Niño (La Niña) é responsável pelo aumento (diminuição) de extremos de vazão na região das bacias do Paraguai e Uruguai (Brasil central) (Tedeschi e Grimm, 2008). A La Niña também exerce influência significativa nos padrões hidrológicos de bacias do norte do Brasil (Boening *et al*., 2012; Andrade *et al*., 2016).

Além das séries históricas das variáveis se tornarem cada vez maiores e com disponibilidades em horizontes temporais mais curtos, a quantidade de novas informações disponíveis sobre sistemas hídricos cresce a cada dia. Nesse contexto, com o intuito de contornar o problema da enorme dimensão e complexidade dos dados hoje disponíveis, é imprescindível o desenvolvimento de técnicas de redução, classificação e filtragem de informações. Muito embora a omissão de características individuais de rios e peculiaridades hidrológicas das bacias aos quais estes pertencem traduza perdas de informação, entende-se que a agregação de grandes massas de dados em conjuntos menores e equivalentes é uma simplificação necessária a certos sistemas para tornar viável seu planejamento.

Dentro do contexto supracitado, as técnicas de análise multivariada de dados surgem como um importante conjunto de ferramentas estatísticas de apoio à decisão. Além de propiciarem o estudo simultâneo do comportamento entre diversas variáveis de interesse numa pesquisa, um grupo específico de técnicas multivariadas permite classificar os elementos de um conjunto de observações em um número restrito de grupos homogêneos, segundo algum critério de homogeneidade. Destacam-se, nesse caso, os chamados métodos fatoriais, nos quais a redução do número de variáveis se dá por meio da construção de novas variáveis sintéticas, denominadas fatores. O objetivo principal

dos métodos fatoriais multivariados é a identificação das medidas responsáveis pelas maiores variações entre os resultados, transformando um conjunto original de informações em outro equivalente, de menor dimensão, sem perdas significativas de informações. A classificação final em fatores, dependendo da natureza do problema abordado, permite a identificação de aspectos importantes, como a geração, a seleção e a interpretação dos fatores investigados.

À luz do acima exposto, e tendo em vista a importância que os rios assumem na economia do país (desde a utilização para consumo humano e a reserva de água potável até à geração de energia elétrica), este trabalho busca identificar características comuns a diferentes rios brasileiros, utilizando como informação principal suas vazões mensais, expressas em m^3/s. Para tanto, busca-se primeiramente obter, por meio da aplicação de diferentes técnicas de análise estatística multivariada, padrões específicos de comportamento dos rios. Em um segundo momento, procura-se associar as especificidades dos padrões obtidos a diferentes modelos climáticos, verificando qual destes melhor se aderem à realidade dos rios brasileiros e permitindo a simplificação de estudos de relações entre clima, chuva e vazão.

O presente trabalho está estruturado em outras 4 seções, além desta introdução. Na Seção 2, um breve referencial teórico é apresentado sobre o tema. A Seção 3 explica a metodologia multivariada empregada no trabalho. A Seção 4 traz os resultados, bem como suas discussões. Por fim, a Seção 5 conclui e indica direções para trabalhos futuros.

2. Referencial Teórico

As informações acerca das magnitudes e frequências de ocorrências de vazões em rios são insumos básicos em diversos estudos, sobretudo, econômicos e ambientais. Essa situação é ainda mais crítica em países com dimensões continentais onde, geralmente, a variabilidade espacial dos padrões hidrológicos é bastante pronunciada. A grande variabilidade e complexidade de dados sobre regimes hidrológicos, por sua vez, limita o uso de métodos estatísticos univariados na identificação e/ou avaliação de padrões específicos. Nesse contexto, o uso de técnicas multivariadas na avaliação de características hidrológicas específicas tem se tornado bastante popular ao longo dos anos (Marengo *et al*., 1995; Pantsar-Kallio *et al*., 1999; Simeonov *et al*., 2002; Arslan, 2009).

O uso de técnicas estatísticas multivariadas para determinação de padrões fluviais utilizando-se vazões de rios como variáveis, entretanto, é relativamente novo. Alguns exemplos de natureza parecida são os trabalhos de Assani *et al*. (2006) e Noori *et al*. (2010). Os primeiros aplicaram a Análise de Componentes Principais (ACP) a um conjunto de 18 variáveis hidrológicas distintas, obtidas para a região do sul de Québec, no Canadá, e extraíram cinco componentes que, segundo sua interpretação, associavam-se a características fundamentais do regime hidrológico de vazões mínimas anuais da região: frequência e magnitude, timing, variabilidade interanual de timing, variabilidade interanual de magnitude e formato da curva de distribuição. Já Noori *et al.* (2010) utilizaram a ACP, juntamente com outras técnicas de aprendizado de máquina, como Redes Neurais Artificiais, para prever o nível de vazão do rio Sofichay, localizado na província do Azerbaijão Leste, no Noroeste do Irã. Para tanto, os autores fizeram uso de variáveis mensais como precipitação, vazão, radiação solar e temperaturas (mínima, máxima e média). Apesar disso, quase nenhum estudo utilizou, como variáveis originais, as vazões de múltiplos rios, visando à identificação de padrões específicos associados à hidrologia de certa região ou país. Uma exceção é o trabalho de Berhanu *et al.* (2015), no qual os autores utilizaram dados de vazão de rios em 208 estações de medição distintas, propondo uma classificação dos rios por clusters segundo o método hierárquico de Ward, uma técnica multivariada de classificação que procura por partições que minimizem a perda associada a cada agrupamento. O presente trabalho possui uma abordagem similar, por considerar como variáveis as vazões mensais em pontos específicos de medição de rios selecionados, porém busca também associar os resultados obtidos com diferentes modelos climáticos, verificando qual destes melhor se aderem à realidade dos rios brasileiros.

3. Dados e Metodologia

Para as suas funções de planejamento da operação do sistema elétrico, o Operador Nacional do Sistema Elétrico (ONS) possui diversos modelos de otimização da operação eletroenergética. Esses modelos utilizam séries de vazões médias diárias, semanais e mensais. De acordo com o ONS, dadas as metodologias e critérios utilizados atualmente na previsão de vazões, pode-se dispensar a disponibilidade de vazões para alguns locais de aproveitamento em operação. Assim, em geral, adota-se a realização de previsão de

vazões para um subconjunto de aproveitamentos de cada bacia, denominados postos base (ONS, 2015).

O presente trabalho tem como objetivo analisar as vazões mensais dos postos supracitados, disponibilizadas pelo ONS, utilizando as seguintes técnicas multivariadas:

- Análise de Componentes Principais (ACP); e
- Análise de Fatores (AF).

O arranjo e o tratamento dos dados bem como a aplicação das técnicas supracitadas foram feitos com o auxílio de dois softwares: MATLAB (R2016b) e R. Espera-se, a partir dessas análises, associar os resultados obtidos a padrões climáticos característicos para o caso brasileiro.

3.1 Seleção das Variáveis

A base de dados original continha dados de vazões médias mensais em 192 postos base, no período compreendido entre 1931 e 2014. Os dados estavam organizados da seguinte forma: para cada posto de medição, uma matriz de vazões médias mensais (em m^3/s) era fornecida, com as colunas representando os meses e as linhas os anos entre 1931 e 2014 (ONS, 2017).

As variáveis de interesse foram definidas como sendo as vazões mensais de cada posto de medição no tempo (série temporal). O primeiro passo, portanto, foi adequar a base de dados original a um novo formato, que permitisse considerar os postos como variáveis. Dessa forma, as matrizes de cada posto foram transformadas em vetores de tamanho 1008 (12 meses x 84 anos) e a nova matriz de dados foi formada utilizando os referidos vetores como colunas (variáveis).

No relatório disponibilizado pelo ONS consta a informação de localização de cada posto de medição, isto é, a qual rio ele pertence e a sua respectiva bacia hidrográfica. A partir disso, observou-se que alguns rios possuíam diversos postos de medição e que a base de dados obtida fornecia, portanto, observações para 62 rios distintos, distribuídos em 22 bacias hidrográficas. Optou-se por eliminar as bacias que possuíam dados disponíveis para apenas um ou dois rios. Com isso, a base de dados selecionada passou a envolver dados de 10 bacias diferentes. Além disso, escolheu-se utilizar apenas um ponto de medição em cada rio, cuja seleção seguiu o critério de maior vazão média em todo o período de medição considerado.

Após a definição do novo arranjo, a matriz de dados selecionada neste trabalho passou a conter 45 variáveis (rios diferentes, distribuídos entre 10 bacias) e 1008 observações. A Tabela 1, a seguir, mostra a relação dos rios escolhidos e as bacias a que pertencem, bem como o código atribuído a cada um dos rios ao longo do trabalho. A nova base de dados selecionada está disponível para consulta na plataforma Kaggle, acessando-se o seguinte link: https://www.kaggle.com/erickmeira/br-monthly-river-flows

Tabela 1 – Relação dos rios e bacias selecionados

Rio	Bacia	Código	Rio	Bacia	Código
Araguari	Amazonas	AM_1	Paraibuna	Paraíba do Sul	PS_3
Apiacás	Amazonas	AM_2	Peixe	Paraíba do Sul	PS_4
Aripuanã	Amazonas	AM_3	Piraí	Paraíba do Sul	PS_5
Comemoração	Amazonas	AM_4	Ribeirão das Lajes	Paraíba do Sul	PS_6
Curuá-Una	Amazonas	AM_5	Paraná	Paraná	PA_1
Guaporé	Amazonas	AM_6	Claro	Paranaíba	PB_1
Jamari	Amazonas	AM_7	Corrente	Paranaíba	PB_2
Jari	Amazonas	AM_8	Corumbá	Paranaíba	PB_3
Madeira	Amazonas	AM_9	Paranaíba	Paranaíba	PB_4
Teles Pires	Amazonas	AM_10	São Marcos	Paranaíba	PB_5
Uatumã	Amazonas	AM_11	Verde	Paranaíba	PB_6
Xingu	Amazonas	AM_12	Preto	São Francisco	SF_1
Doce	Doce	DO_1	São Francisco	São Francisco	SF_2
Piracicaba	Doce	DO_2	Guarapiranga	Tietê	TI_1
Santo Antônio	Doce	DO_3	Pinheiros	Tietê	TI_2
Iguaçu	Iguaçu	IG_1	Tietê	Tietê	TI_3
Jordão	Iguaçu	IG_2	Canoas	Uruguai	UR_1
Correntes	Paraguai	PG_1	Chapecó	Uruguai	UR_2
Itiquira	Paraguai	PG_2	Ijuí	Uruguai	UR_3
Jauru	Paraguai	PG_3	Passo Fundo	Uruguai	UR_4
Manso	Paraguai	PG_4	Pelotas	Uruguai	UR_5
Jaguari	Paraíba do Sul	PS_1	Uruguai	Uruguai	UR_6
Paraíba do Sul	Paraíba do Sul	PS_2			

Fonte: Os autores.

3.2 Análise de Componentes Principais (ACP)

A Análise de Componentes Principais (ACP) é uma técnica de análise multivariada que busca explicar a estrutura de variância-covariância de um conjunto de variáveis através de combinações lineares destas. Seus objetivos gerais são, portanto, a redução de variáveis e a interpretação do conjunto observado.

Embora $\boldsymbol{p}$ componentes sejam necessários para reproduzir a variabilidade total do sistema, normalmente, boa parte dessa variabilidade pode ser explicada por um número pequeno $\boldsymbol{k}$ de componentes principais. Assim, há praticamente a mesma quantidade de informação nos $\boldsymbol{k}$ componentes que nas $\boldsymbol{p}$ variáveis originais. Os $\boldsymbol{k}$ componentes principais podem, então, substituir as $\boldsymbol{p}$ variáveis e o conjunto de dados de $\boldsymbol{n}$ medidas das $\boldsymbol{p}$ variáveis será reduzido a um conjunto de $\boldsymbol{n}$ medidas nos $\boldsymbol{k}$ componentes principais (Johnson e Wichern, 2002). Frequentemente, a ACP revela relações que não eram evidentes anteriormente e, desta forma, permite novas interpretações, como, por exemplo, associações com padrões climáticos específicos para a base de dados deste trabalho.

Algebricamente, as Componentes Principais (CPs) são combinações lineares específicas das $\boldsymbol{p}$ variáveis aleatórias. Geometricamente, essas combinações lineares representam a seleção de um novo sistema de coordenadas obtido através da rotação do sistema original. Os novos eixos representam as direções com máxima variabilidade e fornecem uma descrição mais simples da estrutura de covariância. Em linhas gerais, podemos descrever o modelo de obtenção do vetor de $\boldsymbol{p}$ CPs – $\boldsymbol{CP} = \left[CP_1, CP_2, \dots, CP_p\right]'$ – da seguinte forma:

$$\boldsymbol{CP} = \boldsymbol{A}'\boldsymbol{X} \tag{1}$$

onde $\boldsymbol{A}'$ é a transposta da matriz de autovetores ordenados de forma decrescente pelo valor de seus autovalores associados e $\boldsymbol{X}$ é a matriz $\boldsymbol{n}$ x $\boldsymbol{p}$ de dados. Dessa forma, se representada essa matriz de dados na forma de $\boldsymbol{p}$ vetores $\boldsymbol{n}$ x 1 agrupados – $\boldsymbol{X} = \left[\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_p}\right]$ –, garante-se que as combinações lineares

$$\begin{aligned} CP_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}\boldsymbol{X_1} + a_{12}\boldsymbol{X_2} + \cdots + a_{1p}\boldsymbol{X_p} \\ CP_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}\boldsymbol{X_1} + a_{22}\boldsymbol{X_2} + \cdots + a_{2p}\boldsymbol{X_p} \\ &\dots \\ CP_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}\boldsymbol{X_1} + a_{p2}\boldsymbol{X_2} + \cdots + a_{pp}\boldsymbol{X_p} \end{aligned} \tag{2}$$

são descorrelatadas. Além disso, a $CP_i = \boldsymbol{a}_i'\boldsymbol{X}$, *i = 1,...,p* é a combinação linear que maximiza $Var\ (\boldsymbol{a}_i'\boldsymbol{X})$, sujeito a $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$. Assim, a porcentagem da variância explicada pelas CPs é dada por:

$$\%Var\,(\boldsymbol{X}) \text{ até a q} - \text{ésima CP} = \sum_{i=1}^{q} \lambda_i \Big/ \sum_{j=1}^{p} \lambda_j \qquad (3)$$

onde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ são os autovalores da matriz de covariâncias $\boldsymbol{\Sigma}$ (ou de correlações $\boldsymbol{\rho}$). Normalmente, as primeiras CPs concentram grande parte da variação da amostra, podendo-se desprezar as demais em detrimento da simplificação do problema. Há métodos bem estabelecidos para a determinação do número ótimo de CPs ($\boldsymbol{k}$), que serão comentados na próxima seção.

Neste trabalho, a ACP foi utilizada como uma primeira avaliação das variáveis, para observar como elas são agrupadas nas componentes principais, além de fornecer uma ideia da quantidade de componentes para uma possível redução de variáveis. A observação da matriz de correlação para os rios revela a existência de variáveis altamente correlacionadas entre si, porém com baixa correlação com outras variáveis da matriz, o que indica a possibilidade de redução das variáveis.

3.3 Análise de Fatores (AF)

A Análise de Fatores (AF), outra técnica de análise multivariada, se propõe a descrever as relações de covariância entre diversas variáveis em termos de algumas quantidades aleatórias subjacentes chamadas fatores, que não são observáveis. A AF pode ser considerada como uma extensão da Análise de Componentes Principais (ACP). Ambos podem ser vistos como tentativas de aproximar a matriz de covariância. As técnicas diferem, entretanto, com respeito aos outputs: na AF, os fatores não são uma combinação linear das variáveis originais. Reis (1997) acrescenta que a ACP não busca explicar as correlações existentes entre as variáveis, mas sim encontrar funções matemáticas, entre as variáveis iniciais, que expliquem o máximo possível da variação existente nos dados e permita descrever e reduzir essas variáveis. Já a AF explica a estrutura das covariâncias, entre as variáveis, utilizando um modelo estatístico causal e pressupondo a existência de $\boldsymbol{p}$ variáveis não observadas e subjacentes aos dados. Nesse contexto, os fatores expressam o que existe de comum nas variáveis originais.

O modelo dos fatores postula que o vetor aleatório observável $\boldsymbol{X}$ (com $\boldsymbol{p}$ componentes, média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$) é linearmente dependente de algumas variáveis aleatórias não observáveis: os fatores comuns. Em sua forma mais geral, o modelo é regido pela seguinte equação:

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon} \quad (4)$$

onde $\boldsymbol{L}$ é a matriz de loadings dos fatores (uma medida que reflete a proporção de variação de determinada variável que é explicada pelo fator, ou ainda, o quanto cada variável contribui na formação de cada componente), $\boldsymbol{F}$ é a matriz formada pelos fatores e $\boldsymbol{\varepsilon}$ é um vetor aleatório de erros.

Neste trabalho, o objetivo da análise de fatores, além de reduzir a dimensionalidade da base de dados, é identificar os fatores comuns e analisar de que forma eles sintetizam a informação das variáveis. Em outras palavras, deseja-se entender que tipo de informação os fatores reúnem.

4. Resultados e Discussões

Em uma primeira análise descritiva dos dados envolvidos, como ilustrado na Tabela 2, observa-se que certos rios possuem vazões significativamente elevadas, quando comparados aos demais. As variações absolutas nas medições desses rios, naturalmente, também são grandes. Assim, se as análises fossem feitas com os valores absolutos (em m^3/s), os rios com maiores vazões dominariam a composição das componentes (no caso da ACP) ou dos fatores (no caso da AF), trazendo pouca ou nenhuma contribuição em termos de interpretação do problema.

Em virtude do acima exposto, o primeiro passo (após a obtenção da base de dados já filtrada de acordo com os rios e bacias de interesse) foi a normalização dos dados. Para tanto, utilizou-se o procedimento padrão dos softwares de análise envolvidos – MATLAB (R2016b) e R, como já comentado – que consiste em subtrair, para cada série temporal (cada rio), a média total de suas vazões durante o período abrangido e dividir o resultado pelo desvio padrão das observações.

Tabela 2 – Estatísticas descritivas básicas dos rios selecionados

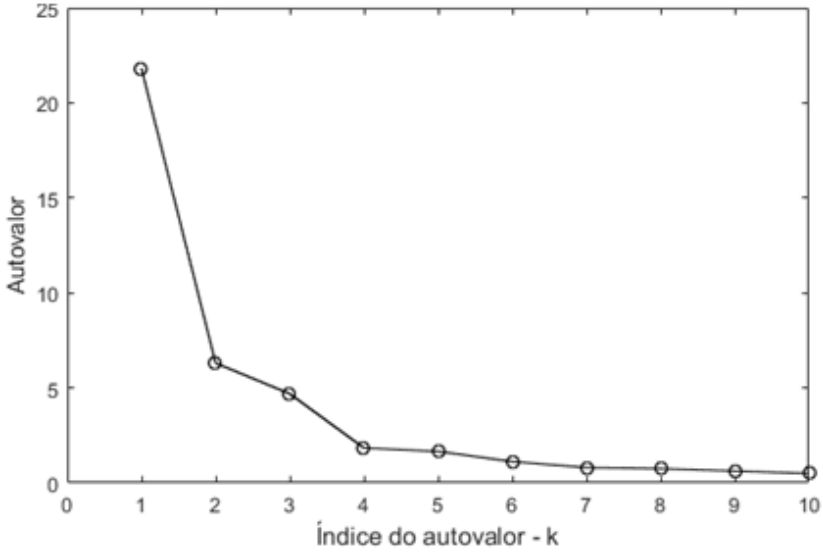
Rio	Mínimo (m^3/s)	Máximo (m^3/s)	Mediana (m^3/s)	Média (m^3/s)	Erro Pad. (m^3/s)	Coef. Assimetria	Coef. Curtose
AM_1	43,37	3368,63	853,37	979,44	709,91	0,63	-0,55
AM_2	3,00	1022,00	92,50	201,02	207,79	1,10	0,36
AM_3	27,36	1515,16	204,89	339,96	315,78	0,90	-0,21
AM_4	33,00	366,00	71,00	85,65	41,85	1,73	4,20

AM_5	44,81	826,77	162,71	214,05	138,12	1,48	2,04
AM_6	25,57	48,48	32,13	33,06	4,09	0,79	0,27
AM_7	9,49	1339,00	240,84	354,12	299,51	0,78	-0,54
AM_8	32,70	5138,43	910,34	1095,44	808,24	1,11	1,26
AM_9	1407,00	54434,60	17287,00	18941,14	11414,67	0,43	-0,88
AM_10	316,00	8150,00	1843,50	2307,42	1543,24	0,78	-0,30
AM_11	19,00	2690,00	489,00	602,19	411,11	1,16	1,32
AM_12	380,00	42442,00	4506,00	8067,52	7790,65	1,16	0,83
DO_1	191,53	5262,00	731,00	960,13	681,95	1,87	4,56
DO_2	16,28	499,00	59,00	77,51	54,18	2,30	7,89
DO_3	25,00	1097,00	123,00	160,13	116,36	2,03	6,73
PS_1	5,15	111,04	24,76	29,54	15,85	1,44	2,41
PS_2	167,00	2708,36	536,83	660,05	383,86	1,44	2,25
PS_3	20,87	280,63	63,58	77,33	42,51	1,72	3,74
PS_4	8,91	141,94	31,70	38,61	21,44	1,69	3,54
PS_5	5,39	196,46	18,19	23,76	16,88	2,78	15,74
PS_6	58,00	195,00	172,50	154,09	39,55	-0,54	-1,15
SF_1	9,00	294,00	45,00	55,01	35,20	1,95	5,49
SF_2	501,00	16102,00	1917,00	2718,16	2029,47	1,65	3,96
PG_1	28,39	156,00	74,00	76,07	20,03	0,49	0,66
PG_2	17,32	244,00	63,00	73,47	32,53	1,29	2,08
PG_3	55,00	154,00	82,00	85,78	17,24	0,93	0,64
PG_4	41,88	727,00	119,00	171,15	120,47	1,42	1,71
PB_1	62,00	654,00	187,87	218,31	108,72	1,02	0,62
PB_2	22,00	179,00	60,00	63,00	20,19	1,06	1,74
PB_3	74,00	1955,00	341,00	453,62	324,86	1,52	2,50
PB_4	450,00	9931,00	1875,52	2400,29	1554,32	1,30	1,49
PB_5	14,57	834,00	135,00	177,71	131,03	1,47	2,25
PB_6	92,00	521,65	178,13	196,57	68,83	1,03	0,85
IG_1	160,00	11670,00	1176,50	1472,18	1126,69	2,38	10,11
IG_2	14,00	841,00	82,00	103,76	82,54	2,78	13,03
PA1	2839,00	31630,00	9091,00	10301,00	4778,67	1,00	1,02
TI_1	3,00	59,64	11,00	12,71	7,06	1,67	4,61
TI_2	5,00	140,00	25,00	29,02	15,94	1,46	3,94
TI_3	165,00	3761,00	660,00	804,76	487,84	1,80	4,73
UR_1	17,00	2932,00	237,00	314,19	260,75	2,49	12,43
UR_2	3,00	621,00	62,00	79,97	65,08	2,05	7,36
UR_3	21,00	1490,00	191,10	249,64	213,09	1,95	5,45
UR_4	2,00	702,00	73,00	96,45	83,86	2,19	7,91
UR_5	44,00	5925,00	573,00	739,10	596,12	2,06	7,63
UR_6	79,00	10048,00	977,50	1277,35	1048,02	2,09	7,40

Fonte: Os autores.

Considerando-se primeiramente a ACP, para se definir quantas componentes principais (CPs) seriam necessárias para se ter uma boa representação da variância total explicada, gerou-se um Scree plot - gráfico que exibe os autovalores (da matriz de correlações $\boldsymbol{\rho}$ dos dados) associados a cada CP (em ordem decrescente) versus o número (índice) da CP -, bem como computou-se a a porcentagem da variância explicada por cada componente. Os resultados para os dados normalizados estão ilustrados na Figura 1 e na Tabela 3, logo abaixo. Ao se observar a Figura 1, nota-se que há um "joelho" (ponto de inflexão) no quarto autovalor, indicando que quatro componentes principais (CP) seriam suficientes. Isso porque, a partir dessa CP, as contribuições adicionais de demais componentes à variância total explicada são pouco relevantes, em comparação com as primeiras. Já a porcentagem da variância total explicada pelas quatro primeiras componentes principais, como ilustrado na Tabela 3, é igual a 85,04%, composição bastante razoável considerando-se a necessidade de se reduzir a dimensão dos dados a número restrito de componentes. A partir desses resultados, escolheu-se analisar as quatro primeiras CPs.

Figura 1 – ACP – Scree plot para os 10 maiores autovalores da matriz de correlação dos dados ($\boldsymbol{\rho}$)



Fonte: Os autores.

Tabela 3 – Porcentagem da variância total explicada para os 10 maiores autovalores de $\boldsymbol{\rho}$

Componente Principal (CP)	% Variância Explicada
CP_1	53,0943
CP_2	16,0844
CP_3	13,3527
CP_4	2,5082
CP_5	2,0969
CP_6	1,5490
CP_7	1,1874
CP_8	1,0990
CP_9	0,9598
CP_{10}	0,8389

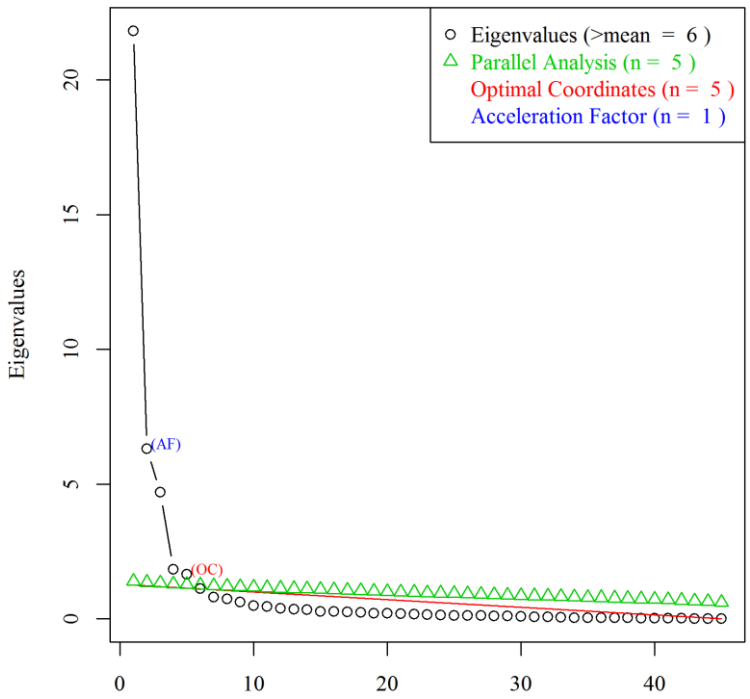
Fonte: Os autores.

Passando-se à análise específica das componentes principais selecionadas, observa-se que a primeira componente principal (CP_1) possui loadings (proporção de variação de cada rio que é explicada pela CP) elevados para muitos rios, à exceção daqueles pertencentes às bacias do rio Iguaçu e do rio Uruguai e de deteminados rios da bacia Amazônica: Araguari, Curuá-Una, Jari e Uatumã. Essas exceções correspondem a rios do extremo sul e do extremo norte do país. Já a CP_2 concentra *loadings* maiores em rios das bacias do Iguaçu e Uruguai e possui valores médios para a bacia dos rios Tietê e Paraná. Essas bacias estão localizadas no sul e sudeste (parte mais ao sul) do país. A terceira CP possui *loadings* mais altos na bacia Amazônica e valores negativos nas bacias dos rios Doce, São Francisco, Paraíba do Sul e Paranaíba – esses últimos são rios do sudeste brasileiro. Finalmente, para a quarta CP, os *loadings* são maiores para a bacia dos rios Tietê e Paraíba do Sul. Para quase todos os outros rios, os valores são negativos.

Em suma, pode-se argumentar que as CPs selecionadas identificam regiões de localização das bacias, que podem estar correlacionadas, por exemplo, pela proximidade (rios que desembocam em outros de bacias próximas) ou pelo mesmo tipo de clima da região. Entretanto, as interpretações das CPs não são muito evidentes: necessita-se do auxílio de especialistas para maior entendimento. Dessa forma, para se conservar espaço, os loadings obtidos em cada uma das CPs não serão aqui ilustrados, em detrimento daqueles obtidos na próxima análise – AF – que demonstrou resultados mais satisfatórios em termos de interpretação imediata.

Passando-se à Análise de Fatores (AF), alguns testes iniciais foram realizados no R e no MATLAB, variando-se o número de fatores comuns na análise, com o intuito de determinar o número ótimo dos mesmos. Todos os modelos foram estimados utilizando-se o método de rotação ortogonal VARIMAX, cujo objetivo é maximizar a variação entre os pesos de cada fator. Um auxílio para a decisão foi o *Scree test*, um conjunto de procedimentos fornecidos pelo R que busca, através de quatro diferentes métodos, indicar um número ótimo de fatores. Para se conservar espaço, os detalhes sobre os referidos métodos não serão aqui comentados. Recomenda-se ao leitor interessado o trabalho de Raîche *et al*. (2013). Os resultados do *Scree test*, como observados na Figura 2, sugerem um número de ótimo de cinco fatores em dois dos quatro procedimentos. O percentual da variação total explicada com cinco fatores é de 77,2%. Para validar a escolha, também foi realizado o teste chi-quadrado de razão de verossimilhança (corrigida por Bartlett) (Lawley, 1956) com a hipótese nula de que o número de fatores comuns selecionados (5, no caso) descrevem adequadamente as relações entre as variáveis originais envolvidas. O teste, realizado no MATLAB 2016R indicou que a hipótese nula não deve ser rejeitada, sugerindo, portanto, que o modelo de cinco fatores pode ser considerado adequado.

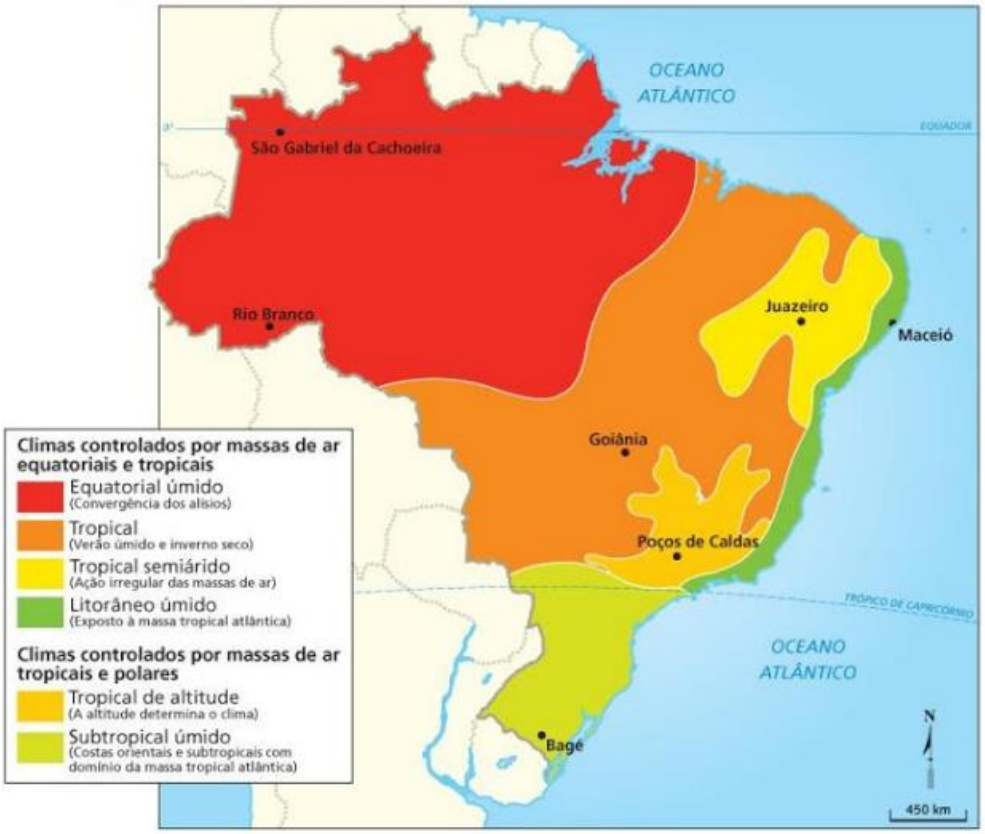
Figura 2 – AF – *Scree test* (número de componentes no eixo horizontal)



Fonte: Os autores.

Para auxiliar as interpretações, e já adiantando os resultados do trabalho, a Figura 3 logo abaixo ilustra o mapa de climas no Brasil definidos por diferentes massas de ar, segundo a classificação de Conti (Conti e Furlan, 2003).

Figura 3 – Climas no Brasil definidos por regimes de massas de ar – Classificação de Conti



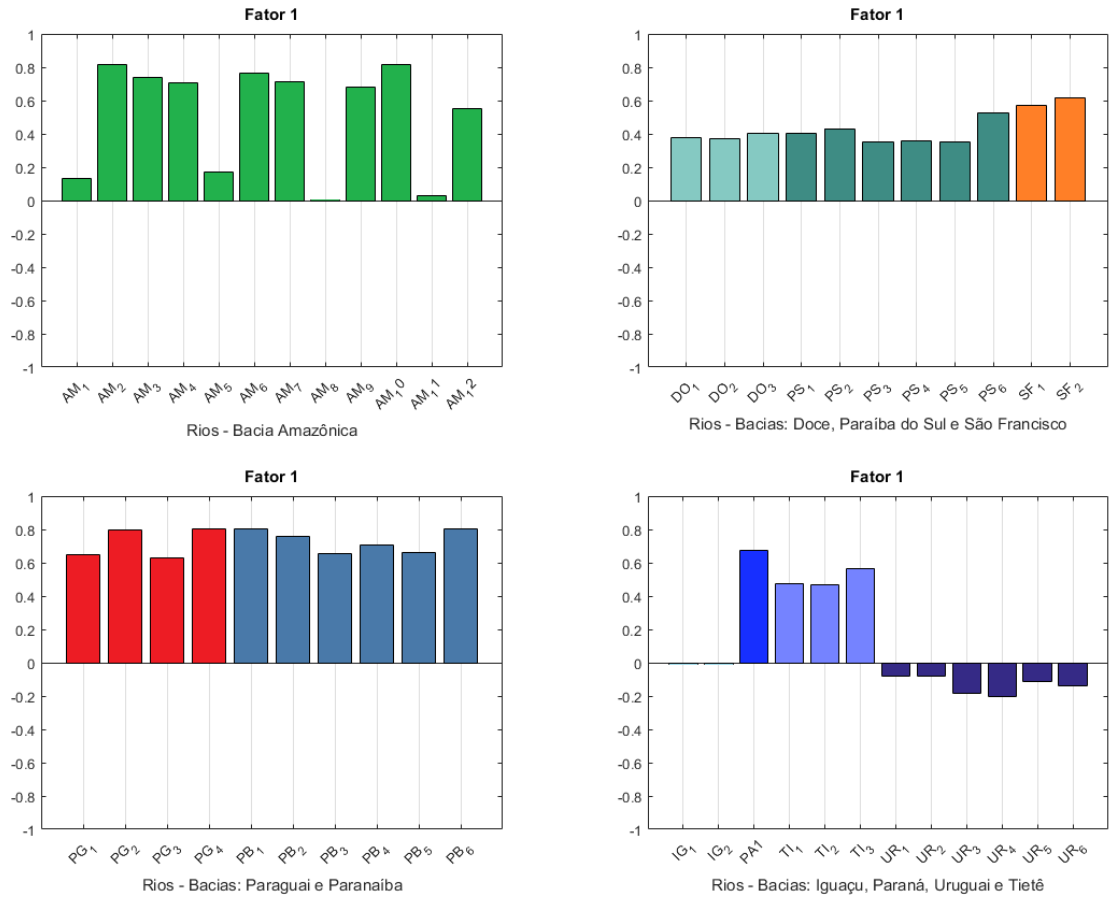
Fonte: Conti e Furlan (2003).

Para facilitar a interpretação da AF, os loadings dos fatores foram computados e disponibilizados em gráficos de barras. A Figura 4 mostra os *loadings* para o fator 1. Nota-se que os loadings são maiores para as bacias Amazônica, Paraguai e Paranaíba. Da mesma forma que ocorreu para as CPs, os rios do extremo norte da bacia Amazônica (Araguari, Curuá-Una, Jari e Uatumã) são exceções. Vale lembrar que a bacia do Rio Paranaíba está nas regiões Sudeste e Centro-Oeste, na região hidrográfica do Rio Paraná. Valores médios para os *loadings* são observados nas bacias dos rios Doce, São Francisco, Paraíba do Sul, Tietê e Paraná. Pode-se dizer que esse fator identifica os rios das regiões norte e centro-oeste, com um pouco de influência do sudeste. Pode-se fazer uma associação com o clima tropical (em laranja no mapa de climas), porém com algumas sobreposições de outros tipos de clima.

Os *loadings* para o fator 2, por sua vez, estão ilustrados na Figura 5. Eles são consideravelmente elevados para as bacias dos rios Uruguai e Iguaçu. Pode-se inferir que esses fatores identificam os rios do sul do país, que possui clima subtropical úmido. Para fins de conservação de espaço, os gráficos contendo os *loadings* dos fatores 3, 4 e 5 não serão aqui ilustrados. Os comentários sobre os *loadings* desses fatores, entretanto, seguem nos próximos parágrafos.

O fator 3, em detrimento dos primeiros, possui *loadings* maiores principalmente para a bacia do rio Paraíba do Sul. Os mesmos também são relativamente altos para as bacias dos rios Tietê e Paraná. Esses rios estão localizados nas regiões sul e sudeste do país, mais especificamente na região de clima tropical de altitude (laranja claro no mapa de climas), com alguma interseção com a região de clima subtropical úmido.

Figura 4 – AF – *Loadings* para o fator 1



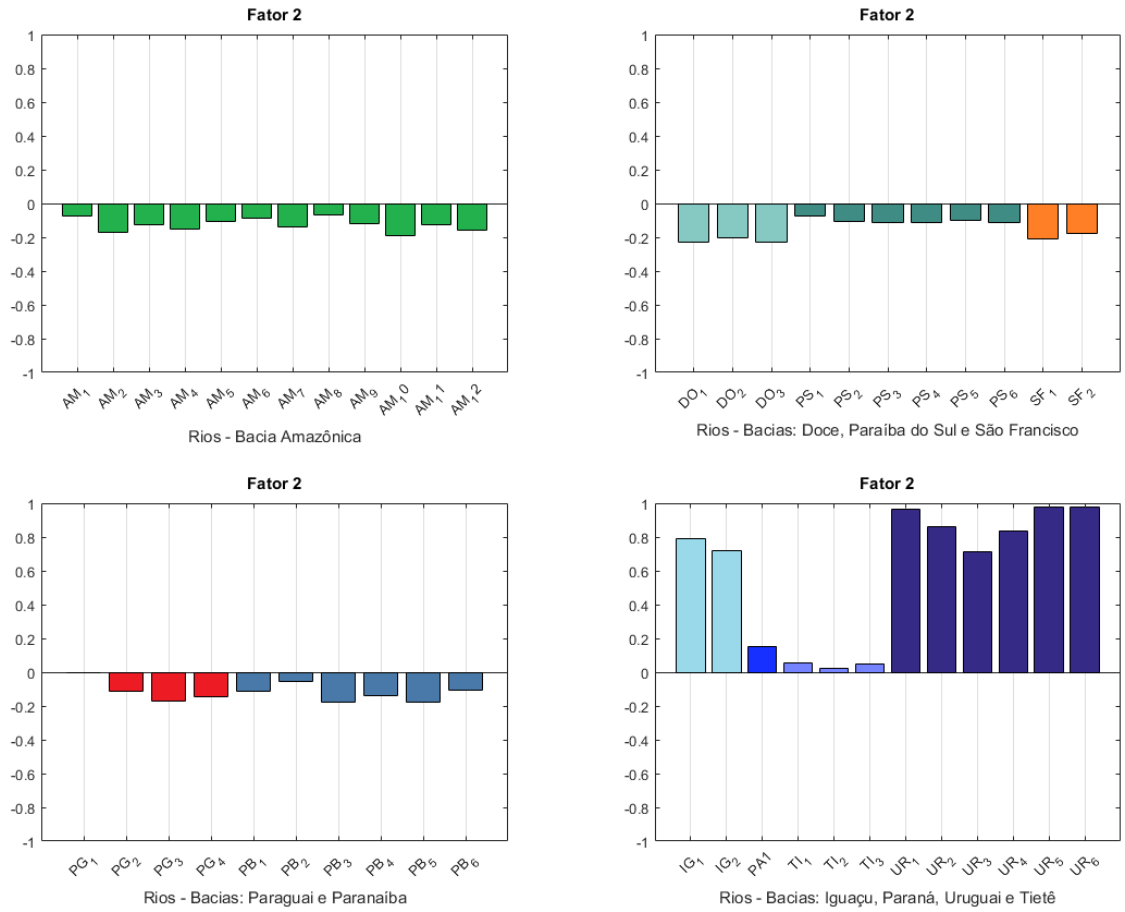
Fonte: Os autores.

O fator 4, por sua vez, concentra os maiores *loadings* na bacia do rio Doce, seguida das bacias dos rios São Francisco, Paraíba do Sul e Paraibuna. Os rios Paraibuna e Peixe

(PS_3 e PS_4) estão mais próximos do rio Doce do que os restantes dos rios considerados na bacia do Paraíba do Sul. É possível afirmar que esse fator identifica rios de bacias do Sudeste, mais ao norte, isto é, a interseção de regiões de clima tropical de altitude e clima tropical.

Finalmente, o fator 5 possui *loadings* maiores para a bacia Amazônica, com destaque para os rios do extremo norte do país.

Figura 5 – AF – *Loadings* para o fator 2



Fonte: Os autores.

Diferente das CPs, os fatores identificaram de maneira mais clara os diferentes padrões de vazão fluvial. A Tabela 4, a seguir, sintetiza os comportamentos identificados por cada fator na Análise Fatorial (AF), contrastando estes com as bacias e os padrões climáticos, segundo a classificação de Conti (Conti e Furlan, 2003). Tal classificação é bastante semelhante à de Arthur Strahler (Strahler, 1951), dado que tem por base a influência das massas de ar em áreas diferenciadas. Ela não trabalha, portanto, com as médias de chuvas e temperaturas, mas com a explicação de sua dinâmica. A diferença

entre as duas classificações reside no clima Tropical de Altitude, regime de clima controlado por massas de ar tropicais e polares que só aparece na classificação de Conti.

Em suma, observou-se que quase 80% da variação total de vazões entre rios brasileiros com potencial de aproveitamento hidrelétrico pode ser representada por um número pequeno de fatores, cujas interpretações sugerem combinações entre regimes hidrológicos das bacias e padrões climáticos brasileiros, com maior destaque para esses últimos.

Tabela 4 – AF – Fatores e interpretações

Fator	Bacias	Climas	Características Principais
1	Amazônica (exceto extremo norte), Paraguai e Paranaíba (*loadings* altos). Doce, São Francisco, Paraíba do Sul, Tietê e Paraná (*loadings* médios)	Tropical, porém com algumas sobreposições de outros tipos de clima	Centro do Brasil, quente, com verão chuvoso e inverno seco
2	Uruguai e Iguaçu	Subtropical úmido	Sul, com verão quente, inverno ameno e chuvas medianas o ano todo
3	Paraíba do Sul (*loadings* bem elevados), Tietê e Paraná	Tropical de altitude, com alguma interseção com clima subtropical úmido	Sudeste, características semelhantes ao Tropical, porém com queda maior de temperatura no inverno
4	Doce (maiores *loadings*), São Francisco, Paraíba do Sul e Paraibuna	Tropical de altitude e tropical	Vide características dos fatores 1 e 3
5	Amazônica (rios do extremo norte)	Equatorial	Norte, quente e úmido, com pequeno período seco

Fonte: Os autores.

5. Conclusões e Direções Futuras

Este trabalho buscou identificar, por meio da aplicação de diferentes técnicas de análise estatística multivariada, características comuns a diferentes rios brasileiros

utilizados para fins de aproveitamento hidrelétrico, utilizando como informação principal suas vazões. Os resultados demonstram que foi possível obter padrões distintos de comportamento dos rios, cujas especificidades estão associadas às bacias que os contemplam e, principalmente, aos regimes de macroclima sob os quais eles estão expostos. Como exemplo, destacamos que a análise de fatores possibilitou subdividir o conjunto base adotado em cinco fatores, que respondiam por grande parte da variabilidade da amostra. Foi possível associar cada fator a características comuns de regime hidrológico e de clima. Já a análise de componentes principais identificou quatro padrões comportamentais distintos que guardam, aparentemente, associações com grandes bacias hidrográficas brasileiras, muito embora as interpretações das componentes não sejam tão claras como as dos fatores.

No geral, pode-se dizer que os objetivos iniciais foram alcançados, dado que foi possível capturar grande parte da variabilidade das vazões de diversos rios brasileiros por meio de um conjunto restrito de fatores, bem como associar estes a padrões climáticos característicos, cujas especificidades se devem, em grande parte, aos diferentes regimes de massas de ar aos quais o território brasileiro está submetido.

É importante destacar que outros padrões comportamentais específicos podem ser identificados, seja por meio do aumento da amostra – por exemplo, com a inclusão de mais rios ou de pequenas bacias (que foram retiradas em detrimento da simplificação do problema inicial) – ou pela aplicação de outras técnicas de análise multivariada. Sugere-se, nesse ponto, a aplicação de diferentes métodos de clusterização, buscando-se encontrar novas formas de se agrupar os dados, bem como a adoção de técnicas específicas de discriminação de grupos (Análise de Discriminantes), de maneira a verificar se *clusters* formados por meio de diferentes métodos trazem ou não informações adicionais à divisão já pré-estabelecida.

Referências

ANA (AGÊNCIA NACIONAL DE ÁGUAS) (2013). **Conjuntura dos Recursos Hídricos no Brasil 2013**. Brasília (DF): Ministério do Meio Ambiente (MMA). Disponível em: <http://arquivos.ana.gov.br/institucional/spr/conjuntura/webSite_relatorioConjuntura/projeto/>. Acesso em: 2017.

ANDRADE, M. P.; MAGALHÃES, A.; PEREIRA, L. C. C.; FLORES-MONTES, M. J.; PARDAL, E. C.; ANDRADE, T. P.; COSTA, R. M. Effects of a La Niña event on hydrological patterns and copepod community structure in a shallow tropical estuary (Taperaçu, Northern Brazil). **Journal of Marine Systems**, 164, 128–143, 2016.

ARSLAN, O. GIS-Based Spatial-Multivariate Statistical Analysis of Water Quality Data in the Porsuk River, Turkey. **Water Quality Research Journal of Canada**, 44, 279–293, 2009.

ASSANI, A. A.; TARDIF, S.; LAJOIE, F. Statistical analysis of factors affecting the spatial variability of annual minimum flow characteristics in a cold temperate continental region (southern Québec, Canada). **Journal of Hydrology**, 328, 753–763, 2006.

BERHANU, B.; SELESHI, Y.; DEMISSE, S. S.; MELESSE, A. M. Flow Regime Classification and Hydrological Characterization: A Case Study of Ethiopian Rivers. **Water**, 7, 3149–3165, 2015.

BOENING, C.; WILLIS, J. K.; LANDERER, F. W.; NEREM, R. S.; FASULLO, J. The 2011 La Niña: so strong, the oceans fell. **Geophysical Research Letters**, 39:L19602, 2012.

CARDOSO, A. O.; CATALDI, M. Relações de índices climáticos e vazão de rios no Brasil. In: XVII CONGRESSO BRASILEIRO DE METEOROLOGIA (CBMET), P25–073; 2012, Gramado. **Anais...** Gramado: ExpoGramado, 2012. P25–073.

CONTI, J. B.; FURLAN, S. A. Geoecologia: o clima, os solos e a biota. In: ROSS, J. L. S. **Geografia do Brasil**. São Paulo: EDUSP, 2003, p. 67–237.

FERRAZ, S. E. T.; CARDOSO, A. O.; CAPOZZOLI, C. R. Variabilidade Espectral de Vazão de Rios Brasileiros. **Ciência e Natura**, Edição Especial/Novembro - VIII Brazilian Micrometeorology Workshop, 467–469, 2013.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. Editora Prentice Hall, New Jersey, 2002.

LAWLEY, D. N. A General Method for Approximating to the Distribution of Likelihood Ratio Criteria. **Biometrika**, 43, 295–303, 1956.

MARENGO, E.; GENNARO, M. C.; GIACOSA, D.; ABRIGO, C.; SAINI, G.; AVIGNONE, M. T. How chemometrics can helpfully assist in evaluating environmental data. Lagoon water. **Analytica Chimica Acta**, 317, 53–63, 1995.

NOORI, R.; KHAKPOUR, A.; OMIDVAR, B.; FAROKHNIA, A. Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. **Expert Systems with Applications**, 37, 5856–5862, 2010.

ONS (OPERADOR NACIONAL DO SISTEMA ELÉTRICO) (2015). **Atualização de séries históricas de vazões – período 1931 a 2014**. Disponível em: <http://www.ons.org.br/operacao/vazoes_ naturais.aspx>. Acesso em: 2017.

ONS (Operador Nacional do Sistema Elétrico) (2017). **Base de dados**. Disponível em <http://www.ons.org.br/operacao/vazoes_naturais.aspx>. Acesso em: 2017.

PANTSAR-KALLIO, M.; MUJUNEN, S-P.; HATZIMIHALIS, G.; KOUTOUFIDES, P.; MINKKINEN, P.; WILKIE, P. J.; CONNOR, M. A. Multivariate data analysis of key pollutants in sewage samples: A case study. **Analytica Chimica Acta**, 393, 181–191, 1999.

RAÎCHE, G.; WALLS, T. A.; MAGIS, D.; RIOPEL, M.; BLAIS, J. Non graphical solutions for the Cattell's scree test. **European Journal of Research Methods for the Behavioral and Social Sciences**, 9, 23–29, 2013.

REIS, E. **Estatística multivariada aplicada**. Edições Sílabo, Lisboa, 1997.

SIMEONOV, V.; EINAX, J. W.; STANIMIROVA, I.; KRAFT, J. Environmental modeling and interpretation of river water monitoring data. **Analytical and Bio Analytical Chemistry**, 374, 898–905,

2002.

STRAHLER, A. **Physical Geography**. Editora John Willey e Sons, New York, 1951.

TEDESCHI, R. G.; GRIMM, A. M. Variações significativas de eventos extremos de precipitação e de vazão durante episódios de El Niño e La Niña. In: XV CONGRESSO BRASILEIRO DE METEOROLOGIA (CBMET); 2008, São Paulo. **Anais...** São Paulo: Frei Caneca Convention Center, 2008.

FLOW REGIME VARIABILITY IN RUN-OF-THE-RIVER PLANTS AND ASSOCIATIONS WITH AIR MASSES

Abstract

River patterns in Brazil are sensitive to a wide range geographical and atmospheric factors and thus can vary greatly from one area of the country to another. Despite the wide variety of information concerning such factors, it is known that the flow of most Brazilian rivers is closely associated with well-defined climatic standards in the country. In this connection, a significant amount of information can be expressed in terms of a small number of elements or factors. This work, therefore, through the application of different multivariate techniques, sought to identify characteristics which are common to different rivers in Brazil, using as base information river outflows. The results indicate that the behavior of the vast majority of Brazilian rivers used for hydro-power generation can be described by a finite set of patterns which are, in turn, associated with the different types of air masses that hover over the country.

Key-words: *Multivariate Analysis; River flows; Air Masses.*