METODOLOGIA DE OPTIMIZAÇÃO DE REDES DE MONITORIZAÇÃO

Ana Sofia GONÇALVES [(1)] ; Maria Alzira SANTOS [(2)]

RESUMO

Este estudo tem por objectivo demonstrar em que medida uma análise em componentes principais pode ser utilizada para identificar grupos de estações de precipitação homogéneas, numa dada bacia hidrográfica, identificando dentro de cada grupo quais as estações redundantes.

O estudo, aplicado a 19 estações pertencentes à Bacia Hidrográfica do Rio Minho, iniciou-se pela análise à homogeneidade e aleatoriedade dos dados de precipitação anuais. Foi, no entanto, necessário proceder-se a outra análise aos dados mensais. O resultado de ambas mostrou que apenas 11 estações estavam em condições de ser utilizadas.

Da aplicação da análise em componentes principais obtiveram-se 3 grupos distintos de estações de precipitação homogéneas, tendo sido possível, para cada um deles, reduzir-se o número das respectivas estações, seleccionando-se apenas as estações mais significativas.

A metodologia apresentada neste artigo faz parte integrante de um estágio em curso no Departamento de Hidráulica do LNEC. Apresentam-se os resultados e conclusões obtidas, salientando-se o facto de que o estudo continuará e será aprofundado no decorrer do estágio.

Palavras chave: precipitação, aleatoriedade, homogeneidade, componentes principais, grupos homogéneos.

1 - INTRODUÇÃO

[(1)] Lic. em Estatística e Investigação Operacional, Estagiária - LNEC, Lisboa, Portugal
[(2)] Lic. em Matemática, M. Sc., Especialista, Investigadora Coordenadora - LNEC, Lisboa, Portugal

O número de estações necessário à caracterização do regime pluviométrico de uma bacia hidrográfica estabelece-se quando a rede é desenhada, identificando-se zonas com clima e fisiografia diferentes, de forma a instalar-se em cada uma delas um ou vários postos de medição. No entanto, ao definirem-se zonas com base nessas características, certas variações climatológicas não são detectadas. Várias razões, entre as quais se incluiem os custos de instalação e alterações induzidas pelo homem podem conduzir a que o número de estações definidas à *priori* possa ser considerado elevado ou deficiente.

Pretende-se com este artigo apresentar uma metodologia de optimização de uma rede de monitorização hidrológica. Com as ferramentas propostas é ainda possível definir regiões homogéneas em termos hidrológicos. Seleccionou-se, como caso de estudo, a bacia hidrográfica do Rio Minho.

A regionalização e a optimização duma rede de monotorização de precipitação podem possibilitar reduções de custos significativos sem diminuir a qualidade de informação necessária. A metodologia proposta é particularmente útil em estudos de caracterização da precipitação em bacias hidrográficas que apresentam um elevado número de estações. Refira-se ainda que a análise em componentes principais pode ser aplicada a outras variáveis, nomeadamente aos parâmetros que definem a qualidade água.

A utilização de séries de precipitação numa análise estatística deve ser precedida pela aplicação de alguns testes de forma a estudar o comportamento dos valores observados relativamente à aleatoriedade e homogeneidade. Utilizam-se, em geral, testes não paramétricos, pois não requerem o conhecimento da distribuição subjacente aos dados. Após a selecção das estações que satisfazem as características de homogeneidade e aleatoriedade pretendidas, utiliza-se a teoria da análise em componentes principais para identificar os grupos de estações que definem regiões homogéneas. Ao identificar-se as regiões, é importante inferir, dentro de cada grupo, quais as estações mais significativas. A escolha das estações mais significativas, dentro de cada região, é feita através de duas abordagens distintas da análise em componentes principais.

2 - ANÁLISE DE DADOS

2.1 - Bacia hidrográfica do Rio Minho

A unidade de estudo escolhida para o teste das técnicas referidas é a bacia hidrográfica do Rio Minho. Localizada entre 41º45’ N e 43º40’ N de latitude e 06º10’ W e 08º55’ W de longitude, cobre uma área de 17081 km^2 , sendo 16235 km^2 em Espanha (95%) e apenas 846 km^2 em Portugal (5%). É limitada a Oeste e a Norte pelas bacias do Rio Ulla, Tambre, Mandeo, Eume, Masma, Eo, Navia e Narcea com Espanha, e a Este e Sul pelas bacias dos rios Douro, Lima e Âncora. A altitude média da bacia é de 683 metros.

Na Figura 1 apresenta-se a rede hidrográfica da bacia portuguesa do Rio Minho, onde se podem observar as estações udométricas existentes.

Figura 1 - Mapa da Bacia Hidrográfica do Rio Minho (Fonte: INAG)

No Quadro 1 apresentam-se as estações e algumas das suas características. É de referir que quatro das estações utilizadas não se encontram representadas no mapa, pois estão localizadas na proximidade e não no interior da bacia.

Quadro 1
Relação dos postos da rede udométrica

Estação	Código da estação	Altitude (m)	Latitude (N)	Longitude (W)	Períodos de exploração
Friestas	01F/01	20	42º 03’	08º 34’	1/10/80 a 1/9/95
Valença	01F/02	20	42º 01’	08º 39’	1/10/79 a 1/9/95
Parada	01G/02	255	42º 01’	08º 28’	1/10/80 a 1/9/95
Merufe	01G/03	160	42º 01’	08º 23’	1/10/80 a 1/9/95
Fonte Boa	01H/03	745	42º 01’	08º 19’	1/10/60 a 1/9/95
Melgaço	01H/04	220	42º 07’	08º 15’	1/10/79 a 1/9/95
Cubalhão	01H/05	770	42º 03’	08º 15’	1/10/80 a 1/9/95
Fiães	01H/06	705	42º 06’	08º 13’	1/10/80 a 1/9/95
Vila Nova de Cerveira	02E/02	205	41º 56’	08º 44’	1/10/79 a 1/9/95
Arga de Baixo	02E/03	470	41º 51’	08º 43’	1/10/80 a 1/9/95
Caminha	02E/04	10	41º 52’	08º 50’	1/10/79 a 1/9/95
Sapardos	02F/01	245	41º 56’	08º 40’	1/10/60 a 1/9/95
Cerdeira	02F/02	600	41º 53’	08º 34’	1/10/60 a 1/9/95
Chã de Virialho	02F/04	670	41º 59’	08º 34’	1/10/79 a 1/9/95
Extremo	02G/03	380	41º 58’	08º 28’	1/10/60 a 1/9/95
Lamas de Mouro*	02G/13	870	42º 03’	08º 11’	1/9/80 a 1/5/90
Aspra*	03D/02	20	41º 48’	08º 50’	1/10/79 a 1/9/95
Valadares*	03E/04	260	41º 47’	08º 46’	1/4/79 a 1/9/95
Montaria*	03E/05	285	41º 48’	08º 44’	1/7/80 a 1/9/95

* Estações que se encontram fora da Bacia.

Os dados de precipitação foram cedidos pelo Instituto da Água (INAG), através do Sistema Nacional de Informação de Recursos Hídricos (SNIRH).

2.2 - Precipitações anuais

Ao utilizarem-se séries de observações de fenómenos hidrológicos deve-se ter cuidado relativamente ao comportamento destas. É desejável que as séries a utilizar sejam, do ponto de vista estatístico, homogéneas e aleatórias. Considera-se que uma série é homogénea quando os factores que condicionam o fenómeno em causa se mantêm constantes ao longo do período de observação. As características acima referidas podem ser avaliadas a partir de um conjunto de testes estatísticos não paramétricos, ou seja, testes independentes da distribuição subjacente à população.

Neste estudo escolheu-se a precipitação como fenómeno de teste. Numa primeira análise da homogeneidade e aleatoriedade das séries consideraram-se os valores anuais de precipitação e não os valores mensais, pois uma eventual quebra de homogeneidade nas séries de precipitação mensais reflecte-se nas séries de precipitação anuais.

Verificando-se a ocorrência de algumas falhas nos registos disponíveis procurou-se preenchê-las a partir da média dos valores de precipitação, correspondentes ao mês em falta, observados nos restantes anos da série. Os postos com o período de observações mais longo (35 anos) são Fonte Boa, Sapardos, Cerdeira e Extremo. A estação de Lamas de Mouro possui um período de funcionamento de 10 anos e as restantes estações variam entre os 15 e 17 anos.

Para verificar a aleatoriedade submeteram-se as séries de precipitação anual aos testes dos chorrilhos (estatísticas Z_1 e Z_2). Utilizaram-se ainda os testes de desvio acumulado da média (estatísticas P e R) e Bayesiano (estatísticas U e A) com o objectivo de estudar a homogeneidade (SINGH, 1989; LINDGREN, 1976; TAYLOR, 1990). Os testes referidos anteriormente encontram-se implementados na aplicação STATWAT (OLIVEIRA, 1996) que funciona como um "*add-in*" ao MSExcel. Assim, de uma forma rápida e eficiente, foi possível seleccionar as estações cujos dados satisfazem a homogeneidade e aleatoriedade desejada.

Apresentam-se nos Quadros 2 e 3 os resultados dos testes e as falhas verificadas, para um nível de significância de 5%, respectivamente.

Quadro 2

Resultados dos testes de aleatoriedade e homogeneidade

	Estações																		
	Fri.	Val.	Par.	Mer.	F. B.	Mel.	Cub.	Fiães	V.N.C	A. B.	Cam.	Sap.	Cer.	C. V.	Ext.	L. M.	Aspra	Val.	Mon.
Média	1244	1241	1290	1611	1912	1267	1476	1114	1613	2037	1274	1896	2509	1898	2247	2372	1277	1770	1852
Máximo	1761	1606	2112	2477	3.283	1723	2014	1579	2207	3287	1626	2878	4023	2527	3899	3157	1746	2644	2672
n	15	16	15	15	35	16	15	15	16	15	16	35	35	16	35	10	16	16	15
Z_1	-0.28	-0.77	-1.39	-0.83	-2.61	-0.26	0.28	1.39	-0.26	-1.94	-0.25	0.52	-1.22	0.77	-2.26	-0.33	1.29	-0.77	-1.94
Z_2	-1.04	-1.46	-2.08	-1.04	-2.71	-0.49	0	1.04	-0.49	-2.08	-0.49	0.68	-1.35	0.49	-2.03	-1.22	1.46	-0.49	-2.08
P	3.69	3.57	4.11	3.1	7.36	3.01	2.49	2.63	2.8	4.95	2.82	3.67	7.76	2.68	8.25	2.13	2.65	5.25	5
R	5.04	5.63	5.89	4.5	7.36	3.01	3.22	5.1	4.28	6.61	4.57	4.89	7.76	4.23	8.86	3.47	4.82	6.68	6.27
U	0.23	0.29	0.25	0.17	0.52	0.25	0.11	0.12	0.06	0.43	0.1	0.09	0.79	0.11	0.65	0.13	0.09	0.39	0.4
A	1.27	1.69	1.78	0.88	3.28	1.71	0.87	0.77	0.36	2.02	0.59	0.53	4.23	0.8	3.17	0.8	0.56	1.85	1.87

Quadro 3

Falhas verificadas nos testes para um nível de significância de 5%

	Estações																		
	Fri.	Val.	Par.	Mer.	F. B.	Mel.	Cub.	Fiães	V.N.C	A. B.	Cam.	Sap.	Cer.	C. V.	Ext.	L. M.	Aspra	Val.	Mon.
Z_1	-	-	-	-	F	-	-	-	-	-	-	-	-	-	F	-	-	-	-
Z_2	-	-	F	-	F	-	-	-	-	F	-	-	-	-	F	-	-	-	F
P	-	-	-	-	F	-	-	-	-	F	-	-	F	-	F	-	-	F	F
R	F	F	F	-	-	-	-	F	-	F	-	-	-	-	-	-	-	F	F
U	-	-	-	-	F	-	-	-	-	F	-	-	F	-	F	-	-	-	-
A	-	-	-	-	F	-	-	-	-	-	-	-	F	-	F	-	-	-	-

Com base na análise do Quadro 3 optou-se por não considerar as estações Fonte Boa (F.B.), Arga de Baixo (A.B.), Extremo (Ext.) e Montaria (Mon.) na continuação do estudo, pois as hipóteses nulas relativas a estas estações são na sua maior parte rejeitadas. Apesar da estação Lamas de Mouro (L.M.) não ser rejeitada pelos testes utilizados, não foi considerada para a análise seguinte por o período de observação ser muito curto.

As estações Caminha (Cam.) e Aspra além de se encontrarem muito próximas apresentam um comportamento bastante semelhante. Por essa razão e pelo facto da estação Caminha satisfazer a hipótese de aleatoriedade e a estação Aspra se encontrar fora da bacia hidrográfica, não se utilizará a estação Aspra no estudo seguinte.

Para continuar o estudo é necessário que as estações a considerar tenham período comum de observações. Como se pode observar no Quadro 1, o período de exploração comum às séries seleccionadas está compreendido entre Outubro de 1980 e Setembro de 1995, correspondendo a 15 anos de observação. Atendendo a que 15 anos de observações é um número muito reduzido para uma qualquer análise estatística, em particular na análise aplicada na regionalização, optou-se por utilizar séries de precipitação mensais, tentando-se assim obter mais informação relativamente aos dados. É este o objectivo da secção 2.3.

2.3 - Precipitações mensais

As séries a analisar correspondem aos valores mensais de precipitação para o período compreendido entre Outubro de 1980 e Setembro de 1995, das estações seleccionadas na análise anterior (13 estações).

As séries mensais de precipitação são caracterizadas por um padrão sazonal bastante nítido. Por existir esta dependência entre os dados, surge a necessidade de estimar e retirar a sazonalidade. Qualquer função pode ser representadas através das suas componentes harmónicas. Neste estudo recorreu-se à análise de Fourier para modelar a componente periódica dos dados. Depois de encontrado o modelo a ajustar, calcularam-se os respectivos resíduos e testou-se a sua aleatoriedade.

Através do periodograma[(1)], para cada estação, verificou-se que as precipitações mensais são razoavelmente explicadas por componentes periódicas de 12 e 6 meses (associadas às 15ª e 30ª frequência, respectivamente). Decidiu-se então ajustar o modelo seguinte às séries de precipitação mensais:

[(1)] Chama-se Periodograma à função dos valores $I(w_j)$ definida por

$I(w_j) = \frac{N}{2}\left[a^2(w_j) + b^2(w_j)\right], j = 0, \ldots, \frac{N}{2}$ com $a(w_j)$ e $b(w_j)$ coeficientes da representação de Fourier.

$$\hat{y}_t = a_0 + \sum_{j=15,30} \left[a_j \cos(w_j t) + b_j sen(w_j t) \right] \quad (1)$$

em que :

$\hat{y}_t$ - estimativa da precipitação mensal no instante t;

a_0 - média da série observada;

a_j e b_j - coeficientes da representação de Fourier;

ω_{15} = 0.524 (ou seja, periodicidade de 12 meses, $\frac{2\pi \times 15}{180}$);

ω_{30} = 1.047 (ou seja, periodicidade de 6 meses, $\frac{2\pi \times 30}{180}$).

Após se terem obtido os valores do modelo ajustado, calcularam-se os resíduos, isto é a função $r_t = y_t - \hat{y}_t$, e é com base nestes novos valores que se prossegue o estudo. Para que seja possível inferir credibilidade às séries agora em estudo, aplicaram-se os testes dos chorrilhos, com o objectivo de verificar a aleatoriedade dos dados. Para a maioria dos postos a hipótese de aleatoriedade não foi rejeitada, sendo as únicas excepções os postos de Fiães e Valadares.

Após estas duas análises, constata-se que apenas 11 estações, das 19 consideradas inicialmente, estão em condições de serem utilizadas na análise seguinte.

3 - REGIONALIZAÇÃO E OPTIMIZAÇÃO DA REDE

3.1 - Metodologia

O objectivo principal desta secção é mostrar como a análise em componentes principais pode ser utilizada para identificar grupos de estações de precipitação homogéneas, isto é, grupos cujas estações apresentam um elevado grau de associação. A teoria da análise em componentes principais não pode ser descrita com detalhe neste artigo, podendo-se encontrar em MARDIA *et al* (1979) uma descrição adequada da técnica a aplicar.

Considere-se um grupo de *p* variáveis, cada uma formada por *n* observações, representadas por um plano *p*-dimensional. O objectivo de uma análise em componentes principais é transformar o plano *p*-dimensional, referido anteriormente, num novo plano, em que os eixos sejam ortogonais (para que as projecções das observações iniciais em cada um dos novos eixos, chamados componentes principais, sejam independentes) e as componentes principais sejam encontradas em ordem decrescente de importância, de forma a que cada componente explique o máximo possível da variância não explicada pelas componentes anteriores.

Para obter as componentes principais constrói-se a matriz *nxp* e calculam-se os respectivos valores próprios, l_i. Os *p* valores próprios encontrados dão-nos a conhecer a proporção de variância explicada por cada uma das componentes sendo este para a componente principal *i* igual a l_i/p. Considerando apenas as *k* primeiras componentes principais ($k < p$) consegue-se reduzir o sistema *p*-dimensional para um *k*- dimensional, com uma perda mínima de informação. O número de componentes a reter não é definido com rigor. Em geral, a escolha baseia-se no total de variância pretendida, obtida a partir das *k* componentes. Uma vez escolhido o número de componentes a ter em conta, calculam-se os coeficientes de correlação, r_{ji}, entre a variável *j* e a componente *i*:

$$r_{ji} = v_{ji} \lambda_i^{1/2} \quad (2)$$

em que:

v_{ji} - componente *j* do vector próprio associado ao valor próprio *i*;
λ_i - valor próprio associado à componente principal *i*.
A análise destes coeficientes possibilita uma valiosa informação relativamente às relações entre as variáveis.

Dado a variância explicada pela primeira componente principal ser máxima, as variáveis apresentam elevados coeficientes de correlação com esta e valores muito reduzidos com as restantes componentes, tornando-se difícil a interpretação destes. Por esta razão, recorre-se a uma rotação dos eixos, rotação *Varimax* Normalizada, que permite distribuir, de forma mais equitativa, a variância pelas componentes, sem, no entanto, afectar o total de variância explicada. Obtém-se assim uma estrutura mais simples, que possibilita uma melhor e mais fácil interpretação das componentes. Determinam-se, como anteriormente, os coeficientes de correlação entre as variáveis e as novas componentes principais, e é a partir destes que se procede ao agrupamento das estações. Este agrupamento é feito com base no maior coeficiente de correlação (em valor absoluto).

3.2 - Aplicação à Bacia do Rio Minho

3.2.1 - Regionalização da rede de monitorização

Como foi referido na secção anterior, os dados em estudo pertencem a 11 estações, sendo cada série formada pelos resíduos (obtidos na secção 2.2) da série de precipitação mensais. O *software* utilizado na análise das componentes principais foi o STATISTICA.

Apresentam-se no Quadro 4 os valores próprios e a percentagem de variância explicada por cada um deles; na última coluna, apresentam-se ainda as percentagens acumuladas. A proporção da variância explicada é de extrema utilidade na identificação das componentes consideradas importantes.

Quadro 4
Valores próprios e variância explicada

	λ_i	% var expl.	% var ac.
1	9.545	86.772	86.772
2	0.375	3.405	90.177
3	0.350	3.183	93.360
4	0.237	2.153	95.512
5	0.158	1.433	96.945
6	0.093	0.848	97.793
7	0.079	0.722	98.515
8	0.055	0.504	99.020
9	0.047	0.425	99.445
10	0.034	0.311	99.755
11	0.027	0.245	100.000

Da análise do quadro anterior, deduz-se que apenas três componentes explicam cerca de 93% da variância, isto é, consegue-se reduzir o número de componentes com uma perda mínima de informação. Neste estudo retiveram-se apenas três componentes, dado o total de variância explicada ser já elevado. No entanto, refira-se que mais ou menos componentes poderiam ter sido utilizadas. No Quadro 5 apresentam-se os coeficientes de correlação entre os resíduos da precipitação mensal em cada uma das estações e as três primeiras componentes

principais. É difícil, a partir da análise do Quadro 5, identificar estações com comportamento semelhante, pois como foi referido anteriormente, as variáveis são fortemente correlacionadas com a primeira componente, apresentando com as restantes valores reduzidos. Contudo, a representação gráfica da 2ª componente *versus* 3ª componente, possibilita já, de uma forma algo arbitrária, identificar grupos de estações com comportamento análogo (Figura 2).

Quadro 5

Coeficientes de correlação entre as estações e as componentes principais que explicam cerca de 93% da variância total

	1ª Comp. Principal	2ª Comp. Principal	3ª Comp. Principal
Friestas	0.969	0.128	-0.018
Valença	0.921	0.160	-0.034
Parada	0.841	0.155	-0.498
Merufe	0.880	-0.291	-0.155
Melgaço	0.936	-0.202	0.059
Cubalhão	0.891	-0.374	-0.004
Vila N. C.	0.961	0.061	0.151
Caminha	0.945	0.171	0.159
Sapardos	0.980	0.029	0.104
Cerdeira	0.933	0.064	0.107
Chã de Vir.	0.977	0.070	0.056
λ_i	9.545	0.375	0.350
% var expl.	86.772	3.405	3.183

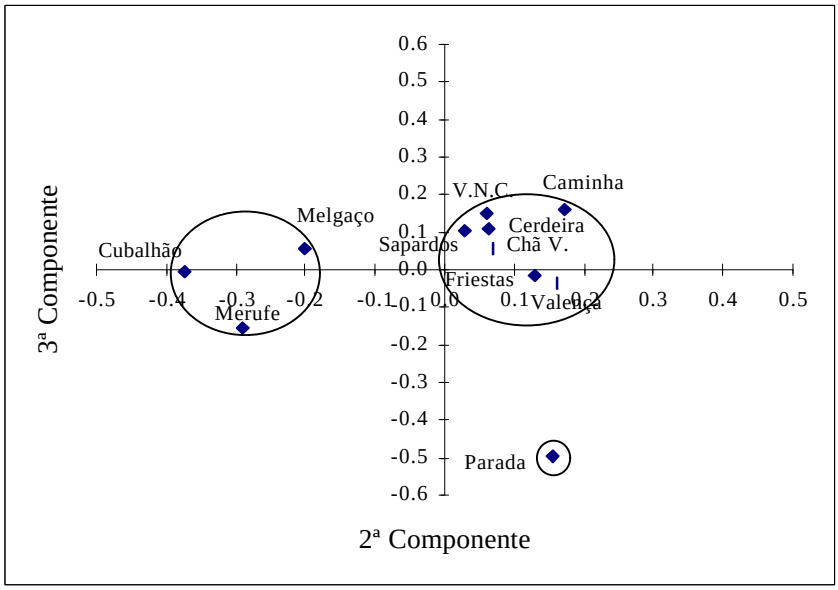

Figura 2 - Representação gráfica dos coeficientes de correlação entre as 11 estações e as 2ª e 3ª componentes

Da análise da figura anterior constata-se a existência de três grupos distintos:

- estações Cubalhão, Melgaço e Merufe;

- estações Friestas, Valença, Vila Nova de Cerveira, Caminha, Sapardos, Cerdeira, Chã de Virialho;
- estação Parada.

Uma forma de definir os grupos mais objectivamente é rodando os eixos, sendo possível retirar alguma ambiguidade existente na análise anterior e ainda aumentar a informação relativamente aos dados. É de esperar que os grupos obtidos através da análise gráfica coincidam com os obtidos na análise das componentes após aplicada a rotação. No Quadro 6 apresentam-se os coeficientes de correlação entre as estações e as componentes principais, após aplicada a rotação *Varimax* Normalizada. Agruparam-se as estações com base no maior coeficiente de correlação; as estações que apresentam um maior coeficiente de correlação com o primeiro factor formam um grupo, e as estações que, por sua vez, apresentam maior coeficiente de correlação com o segundo factor, constituem outro grupo e assim sucessivamente.

Quadro 6

Coeficientes de correlação entre as estações e as componentes principais que explicam cerca de 93% da variância total, depois de aplicada a rotação *Varimax* Normalizada

	1ª Comp. Principal	2ª Comp. Principal	3ª Comp. Principal
Friestas	0.734	0.438	0.475
Valença	0.707	0.385	0.477
Parada	0.402	0.350	0.834
Merufe	0.401	0.737	0.424
Melgaço	0.596	0.692	0.295
Cubalhão	0.450	0.810	0.275
Vila N. C.	0.786	0.488	0.307
Caminha	0.832	0.387	0.327
Sapardos	0.760	0.525	0.346
Cerdeira	0.744	0.470	0.333
Chã de Vir.	0.752	0.490	0.398
λ_i	4.924	3.264	2.081
% var expl.	44.763	29.674	18.922

Como se constata da análise do Quadro anterior as estações Friestas, Valença, Vila Nova de Cerveira, Caminha, Sapardos, Cerdeira, e ainda Chã de Virialho apresentam os maiores coeficientes de correlação com o primeiro factor. Os mais elevados coeficientes de correlação das estações Melgaço, Cubalhão e Merufe surgem no segundo factor. A estação Parada apresenta o seu maior coeficiente de correlação com o terceiro factor. Os seguintes grupos[(1)]foram então definidos:

Grupo 1 - estações Friestas, Valença, Vila Nova de Cerveira, Caminha, Sapardos, Cerdeira, e Chã de Virialho;

Grupo 2 - estações Cubalhão, Melgaço e Merufe;

Grupo 3 - estação Parada.

Como se verifica, os grupos obtidos pela análise das componentes principais após aplicada uma rotação dos eixos, coincidem com os obtidos graficamente.

[(1)] Os sublinhados na coluna de cada componente principal definem as estações pertencentes a cada um dos grupos.

Representando-se geograficamente, na Figura 3, os grupos obtidos, é possível considerar, na Bacia Hidrográfica do Rio Minho, três regiões homogéneas.

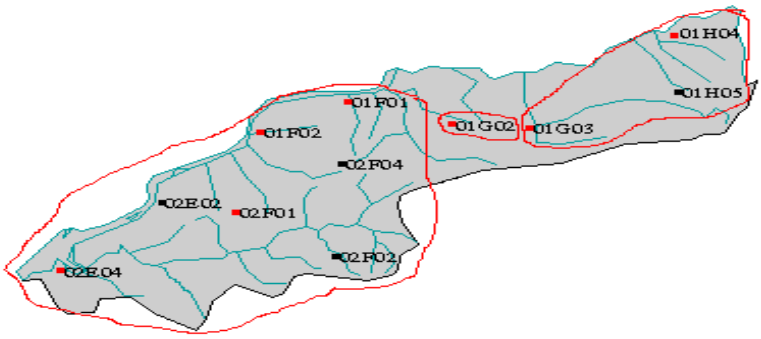

Figura 3 - Representação geográfica dos grupos

3.2.2 - Optimização da rede de monotorização

Aplicou-se a cada região de estações homogéneas a teoria das componentes principais com o objectivo de identificar qual a estação ou estações que melhor a caracterizam. Apresentam-se duas metodologias para a escolha das estações mais significativas.

O primeiro método a ser aplicado consiste em considerar os factores que explicam cerca de 95% da variância e associar a cada um destes a estação, entre aquelas que não foram seleccionadas previamente, que apresenta maior valor de correlação. A mesma análise deve ser feita aplicando-se a rotação aos eixos, para que não seja possível desprezarem-se estações importantes (G. GALEATI *et al*, 1986). Seleccionaram-se como representativas da região, as estações mais fortemente correlacionadas com os factores considerados.

O método utilizado por DYER (1975) baseia-se nos valores próprios. A estação que apresenta maior coeficiente de correlação associado ao menor valor próprio é desprezada. Analisam-se de seguida os coeficientes de correlação associados ao segundo menor valor próprio desprezando-se outra estação, repete-se esta análise até ter-se encontrado os valores próprios superiores a 0.6. As estações que não foram eliminadas são submetidas novamente ao método descrito, e repete-se este procedimento até que não seja possível reduzir o número de estações. Quando o processo termina para uma região as estações retidas são aquelas consideradas capazes de descrever as estações como um todo nessa região.

O primeiro grupo a ser analisado foi o formado pelas estações: Friestas, Valença, Vila Nova de Cerveira, Caminha, Sapardos, Cerdeira e Chã de Virialho. O número de componentes necessário para que se consiga explicar 95% da variância total é dois como se pode ver através do Quadro 7.

Com base na análise do Quadro 8, os coeficientes de correlação, entre as estações e as primeiras duas componentes, mais elevados verificam-se nas estações: Sapardos e Valença. A mesma análise foi repetida utilizando-se a rotação *Varimax* Normalizada de forma a que não se rejeitem estações significativas não encontradas na análise sem rotação. Os resultados são apresentados no Quadro 9.

Quadro 7

Valores próprios e percentagens acumuladas da variância explicada
por cada componente principal

	λ_i	% var expl.	% var ac.
1	6.502	92.882	92.882
2	0.183	2.621	95.503
3	0.119	1.693	97.196
4	0.071	1.019	98.215
5	0.056	0.803	99.018
6	0.039	0.555	99.573
7	0.030	0.427	100.000

Quadro 8

Coeficientes de correlação entre as estações e as componentes principais que explicam cerca de 95% da variância total

	1ª Comp. Principal	2ª Comp. Principal
Friestas	0.973	-0.099
Valença	0.929	-0.350[(2)]
Vila N. C.	0.970	0.083
Caminha	0.965	0.078
Sapardos	0.984[(2)]	0.031
Cerdeira	0.943	0.181
Chã de Vir.	0.981	0.066
λ_i	6.502	0.183
% var expl.	92.882	2.621

Quadro 9

Coeficientes de correlação entre as estações e as componentes principais que explicam cerca de 95% da variância total, depois de aplicada a rotação *Varimax* Normalizada

	1ª Comp. Principal	2ª Comp. Principal
Friestas	0.692	0.692
Valença	0.499	0.858[(2)]
Vila N. C.	0.804	0.548
Caminha	0.797	0.549
Sapardos	0.782	0.598
Cerdeira	0.845[(2)]	0.456
Chã de Vir.	0.802	0.568
λ_i	3.980	2.705
% var expl.	56.859	38.644

Como se constata da análise dos coeficientes de correlação, após aplicada a rotação (Quadro 9), as estações Cerdeira e Valença são consideradas as mais significativas. Assim, as estações consideradas significativas, para a descrição da precipitação nesta região, a partir deste método são Sapardos, Valença e Cerdeira.

No Quadro 10 apresentam-se os coeficientes de correlação entre as estações e todas as componentes após aplicada a rotação *Varimax* Normalizada, de forma a encontrarem-se as estações mais significativas através do segundo método descrito. Analisaram-se os valores

[(2)] Estações mais significativas.

próprios menores de 0.6, de forma crescente, e identificou-se, para cada um deles, a estação que apresentava o maior coeficiente de correlação. As estações consideradas significativas por este método são: Valença, Vila Nova de Cerveira, Caminha e Cerdeira. Submetemos novamente as estações não rejeitadas à análise anterior, contudo não se conseguiu encontrar nenhuma estação redundante (Quadro 11), pois todos os valores próprios são superiores a 0.6.

Quadro 10

Coeficientes de correlação entre as estações e todas as componentes principais, depois de aplicada a rotação *Varimax* Normalizada

	1ª C. Pr.	2ª C. Pr.	3ª C. Pr.	4ª C. Pr.	5ª C. Pr.	6ª C. Pr.	7ª C. Pr.
Friestas	0.343	0.552	0.431	0.463(3)	0.391	0.134	0.079
Valença	0.294	0.802	0.360	0.175	0.306	0.113	0.067
Vila N. C.	0.385	0.432	0.440	0.214	0.641	0.111	0.049
Caminha	0.629	0.434	0.439	0.210	0.398	0.120	0.076
Sapardos	0.375	0.482	0.482	0.252	0.482	0.164	0.268(3)
Cerdeira	0.315	0.396	0.763	0.195	0.330	0.108	0.066
Chã de Vir.	0.395	0.447	0.493	0.278	0.445	0.342(3)	0.092
λ_i	1.143	1.912	1.758	0.514	1.356	0.213	0.104
% var expl.	16.334	27.319	25.116	7.338	19.366	3.046	1.481

Quadro 11

Coeficientes de correlação entre as estações não rejeitadas e todas as componentes principais, depois de aplicada a rotação *Varimax* Normalizada

	1ª Comp. Principal	2ª Comp. Principal	3ª Comp. Principal	4ª Comp. Principal
Valença	0.376	0.813	0.316	0.314
Vila N. C.	0.457	0.444	0.411	0.652
Caminha	0.454	0.446	0.655	0.407
Cerdeira	0.778	0.406	0.340	0.338
λ_i	1.161	1.223	0.813	0.803
% var expl.	29.026	30.563	20.331	20.080

As estações Cerdeira e Valença são consideradas importantes nos dois métodos não surgindo qualquer dúvida em mantê-las. A estação Caminha, apesar de só ser considerada importante no segundo método é escolhida como uma das estações necessárias para a correcta descrição da precipitação, pois caracteriza a precipitação na parte mais oeste da região. O problema surge relativamente às estações Sapardos e Vila Nova de Cerveira, pois foram consideradas representativas apenas por um dos métodos, o primeiro e o segundo método, respectivamente, e no entanto estas duas estações encontram-se à mesma altitude e apresentam características semelhantes. Por estas razões optou-se por considerar apenas uma delas. Escolheu-se a estação Vila Nova de Cerveira devido a localização geográfica que apresenta.
Assim, as estações que interessa manter para a caracterização da precipitação nesta região são Cerdeira, Valença, Caminha e Vila Nova de Cerveira. Através da Figura 4 pode-se ver como as estações, que melhor representam a precipitação, estão distribuídas de uma forma homogénea ao longo da região.

(3) Estações rejeitadas.

Encontradas as estações representativas do primeiro grupo passa-se à análise do grupo formado pelas estações Merufe, Melgaço e Cubalhão. De forma a explicar-se 95% do total da variância é necessário considerar-se duas componentes (Quadro 12). A análise do Quadro 13 e 14 não é muito conclusiva, sendo difícil optar entre as estações de Melgaço e Cubalhão. Por outro lado, o 2º método, cujo resultado é apresentado no Quadro 15, permite inferir não existirem estações supérfulas. Decide-se então, reter todas as estações pois todas parecem contribuir para a definição da precipitação nessa região.

Quadro 12

Valores próprios e percentagens acumuladas da variância explicada por cada componente principal

Grupo 2	λ_i	% var ac.
1	2.703	90.108
2	0.200	96.791
3	0.096	100.000

Quadro 13

Coeficientes de correlação entre as estações e as componentes principais que explicam cerca de 95% da variância total

	1ª Comp. Principal	2ª Comp. Principal
Merufe	0.9293	-0.3694[(4)]
Melgaço	0.9593[(4)]	0.1765
Cubalhão	0.9588	0.1814
λ_i	2.7033	0.2005
% var expl.	90.11	6.68

Quadro 14

Coeficientes de correlação entre as estações e as componentes principais que explicam cerca de 95% da variância total, depois de aplicada a rotação *Varimax* Normalizada

	1ª Comp. Principal	2ª Comp. Principal
Merufe	0.4829	0.8757[(4)]
Melgaço	0.8529	0.4733
Cubalhão	0.8557[(4)]	0.4691
λ_i	1.6929	1.2108
% var expl.	56.4302	40.3610

Quadro 15

Coeficientes de correlação entre as estações e todas as componentes principais, depois de aplicada a rotação *Varimax* Normalizada

[(4)] Estações mais significativas.

	1ª Comp. Principal	2ª Comp. Principal	3ª Comp. Principal
Merufe	0.370	0.853	0.369
Melgaço	0.465	0.433	0.772
Cubalhão	0.775	0.431	0.462
λ_i	0.954	1.101	0.945
% var expl.	31.799	36.697	31.504

Na Figura 4 apresenta-se as estações representativas de cada região.

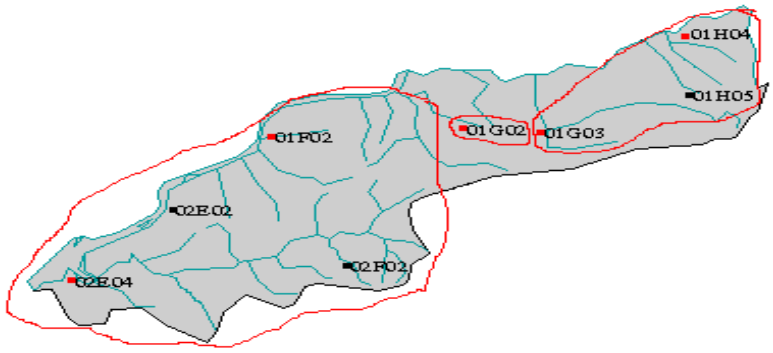

Figura 4 - Representação das estações significativas dentro de cada grupo

4 - CONCLUSÕES

Este estudo permitiu mostrar como se pode utilizar, de forma interessante, a análise em componentes principais para agrupar e optimizar uma rede de estações de precipitação.

Feita uma análise preliminar à aleatoriedade e homogeneidade dos dados de precipitação anual verificou-se que apenas 14 estações, dentre as 19 iniciais, satisfaziam as características pretendidas. Esta análise foi baseada nos dados mensais dado o reduzido número de anos disponíveis. Dado as séries de dados mensais apresentarem uma forte sazonalidade utilizou-se a análise de Fourier para a extrair. Aplicaram-se os testes de aleatoriedade aos resíduos e verificou-se que a hipótese de aleatoriedade foi rejeitada para as estações de Fiães e Valadares.

Através da análise em componentes principais das 11 estações resultantes obtiveram-se os seguintes grupos de estações homogéneas:

Grupo 1 - Friestas, Valença, Vila Nova de Cerveira, Caminha, Sapardos e Cerdeira, Chã de Virialho;

Grupo 2 - Cubalhão, Melgaço e Merufe;

Grupo 3 - Parada.

As estações mais significativas de cada grupo foram identificadas aplicando, de novo, a cada um dos grupos, a análise em componentes principais. No primeiro grupo de estações as consideradas mais importantes, tendo em conta as duas metodologias referidas, foram nomeadamente: Valença, Vila Nova de Cerveira, Cerdeira e Caminha. Relativamente ao grupo dois, não foi possível reduzir o número de estações. As três estações que constituem este grupo são necessárias para a correcta descrição da precipitação nessa região. As estações que não foram desprezadas devem ser controladas de forma a manter a qualidade dos respectivos dados.

A metodologia utilizada neste artigo pode ser aplicada a qualquer outra bacia, e serve ainda para estudar outros fenómenos.

O estudo exposto neste artigo será continuado e aprofundado no decorrer do estágio que se encontra a realizar no Departamento de Hidráulica do LNEC.

AGRADECIMENTOS

Agradece-se ao INAG (Direcção de Serviços de Recursos Hídricos) a disponibilização dos dados utilizados neste estudo. Agradece-se ainda à Eng. Elsa Alves, do Núcleo de Hidráulica de Estruturas, o apoio e a valiosa troca de informação no decorrer do estudo.

BIBLIOGRAFIA

DYER, T.G.J. - "Assignment of rainfall stations into homogeneous groups: An application of principal component analysis". Quart. J. Roy. Meteorol. Soc., **101**, 1975, pp. 1005-1013.

GALEATI, G.; ROSSI, G.; PINI, G.; ZILLI, G. - "Optimization of a snow network by multivariate statistical analysis". Journal des Sciences Hydrologiques, **31**, 1986, pp. 93-108.

JOLLIFFE, I.T. – "Discarding Variables in a Principal Component Analysis. I: Artificial Data". Applied Statistics, **21**, 1972, pp. 160-173.

LINDGREN, B.W. - *Statistical Theory*. New York, Macmillan, 1976.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. - *Multivariate Analysis*. London, Academic Press, 1979.

MORIN, G.; FORTIN, J.; SOCHANSKA, W.; LARDEAU, J. - "Use of principal component analysis to identify homogeneous precipitation for optimal Interpolation". Water Resources Research, **15**, 6, Dezembro 1979, pp. 1841-1850.

MURTEIRA, B.; MÜLLER, D.; TURKAM, K. - *Análise de Sucessões Cronológicas*. Lisboa, McGRAW-HILL, 1993.

OLIVEIRA, R. - STATWAT- Statistics for Water Resources - Manual de Utilização, versão 1.0. Lisboa, 1997.

SINGH, V.P. - *Hydrologic Systems: Watershed Modeling*. Vol. II, New Jersey, Prentice-Hall, 1989.

TAYLOR, J.K. - *Statistical Techniques for Data Analysis*. Michigan, Lewis, 1990.