UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Dissertação

Proposta de métodos de clusterização de dados com validação por testes de heterogeneidade e discordância aplicados à regionalização de bacias hidrográficas

Leroi Floriano de Oliveira

Pelotas, 2018

Leroi Floriano de Oliveira

Proposta de métodos de clusterização de dados com validação por testes de heterogeneidade e discordância aplicados à regionalização de bacias hidrográficas

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciência da Computação

Orientador: Prof. Dr. Marilton Sanchotene de Aguiar
Coorientador: Prof. Dr. Samuel Beskow

Pelotas, 2018

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

O48p Oliveira, Leroi Floriano

Proposta de métodos de clusterização de dados com validação por testes de heterogeneidade e discordância aplicados à regionalização de bacias hidrográficas / Leroi Floriano Oliveira ; Marilton Sanchotene de Aguiar, orientador ; Samuel Beskow, coorientador. — Pelotas, 2018.

63 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2018.

1. Aprendizado não-supervisionado. 2. Regionalização hidrológica. 3. Clusterização. I. Aguiar, Marilton Sanchotene de, orient. II. Beskow, Samuel, coorient. III. Título.

CDD : 005

Elaborada por Maria Inez Figueiredo Figas Machado CRB: 10/1612

Dedico este trabalho aos meus lindos filhos, Nick e Liz, que são o motivo de tudo que faço hoje. Ao meu pai, que, quando eu era pequeno, me levou na faculdade onde ele estudava para ajudá-lo a empalhar uma pomba; e à minha mãe, de quem me lembro encontrar diversas vezes dormindo com minha irmã pequena no colo, sentada de frente para a máquina de escrever, tentando terminar trabalhos da faculdade, ambos exemplos para mim.

AGRADECIMENTOS

Primeiramente gostaria de agradecer a meus pais e irmãos pelo grande exemplo para mim, eles foram fonte de inspiração para buscar sempre cada vez mais conhecimento. Agradeço meu orientador, Marilton, que teve uma enorme paciência com todas as minhas falhas e erros, e sempre esteve prontamente presente para me ajudar em todas as minhas dificuldades para realizar este trabalho. Agradeço também meu co-orientador Samuel e seu bolsista Felício que sempre se mostraram dispostos a me explicar e me fazer entender tudo o que precisei saber sobre o problema. Agradeço também à Fran, que me acompanhou nesta jornada, sempre repetindo que faltava só um pouco e que eu conseguiria. Por fim, agradeço meus filhos, que me acompanharam nessa jornada, sempre me perguntando se eu já era mestre.

É a verdade o que assombra,
o descaso o que condena,
a estupidez o que destrói...
— RENATO RUSSO

RESUMO

OLIVEIRA, Leroi Floriano de. **Proposta de métodos de clusterização de dados com validação por testes de heterogeneidade e discordância aplicados à regionalização de bacias hidrográficas**. 2018. 63 f. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2018.

Através da regionalização de bacias hidrográficas, é possível, dentre outras aplicações, fazer previsões estatísticas de vazões máximas e mínimas em cursos d'água. Diversos estudos demonstram bons resultados na utilização de clusterização para a formação de melhores regiões do ponto de vista hidrológico. Este trabalho aplica técnicas de aprendizado de máquina para a formação de regiões hidrologicamente homogêneas. Mais especificamente, neste trabalho foi explorada a utilização dos métodos: k-means, *affinnity propagation*, *aglomerative clustering* e *regions of influence* para a formação de regiões, fazendo-se, portanto, uma comparação entre os métodos e a utilização de técnicas de seleção de atributos. Ainda, neste trabalho também são propostos três métodos para a solução do problema, utilizando ajuste dos *clusters* com base nas medidas de heterogeneidade e discordância de Hosking. Dois destes métodos utilizam o algoritmo k-means fazendo variações nos *clusters* iniciais de forma a buscar centroides que melhor representem regiões hidrologicamente homogêneas. O outro método combina resultados de clusterização com o método *regions of influence*. Com os métodos propostos, foi possível alcançar uma melhora, de 63,2% para 90,5% de aproveitamento das regiões formadas para a aplicação da análise de frequência regional. Com este trabalho, concluiu-se que os atributos selecionados apresentaram melhores resultados que a utilização de todos os atributos; e, que os métodos propostos demonstram grande potencial, visto que apresentaram melhores resultados que outros métodos já existentes.

Palavras-Chave: aprendizado não-supervisionado; clusterização; regionalização hidrológica

ABSTRACT

OLIVEIRA, Leroi Floriano de. **Proposal of methods for data clustering with validation by tests of heterogeneity and disagreement applied to the regionalization of watersheds**. 2018. 63 f. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2018.

Through the regionalization of river basins, it is possible, among other applications, to make statistical forecasts of maximum and minimum flows in watercourses. Several studies show good results in the use of clustering for the formation of better regions from the hydrological point of view. This work applies machine learning techniques to obtain hydrologically homogeneous regions. More specifically, in this work we have explored the use of the methods: k-means, affinnity propagation, aglomerative clustering and region of influence for the formation of regions, which makes a comparison between the methods and the use of attributes selection techniques. In this work three methods are proposed for the solution of the problem, using clusters adjustment based on the measures of heterogeneity and discordance of Hosking. Two of these methods use the k-means algorithm making variations in the initial clusters in order to look for centroids that best represent hydrologically homogeneous regions. And the other method combines clustering results with regions of influence. With the proposed methods it was possible to improve, in the best result, from 63.2 % to 90.5 % of the utilization of the regions formed for the application of RFA. With this work it was concluded that the selected attributes presented better results than the use of all the attributes, and that the proposed methods show great potential since they presented better results than other already existing methods.

Keywords: unsupervised learning; clustering; regional flood frequency

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	*Aglomerative Clustering*
AM	*Aprendizado de Máquina*
ANA	Agência Nacional de Águas
AP	*Affinity Propagation*
IA	Inteligência Artificial
LSHPD	Laboratório de Simulação Hidrológica e Processamento de Dados
RFA	*Regional Frequency Analysis*
ROI	*Regions Of Influence*

SUMÁRIO

1 INTRODUÇÃO

A regionalização de bacias hidrográficas é um método utilizado na área da hidrologia aplicada, com o intuito de agrupar bacias hidrográficas com comportamentos e padrões hidrológicos semelhantes. O agrupamento de bacias hidrográficas permite, dentre outras aplicações, a previsão de comportamentos de cheias, estiagens e vazões em cursos d'água. Esta técnica de observação e previsão é chamada de *Regional Frequency Analysis* e fornece uma estimativa da frequência que determinado evento hidrológico ocorre em uma região, utilizando cálculos estatísticos (HOSKING; WALLIS, 1997).

Cada região é formada por um conjunto de bacias hidrográficas que consistem em áreas de drenagens de cursos de água e seus rios afluentes. Para que haja uma regionalização eficiente, é necessário que uma região atenda a critérios de homogeneidade. Um dos principais critérios é o teste de Heterogeneidade proposto por HOSKING; WALLIS (1997), onde, a partir de séries de dados relacionados a um comportamento, como dados de vazões máximas em cursos d'água ou dados de estiagem, é possível identificar se a região formada pode ser usada para a previsão de comportamentos em outros pontos não monitorados.

Cada área de drenagem possui diferentes características hidrológicas, tais como: área da bacia, comprimento do curso d'água, altura em relação ao nível do mar, declividade, etc. A utilização de métodos de clusterização, tendo como atributos estas características, permite formar grupos de áreas de drenagens com comportamento hidrológico semelhante, como observado nos trabalhos de RAO; SRINIVAS (2008); CORREA (2014); NOTO; LOGGIA (2009); KINGSTON et al. (2011); BESKOW et al. (2016).

A **clusterização** é uma técnica de formação de agrupamentos classificada como aprendizagem não supervisionada; dá-se este nome por não haver rótulos que definam cada um dos grupos a serem formados. Esta categorização de grupos de dados ou informações faz parte da organização de diversas áreas, como, por exemplo, a organização das diferentes espécies de seres vivos, a categorização das estrelas do cosmos, os tipos de música com base no interesse de usuários, os comportamentos

de consumo em lojas virtuais, etc. Todos estes problemas têm em comum a busca por agrupamentos, não sabendo de antemão quantos e/ou quais grupos se busca.

O emprego de métodos de clusterização à regionalização de bacias hidrográficas demonstra grande potencial para a busca de melhores regiões do ponto de vista de homogeneidade hidrológica se comparados com o método de conveniência geográfica que é tradicionalmente utilizado, uma vez que a formação destas regiões leva em conta as características e os comportamentos hidrológicos e não apenas características geográficas.

Neste contexto, esta dissertação tem como objetivo geral avaliar diferentes métodos de clusterização para séries de dados hidrológicos e propor métodos que inclua a validação dos *clusters* por testes de heterogeneidade e de discordância propostos por (HOSKING; WALLIS, 1997), tendo como estudo de caso a regionalização das bacias hidrográficas do estado do Rio Grande do Sul.

Para alcançar o objetivo geral, foram elaborados objetivos específicos nesta dissertação, os quais são: i) estudo do referencial teórico referente à classificação não supervisionada e clusterização; ii) Estudo dos fundamentos referente a regionalização de bacias hidrográficas e como determinar a homogeneidade das mesmas; iii) Estudar o estado da arte quanto à aplicação de técnicas de clusterização para a regionalização de bacias hidrográficas, bem como outras técnicas de IA; iv) Aplicar diferentes métodos de clusterização ao estudo de caso; v) Elaborar métodos de clusterização visando melhores resultados para a formação de regiões de bacias hidrográficas em comparação a outros métodos tradicionais; e, vi) Comparar os resultados dos diferentes métodos trabalhados.

Esta dissertação está organizada em 6 Capítulos. No Capítulo 1, é exposto o tema do trabalho bem como suas motivações e objetivos. No Capítulo 2, é abordado o referencial teórico que embasou o trabalho explorando os diferentes tipos de clusterização. No Capítulo 3, é apresentado o estudo do estado da arte em que são observados as diferentes técnicas de clusterizações aplicadas a regionalização de bacias hidrográficas. A Metodologia, estudo de caso e seleção de atributos são apresentados no Capítulo 4. No Capítulo 5, são expostos os resultados e as observações dos mesmos, que culminam nas principais conclusões e nas possibilidades de trabalhos futuros, apresentados no Capítulo 6.

2 CLUSTERIZAÇÃO DE DADOS

A classificação de dados em si tem aplicação em quase todos os campos da ciência, visto que organizar os dados torna possível uma identificação de padrões e comportamentos, além de permitir encontrar soluções e métodos para a resolução de problemas. Como mencionado na introdução, a Clusterização - também chamada de classificação não supervisionada - é uma técnica de aprendizado de máquina que busca semelhanças em conjuntos de dados formando grupos sem uma definição prévia de quais ou quantos grupos serão gerados.

2.1 Definição de *Cluster*

Como observado por EVERITT et al. (2011), um *cluster* pode ser definido por sua coesão interna (semelhança de seus itens) e por sua isolação externa (diferença dos itens presentes nos demais *clusters*). A maneira mais simples de identificar essas semelhanças e diferenças matematicamente é calculando a distância entre os itens; quando expostos em um plano, a identificação destas semelhanças e diferenças, como disposto na Figura 1, é facilitada. Pode-se identificar, na referida Figura, que há três grupos formados com itens semelhantes e com bastante distinção dos demais grupos. Na maioria das vezes, entretanto, os dados são mais complexos que o exemplo dado, tendo mais do que apenas dois atributos para comparar, e tendo muito mais proximidades do que os grupos neste exemplo.

Na Figura 2, expõem-se duas clusterizações aplicadas ao mesmo conjunto de dados em uma conjuntura um pouco mais complexa. Neste cenário, os itens estão mais distribuídos, tornando não tão trivial a decisão de a qual grupo cada item deve fazer parte ou a quantidade de grupos que deve formar-se. Uma das características da clusterização é que um mesmo conjunto de dados pode ser classificado com diferentes quantidades de *clusters*, sendo um dos desafios descobrir a quantidade mais adequada para cada objetivo. Outro desafio é a escolha de quais atributos melhor representam as características que se busca destacar entre os diferentes grupos, e qual técnica aplicar para medir a similaridade entre essas diferentes características.

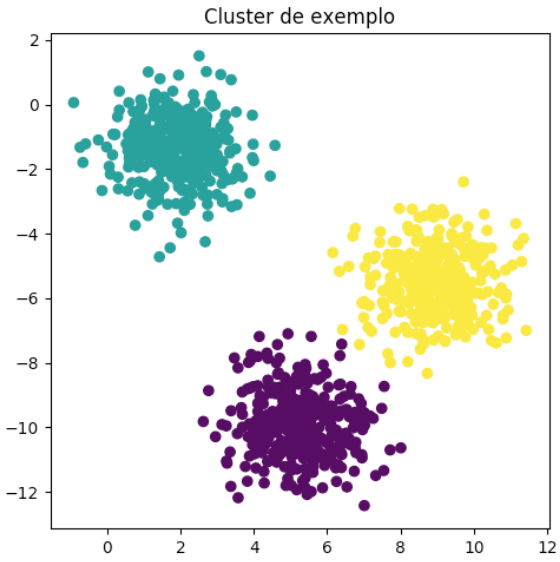

Figura 1 – Exemplo de agrupamento de dados em 3 *clusters*

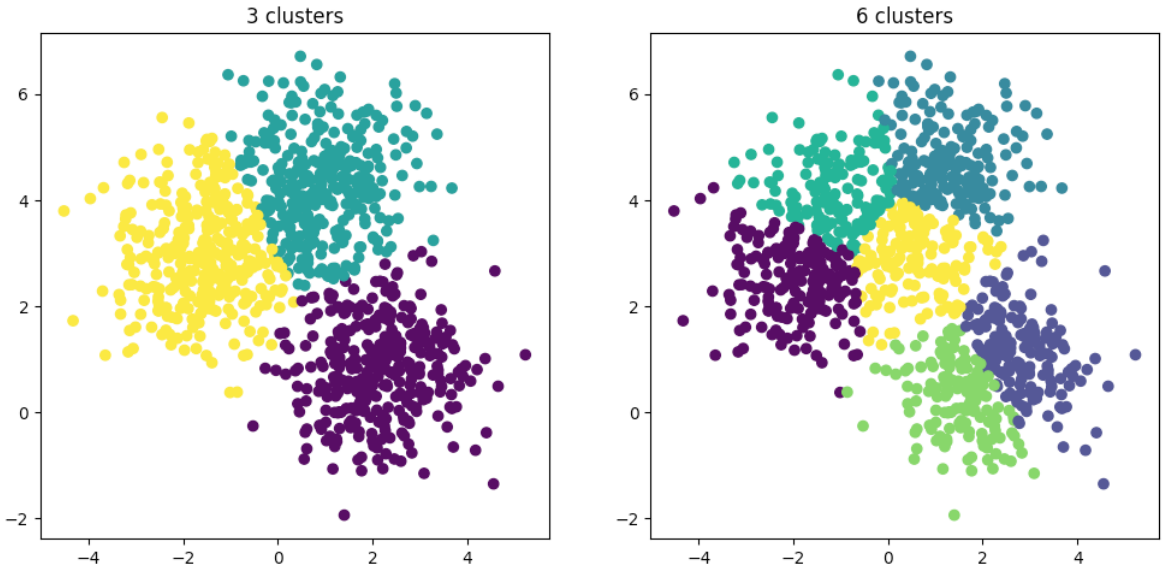

Figura 2 – Exemplo de agrupamento de dados em 3 e 6 *clusters*

2.2 Medidas de Similaridade

Sendo de central importância na identificação de *clusters*, a medida de similaridade consiste em avaliar o quão próximos (ou distantes) estão os itens uns dos outros. Por conseguinte, são considerados próximos dois itens que tenham dissimilaridade, ou distância pequena, ou, ainda, tenham grande similaridade (EVERITT et al., 2011). Quando comparados valores numéricos, as medidas mais comuns de proximidade são as métricas de Minkowski e suas variações, tais como: **Euclidiana**, **Manhattan** e **'sup'**. Já, quando comparados valores nominais, são criados vetores binários e definidos os coeficientes de combinação. As duas principais maneiras de calcular estes

coeficientes são os de **combinação simples** e de **Jaccard** (JAIN; DUBES, 1988). A forma de determinar a similaridade entre os *clusters* pode ser dividida em duas partes, sendo uma a similaridade entre os dados e a outra a matriz de similaridade que observa o grupo inteiro de dados.

2.2.1 Medidas de Distância

O princípio da comparação de similaridade ou distância entre os *clusters* tem base no cálculo da distância entre dados de dois itens dentre um grupo de itens que se busca classificar. Dado um conjunto de dados *x* e outro conjunto de dados *y*, ambos com a mesma quantidade de elementos, onde *n* é a quantidade de atributos, e os conjuntos têm exatamente os mesmos atributos que se pretende comparar.

A seguir está a representação matemática das principais distâncias para comparações de valores reais (JAIN; DUBES, 1988):

Minkowski:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Euclidiana:

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan:

$$\sum_{i=1}^{n} |x_i - y_i|$$

As distâncias Manhattan *(quando p=1)* e Euclidiana *(quando p=2)* são variações da distância Minkowski. Outras medidas de distância, que estão relacionadas a seguir, também são empregadas para medidas de similaridade na clusterização de dados (RAO; SRINIVAS, 2008; CHA, 2007):

Mahalanobis: (MAESSCHALCK; JOUAN-RIMBAUD; MASSART, 2000)

$$\sqrt{(x_i - y_i)^T S^{-1} (x_i - y_i)}$$

onde T é a transposição da matriz e S é a matriz de covariância para o vetor multivariado.

Caberra: (LANCE; WILLIAMS, 1966)

$$\sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Chebychev: (CHA, 2007)

$$\max_{1 \leq i \leq n} |x_i - y_i|$$

Como observado no trabalho de CHA (2007), as diferentes medidas de distância resultam em diferentes *clusters* formados, e podem apresentar melhores ou piores resultados dependendo de sua aplicação. Portanto, é necessário buscar a melhor combinação entre a medida de distância, método de clusterização e grupo de dados que se pretende clusterizar.

2.2.2 Medida de Proximidade Intergrupo

Após calculada a proximidade entre dois itens individuais dentre um grupo de dados, para se definir quais *clusters* formar, compara-se as distâncias entre todos elementos do conjunto observado para assim buscar quais se assemelham e quais são distintos. Segundo EVERITT et al. (2011), primeiramente a proximidade entre dois grupos pode ser definida por um resumo adequado das proximidades entre indivíduos de qualquer grupo. Em segundo lugar, cada grupo pode ser descrito por uma observação representativa ao escolher uma estatística de resumo adequada para cada variável, e, então, define-se a proximidade intergrupo como a proximidade entre as observações representativas.

Por exemplo, a base de uma das técnicas de clusterização, conhecida como *single linkage*, é tomar a menor dissimilaridade entre quaisquer dois itens individuais, um de cada grupo, que é referida como a *distância do vizinho mais próximo*. Também pode ser empregado o oposto a esta, chamada de *distância do vizinho mais distante*, no qual se buscam os itens individuais mais distantes de cada grupo. Outro método para a construção das medidas de dissimilaridades intergrupo é baseado no *means group*, também conhecida como centroide, no qual se faz uma média de todas as variáveis do grupo, definindo um valor para cada variável que represente aquele grupo; a distância, então, de um item individual para aquele grupo pode ser facilmente calculada a partir das métricas de distância anteriormente expostos (EVERITT et al., 2011).

Diferentes abordagens são utilizadas para fazer a classificação intergrupo, sendo isso o que tornam distintos os diferentes tipos de clusterização. Serão detalhados a seguir os principais métodos de clusterização conforme o tipo de classificação dos grupos juntamente com a abordagem utilizada para a definição dos *clusters*.

2.3 Clusterização Hierárquica

Na clusterização hierárquica, a formação dos *clusters* ocorre pela formação de uma hierarquia, onde cada passo das novas classificações dos dados é uma derivação decorrente do passo anterior, assim estruturando a hierarquia já mencionada. A clusterização hierárquica pode utilizar um método aglomerativo, o qual tem seu início com um *cluster* para cada elemento do grupo de dados que se pretende classificar, sendo, posteriormente, unidos até ter apenas um *cluster* com todos os elementos dispostos. Outro método é o divisivo, que, ao contrário do aglomerativo, inicia com todos os elementos em um único *cluster* e, então, são feitos particionamentos posteriores até a máxima divisão de um *cluster* por elemento. Em ambos os casos, não é necessário esgotar todos os passos e possibilidades, uma vez que o processo pode ser interrompido antes da exaustão com base em algum critério de parada, conforme o objetivo da clusterização, por exemplo: número de *clusters* (mínimos ou máximos) esperados; ou coesão entre os elementos do *clusters* (JOHNSON, 1967). Outros critérios serão explorados na sequência deste texto.

2.3.1 Algoritmos de Clusterização Hierárquica

Há diversos métodos de clusterização hierárquica; entretanto, o que os diferencia é a medida de similaridade utilizada para a divisão ou aglomeração dos *clusters* em cada etapa de clusterização. A seguir, estão listados os principais algoritmos de clusterização hierárquica (EVERITT et al., 2011):

Centroid Clustering : neste algoritmo são calculados centroides (média aritmética de cada atributo) para cada *cluster* e, então, são unidos os *clusters* com menor distância entre os centroides, dentre todos os *clusters* diferentes.

Single Linkage : também chamado de *nearest neighbor*, este algoritmo em cada etapa do processo de nova clusterização, ele une *clusters* com a menor distância entre os pares de dados dentre *clusters* diferentes.

Complete Linkage : também chamado de *furthest neighbor*, este algoritmo, ao contrário do *Single Linkage*, utiliza a distância entre os pares mais distantes para identificar a similaridade entre os *clusters*.

Group Average Linkage : neste algoritmo, para definir a distância entre os *clusters*, é calculada a média da distância entre todos os pares de dados formados pelos elementos dos *clusters* comparados. Então, a cada etapa, são unidos os *clusters* com menor distância.

Agglomerative Clustering : neste algoritmo, são combinados os pares de *clusters*

que minimamente aumenta uma determinada distância de ligação de forma recursiva.

2.3.2 Representação da Clusterização Hierárquica

Os resultados deste tipo de clusterização, tanto em métodos aglomerativos como em divisivos, podem ser representados por um diagrama bidimensional chamado de *dendrograma*, conforme pode ser observado na Figura 3 (EVERITT et al., 2011).

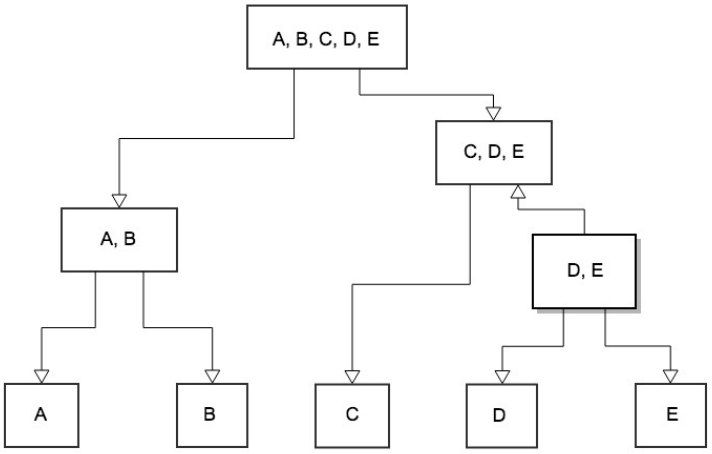

Figura 3 – Exemplo de dendrograma.

Uma vantagem da clusterização hierárquica é a de gerar uma hierarquia com os passos realizados pelo algoritmo, facilitando a leitura e compreensão tanto do resultado quanto do processo, sendo este último o que o algoritmo transitou para chegar a tal resultado.

Uma desvantagem na utilização de métodos hierárquicos é a utilização de uma abordagem gulosa, o que faz com que seja irreversível uma má escolha no algoritmo em alguma etapa; ou seja, uma vez que um nível da arvore hierárquica é montado, ele não é mais modificado pelo algoritmo refletindo nas escolhas futuras de divisões ou de aglomerações.

2.4 Clusterização Particional

Diferente da clusterização hierárquica, a clusterização particional divide os *clusters* em uma quantidade definida dos mesmos, realocando os itens para chegar ao melhor resultado de clusterização. Em outras palavras, para um conjunto de dados de x itens, o método de clusterização particional cria K partições, no qual cada partição corresponde a um *cluster*. Nesse método, cada *cluster* deve ter ao menos um item, e cada item pertencerá a apenas um cluster. A seguir estão dispostos alguns métodos de clusterização particional.

2.4.1 Algoritmo k-means

Descrito por MACQUEEN (1967), este algoritmo inicialmente define *K clusters* contendo um item aleatório de cada um dos *X* itens de um grupo de dados, onde cada item tem *N* atributos. Depois desta definição inicial, o algoritmo executa então dois processos.

No primeiro processo, após definidos os *k clusters* iniciais, são calculados os centroides de cada *cluster* com base nos atributos e, então, a cada iteração, um item é alocado ao *cluster* com centroide mais próximo utilizando uma medida de discordância. A cada interação, são calculadas novamente os centroides dos *clusters* formados com os itens presentes nos então novos *clusters* formados. Este processo se segue até que todos os *x* itens sejam atribuídos para um dos *k clusters*.

Em decorrência do primeiro, no segundo processo são calculados e armazenados os centroides dos *clusters* formados no primeiro passo, são removidos todos os itens dos *clusters* e são formados novos *clusters* alocando novamente cada um dos *x* itens para o *cluster* com menor medida de discordância. Este processo se repete até que não haja mudanças entre os *clusters* do início deste processo e o *cluster* formado ao final deste processo.

No algoritmo k-means, a comparação dos itens com cada *cluster* é feita com o *means* (ou a média) dos atributos de todos os itens presentes no *cluster*, característica esta que dá o nome ao algoritmo.

2.4.2 Algoritmo k-means++

O k-means++, proposto por ARTHUR; VASSILVITSKII (2007), é uma variação do algoritmo k-means mudando apenas o início do algoritmo que, ao invés de definir aleatoriamente os *clusters* iniciais, propõe uma heurística probabilística para definir centroides iniciais que descrevam melhor as diferentes características no conjunto de dados que se pretende classificar.

Primeiro, é definido um centroide inicial como sendo um dos itens, presente no conjunto de dados, escolhido de forma aleatória. Para a definição do item que irá representar o próximo centroide, se calcula uma medida de discordância entre todos os elementos, para aumentar probabilidade de escolha dos itens mais distantes para definir o próximo centroide. Incorporando então este novo centroide a um grupo de centroides e ajustando então a probabilidade de forma a buscar novamente itens distantes dos já escolhidos. Este processo é repetido até que todos os *k* centroides sejam definidos.

Após este processo de inicialização, o algoritmo k-means é executado de forma original. O algoritmo k-means++ apresenta melhores resultados, em comparação ao k-means, tanto em relação ao desempenho e quanto a qualidade dos *clusters* formados, principalmente quando classificados grandes conjuntos de dados (ARTHUR;

VASSILVITSKII, 2007).

2.4.3 Algoritmo k-medoids

O k-medoids é outro método de agrupamento onde, assim como o k-means, um conjunto de dados de *n* itens é agrupado em *K clusters* previamente definidos. Proposto por KAUFMAN; ROUSSEEUW (1987), este método funciona em princípio para minimizar os ruídos gerados por itens no conjunto de dados que tenham dados muito discrepantes do grupo. O algoritmo k-means é sensível a estes casos: uma vez que um item apresente valor extremamente grande pode distorcer substancialmente a distribuição de dados, o que é exacerbado devido ao uso do critério de erros quadrados. O método de k-medoids é, portanto, a forma modificada do k-means, visando diminuir esta sensibilidade. Dessa forma, ao contrário do método k-means, em vez de calcular os valores médios dos objetos em um *cluster* como ponto de referência, um item real dos dados é selecionado para representar o centroide do *cluster*. Este ponto é chamado de *medoide* ou *objeto representativo*. O medoide é o item mais central localizado dentro do *cluster*.

O método de particionamento é então realizado com base no princípio de minimizar a soma das dissimilaridades entre cada item e seu correspondente *objeto representativo* chamado *critério de erro absoluto*. Em termos mais simples, ao se estabelecer o critério de erro absoluto, a distância média do objeto representativo para todos os outros objetos do mesmo *cluster* são minimizados. O critério de erro absoluto usado é definido pela Equação 1, onde E é a soma do erro absoluto para todos os objetos no conjunto de dados, P é o ponto que representa um determinado objeto e o_i é o objeto representativo do cluster c_i (MILLER; HAN, 2001).

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_i| \tag{1}$$

2.5 Clusterização por propagação de afinidade

O algoritimo *Affinity Propagation* (AP) é um algoritmo de agrupamento baseado no conceito de “passagem de mensagem” entre pontos de dados, proposto por FREY; DUECK (2007). Diferente dos algoritmos de agrupamento como k-means ou k-medoids, o AP não exige que o número de *clusters* seja determinado ou estimado antes de executar o algoritmo. Semelhante ao k-medoids, o método de propagação de afinidade encontra “exemplares”, membros do conjunto de entrada, que são representativos de *clusters*.

Para a passagem de dados, são considerados dois aspectos: responsabilidade e disponibilidade. Considerando $x_1 \ldots x_n$ como um conjunto de itens de dados e sendo s uma função que quantifica a semelhança entre os dois pontos, então $s(x_i, x_j) >$

$s(x_i, x_k)$, se x_i é mais parecido com x_j do que com x_k.

O algoritmo prossegue alternando duas etapas de passagem de mensagens. Para atualizar duas matrizes: a) a matriz de responsabilidade R possui os valores $r(i,k)$ que quantificam o quão bem x_k serve como exemplo para x_i, em relação a outros exemplos de candidatos para x_i; e b) A matriz de disponibilidade A contém valores $a(i,k)$, que representam como apropriado para x_i escolher x_k como exemplar, levando em consideração a preferência de outros pontos para x_k como exemplar. Ambas as matrizes são inicializadas com zeros e podem ser vistas como tabelas de probabilidade logarítmica.

O algoritmo executa as seguintes atualizações iterativamente. Primeiro, as atualizações de responsabilidade são enviadas em torno de $r(i,k)$ representados na Eq. 2, após a disponibilidade é atualizada pelas Eqs. 3 e 4 para $i \neq k$.

$$r(i,k) = s(i,k) - \max_{k' \neq k} a(i,k') + s(i,k') \tag{2}$$

$$a(i,k) = \min \left(0, r(k,k) + \sum_{i' \notin i,k} \max\left(0, r(i',k)\right) \right) \tag{3}$$

$$a(k,k) = \sum_{i' \neq k} \max(0, r(i',k)) \tag{4}$$

As iterações são realizadas até que os limites do *cluster* permaneçam inalterados ao longo de várias iterações ou após algum número predeterminado de iterações. Os exemplares são extraídos das matrizes finais como aqueles cuja "responsabilidade + disponibilidade" para si seja positiva, ou seja, $(r(i,i) + a(i,i)) > 0$.

2.6 Clusterização híbrida

O método de clusterização híbrida se origina da combinação de dois outros métodos de clusterização, os hierárquicos e particionais, que já foram mencionados anteriormente neste capítulo. Esta combinação permite que um método complemente o outro em suas deficiências conforme suas características (RAO; SRINIVAS, 2008). Na clusterização hierárquica não há influência na inicialização, já nos algoritmos de clusterização particionais a definição inicial do número de *clusters* e dos *clusters* escolhidos na primeira inicialização influenciam diretamente no resultado.

Por outro lado, a clusterização particional apresenta vantagens em relação a hierárquica por possuir dinamicidade de movimentação dos objetos entre os clusters ao longo da execução do algoritmo - sendo o que a clusterização hierárquica não possui, pois não movimenta objetos entre os *clusters* perpetuando erros de decisão nas divisões ou agrupamentos nas novas hierarquias formadas.

Na clusterização híbrida proposta por RAO; SRINIVAS (2008), primeiro é execu-

tado um algoritmo de clusterização hierárquica no conjunto de dados para se obter uma definição inicial dos *clusters*. Partindo disso, a definição será utilizada como inicialização de um algoritmo de clusterização particional.

2.7 Clusterização por densidade

Neste método de clusterização, se observa a densidade de objetos em relação a medida de discordância e formando *clusters* com maior densidade de objetos, separando um *cluster* do outro pelas regiões de baixa densidade (ESTER et al., 1996).

Os métodos baseados em densidade apresentam como característica a utilização do critério de clusterização local ao considerarem a densidade de ligações entre os dados e também a não necessidade da definição de um número inicial de *clusters* que possibilita ao algoritmo encontrar os *clusters* de forma arbitrária. O principal algoritmo utilizado, o método de clusterização por densidade, é o DBSCAN, descrito a seguir.

2.7.1 DBSCAN

O algoritmo DBSCAN (do inglês, *Density Based Spatial Clustering of Applications with Noise*) proposto por ESTER et al. (1996), identifica as áreas de densidade através de uma abordagem baseada em centro, onde se buscam objetos ou pontos que tenham um número mínimo de vizinhos dentro de um determinado raio.

Sendo Eps o raio máximo da vizinhança formado por um número mínimo de pontos, denotados por $MinPts$, e a medida de distância entre os objetos sendo x_i e x_j, identificam-se as regiões densas; assim, os pontos são classificados conforme regras descritas a seguir:

- x_i é um ponto central, se $|N_{Eps}(x_i)| \geq MinPts$
- x_i é um ponto periférico, se pertence a vizinhança de um ponto central x_j, ou seja, $x_i \in N_{Eps}(x_i)$
- x_i é considerado ruído, se não atender as duas regras anteriores

A formação do *cluster* é feita unindo os pontos densamente conectados. Se a distância entre dois pontos for menor que Eps, estes farão parte do mesmo *cluster*: os pontos periféricos são colocados no mesmo *cluster* que os pontos centrais correspondentes; já, os pontos ruidosos, por não pertencerem a nenhum *cluster*, não são classificados em nenhum *cluster*. Os parâmetros de entrada neste algoritmo são o tamanho da vizinhança Eps, o número mínimo de pontos $MinPts$ e o conjunto de dados.

2.8 Clusterização por Algoritmos Genéticos

Os algoritmos genéticos, propostos por HOLLAND (1992) inspirados na teoria da evolução das espécies de Darwin, são algoritmos voltados para buscar soluções em problemas de otimização e busca. Seguem o princípio de gerar populações de indivíduos, onde cada indivíduo é uma possível solução para o problema, e os "melhores" indivíduos, ou seja, aqueles que melhor se adaptarem ao seu meio, terão mais chances de sobreviver e gerar descentes com suas características hereditárias.

O cruzamento dos sobreviventes geram uma nova população de soluções, podendo ainda ser incluído o princípio de mutação, que altera de forma sucinta alguns indivíduos desta geração de forma aleatória. O algoritmo segue formando novas gerações a cada iteração até obedecer um critério de parada, que normalmente é definido arbitrariamente em um número de repetições, com base no tempo que se pretende executar, no resultado mínimo esperado ou na estagnação da busca (quando há pouca variação na solução entre as gerações).

Uma das aplicações mais comuns dos algoritmos genéticos para a clusterização é para melhorar o desempenho e qualidade dos resultados de outros algoritmos de clusterização como o exemplo do *genetic k-means algorithm* (GKA).

2.8.1 *Genetic k-means algorithm*

O *genetic k-means algorithm*, proposto por KRISHNA; MURTY (1999), utiliza um AG em conjunto com o algoritmo k-means para fazer a clusterização do conjunto de dados. Apesar de apropriar-se de praticamente todas as etapas de um AG tradicional, a etapa de cruzamento foi substituída pela aplicação do k-means em cada cromossomo da população, sendo utilizado como operador de busca, reduzindo a complexidade do algoritmo. Tendo em vista que algoritmos genéticos utilizados para clusterização apresentam alto custo em suas operações de cruzamento de cromossomos e na função objetiva, esta combinação visa otimizar o processo de clusterização.

O GKA recebe, como parâmetros de entrada, o número de *clusters k* a serem formados, o tamanho da população, a probabilidade de mutação e o número de gerações, ou seja, o número de iterações do algoritmo. Cada cromossomo pode ser visto como um vetor de tamanho *n*, onde cada posição representa um objeto do conjunto de dados e o valor dessa posição indica o cluster a qual o objeto está associado.

Inicialmente, os cromossomos (soluções) da população são inicializados aleatoriamente. O algoritmo não permite que existam soluções com *clusters* vazios; sendo assim, é empregado um esforço adicional para satisfazer esta propriedade. Na etapa de mutação, é aplicada a cada cromossomo da população o valor de cada alelo (objeto) do cromossomo, o qual pode ser substituído dependendo da distância entre esse objeto e os centroides dos *clusters*. O centroide mais próximo ao dado possui maior

probabilidade de ser escolhido como novo valor para o alelo. A probabilidade de mutação define a chance de um alelo ser substituído. É indicado que esse fator seja baixo, na faixa entre 0.01 e 0.05, pois fatores maiores tendem a alterar o comportamento do algoritmo, deixando-o com características exageradamente aleatórias.

Na etapa final, para cada cromossomo é aplicado o algoritmo k-means na tentativa de melhorar as soluções, visto que nas etapas anteriores o refinamento dá-se apenas com base em probabilidades. O algoritmo termina sua execução quando atingir o número de gerações estabelecidas.

2.9 Validação dos *clusters*

Os resultados da clusterização podem variar de acordo com algumas escolhas feitas ao longo do processo, como: método e algoritmo utilizado, medida de similaridade, atributos escolhidos para a classificação, quantidade de *clusters* definida, entre outros. Para que seja encontrado o melhor resultado, normalmente são executados diferentes métodos de clusterização variando as combinações das escolhas anteriormente mencionadas. Após a exploração das diferentes combinações de dados de entrada e métodos de clusterização, é necessário aplicar métricas de validação dos *clusters* para identificar quais obtiveram melhores resultados. Além disso, o mesmo cenário pode apresentar parte dos *clusters* com a qualidade esperada e outra parte não.

Para determinar a relevância dos resultados obtidos, geralmente utilizam-se índices estatísticos que medem quantitativamente a qualidade de um *cluster*, chamado de critério de validação. Os principais critérios de validação, apresentados por JAIN; DUBES (1988), estão descritos a seguir:

Critérios externos : a qualidade do agrupamento é medida de acordo com a divisão previamente estabelecida do conjunto de dados. Normalmente um especialista que detenha conhecimento de como classificar os dados faz esta divisão. Após são realizados testes de hipótese, com o intuito de comprovar a hipótese pré-definida. Uma das maneiras de quantificar esta qualificação é através da análise de Monte Carlo (SMYTH, 1996).

Critérios internos : neste critério, os dados internos presentes no conjunto de dados são trabalhados para medir a qualidade do *cluster*. Ou seja, são calculados índices com base nas medidas de similaridade dos *clusters* formados para validar o quanto um *cluster* está distinto do outro e similar entre si. Ao se observarem os índices internos, é possível identificar características muito importantes para a qualidade da clusterização, como, por exemplo, definir o número real de *clusters* presentes no conjunto de dados ou então avaliar e comparar os tamanhos dos *clusters* formados.

2.9.1 Melhor quantidade de *clusters*

Um dos principais problemas a ser resolvido ao aplicar algoritmos de clusterizações em conjuntos de dados é definir a melhor quantidade de *clusters* para se obter o melhor agrupamento. Esta decisão será peculiar a cada problema, variando conforme o conjunto de dados e o resultado esperado com o agrupamento.

Um método bastante utilizado para determinar o melhor número de *clusters* é o *Elbow Method* ou Método do Cotovelo, e é um dos mais antigos para determinar o número ideal de *clusters* em um conjunto de dados. Sendo este um método visual, se inicia $k = 2$ (número de *clusters*) e a cada etapa se incrementa em 1 o valor de k, calculando seus *clusters* e o custo de processamento. Com algum valor para k, atinge um platô do custo. Ao atingir esta estabilidade de custo, indica que aquele k é o número de *clusters* ideal (KODINARIYA; MAKWANA, 2013).

2.9.2 Seleção de Atributos

No processo de clusterização, os atributos selecionados para o conjunto de dados irão determinar diretamente a formação dos *clusters*. Normalmente, quanto maior a quantidade de atributos, mais características temos disponíveis para identificar as peculiaridades de cada objeto do conjunto de dados, e assim permitir ao algoritmo identificar padrões que humanamente seria inviável. Alguns atributos, porém, podem dificultar a clusterização e, por essa razão, uma maneira eficiente de lidar com o problema é criar um subconjunto de dados com atributos mais importantes (DASH; LIU, 1997).

Para determinar os atributos mais significativos, DASH; LIU (1997) descreve um método contendo dois passos: *generation procedure* (ou processo de geração) e *evaluation function* (ou função de avaliação).

2.9.2.1 Processo de geração

Dado um conjunto de dados com um número N de atributos, o total de subconjuntos possíveis será 2^N. Esta quantidade pode ser grande mesmo para um conjunto de dados com quantidade moderada de atributos. A seguir, estão listadas algumas abordagens para solucionar esse problema.

Completa : o qual faz uma busca completa pelo melhor subconjunto de atributos com base na função de avaliação.

Heurística : no qual, a cada iteração, os atributos ainda não selecionados são selecionados através de uma heurística. Algumas heurísticas bastante utilizadas para neste processo são: *sequential forward selection* (SFS) e *sequential backward selection* (SBS), *sequential backward selection-SLASH* (SBSSLASH), *(p,q) se-*

quential search (PQSS), *bi-directional search* (BDS), *Schemata Search* and *relevance in context* (RC) (DASH; LIU, 1997).

2.9.2.2 Função de avaliação

A função de avaliação é o método utilizado para definir qual subconjunto de atributos é melhor para caracterizar um *cluster*. Quanto mais um atributo contribuir para distinguir um grupo de objetos de outro, dentro do objetivo esperado, mais significativo será este atributo para a classificação. Diversas funções de avaliações podem ser utilizadas para a seleção dos atributos mais significativos.

Diversos métodos podem ser utilizados como função de avaliação (DASH; LIU, 1997), como, por exemplo, o método *Relief* que usa de estatística para selecionar os atributos relevantes baseado em algoritmos *instance-based learning*. Este método, em primeiro momento, seleciona aleatoriamente amostras de instâncias dentro de um conjunto de dados de treinamento. Para cada instância, são buscados o *Near Hit* e o *Near Miss*, com base na medida de distância euclidiana. *Near Hit* é a instância de distância euclidiana mínima entre todas as instâncias da mesma classe da instância escolhida; já o *Near Miss* é a instância com distância euclidiana mínima entre todas as instâncias de diferentes classes.

São então atualizados os pesos dos recursos inicializados com zero, considerando que uma característica é mais relevante se distingue uma instância de sua *Near Miss*; e menos relevante se distingue uma instância de seu *Near Hit*. Depois de esgotar todas as instâncias na amostra, são elegidas todas as características com peso maior ou igual a um limite, o qual pode ser automaticamente avaliado usando uma função que usa o número de instâncias na amostra; ou, também pode ser determinado por inspeção (todos os recursos com pesos positivos são selecionados) (DASH; LIU, 1997).

3 REGIONALIZAÇÃO DE BACIAS HIDROGRÁFICAS

Uma bacia hidrográfica é uma área geográfica delimitada onde ocorre escoamento de um rio central e seus afluentes. Este escoamento acontece pela drenagem da água da chuva de forma superficial gerando rios ou riachos, ou de forma que a água se infiltra no solo formando nascentes ou indo diretamente para o lençol freático (SCHIAVETTI; CAMARGO, 2002). Cada bacia hidrográfica possui diversas sub-bacias ou áreas de drenagem de cada afluente. Na Figura 4, pode-se ver o mapa contendo as áreas de drenagem, representadas pelos polígonos demarcados em vermelho, encontradas no estado do Rio Grande do Sul, Brasil.

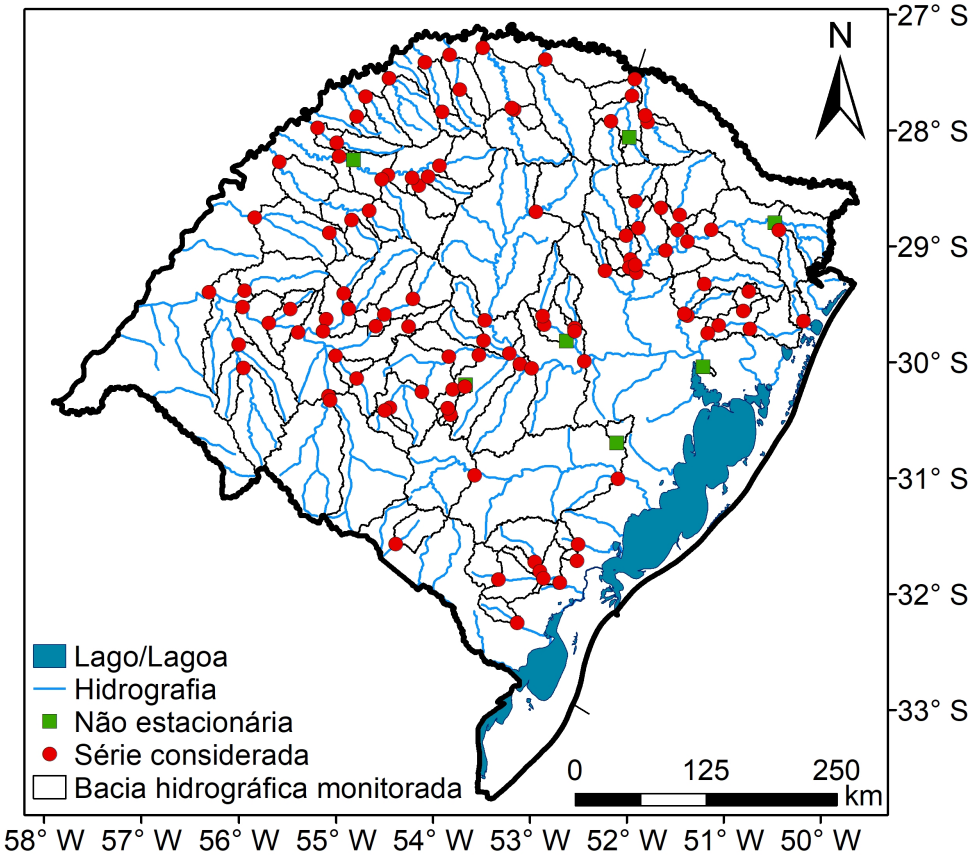

Figura 4 – Mapa das áreas de drenagem do estado RS

A regionalização de bacias hidrográficas é um método utilizado na área da hidrologia aplicada com o intuito de agrupar bacias hidrográficas com comportamentos e padrões hidrológicos semelhantes. O agrupamento de bacias hidrográficas permite, dentre outras aplicações, a previsão de comportamentos de cheias, estiagens e vazões em cursos d'água. Esta técnica de observação e previsão é chamada de *Regional Frequency Analysis* (RFA) e fornece uma estimativa da frequência que determinado evento hidrológico ocorre em uma região, utilizando-se de cálculos estatísticos (HOSKING; WALLIS, 1997).

Uma das possíveis aplicações ao se utilizar a RFA é fazer a previsão estatística de vazões máximas em cursos de água, previsão esta que será o enfoque deste trabalho. O cálculo estatístico leva em consideração as séries históricas de dados de vazão de água obtidos em um ponto de controle no início do curso d'água principal da área de drenagem. Na Figura 4, podem ser observados os pontos de controle, identificados pela cor verde, em suas respectivas áreas de drenagem.

Através dos métodos de *index-flood* – proposto por DALRYMPLE (1960) – e de momentos-L – proposto por HOSKING; WALLIS (1997) –, é possível identificar quais regiões, formadas por *n* áreas de drenagem, são hidrologicamente homogêneas, permitindo a previsão estatística com transposição de dados para pontos em cursos d'água que não tenham dados históricos de vazão.

HOSKING; WALLIS (1997) estabelecem a realização de quatro etapas no método momentos-L que são: triagem inicial dos dados, identificação de regiões hidrologicamente homogêneas, escolha da função densidade de probabilidades e estimativa da função densidade de probabilidades. Dentre estas etapas, a tarefa mais importante para o método é a identificação das regiões hidrologicamente homogêneas, e, portanto, merece mais atenção (HUSSAIN; PASHA, 2009).

3.1 Regiões hidrologicamente homogêneas

Para determinar se uma região é hidrologicamente homogênea, é utilizada a medida de heterogeneidade *H*, proposta por HOSKING; WALLIS (1997). Esta medida se baseia na análise da variabilidade estatística das séries de dados históricos da região formada em comparação com o que seria uma região ótima do ponto de vista de homogeneidade hidrológica. A medida de heterogeneidade *H* é descrita resumidamente pela Equação 5.

$$H = \frac{(V - \mu_V)}{\sigma_V} \tag{5}$$

onde V é o desvio padrão ponderado do coeficiente de variação L; μ_V e σ_V são o desvio padrão e médio das simulações. Uma região pode ser considerada considerada hidrologicamente homogênea se $H < 1$, possivelmente homogênea se $1 \leq H \leq 2$ e

heterogênea se $H > 2$.

Segundo HOSKING; WALLIS (1997), no entanto, o critério de conveniência geográfica – o qual é uma das abordagens utilizadas para formar as regiões –, normalmente não formam as melhores regiões, do ponto de vista de homogeneidade hidrológica. Com base nisso, diversos estudos (RAO; SRINIVAS, 2008) demonstram bons resultados na utilização de técnicas de clusterização para a regionalização de bacias hidrográficas.

3.1.1 Ajuste das Regiões

Após a formação das regiões e da verificação de sua homogeneidade é possível que alguma região formada pela clusterização não seja estatisticamente homogênea em relação às vazões. A razão pela qual más regiões são formadas é justificada por conta da utilização de atributos não exaustivos. Entretanto, as modificações não serão substanciais se selecionados atributos correlacionados com comportamentos de cheias em bacias hidrográficas e se selecionado um algoritmo eficiente de clusterização (RAO; SRINIVAS, 2008).

HOSKING; WALLIS (1997) sugerem as seguintes opções para ajuste de regiões: (i) eliminar (ou excluir) uma ou mais áreas do conjunto de dados; (ii) transferir (ou mover) uma ou mais áreas de uma região para outras regiões; (iii) dividir uma região para formar duas ou mais novas regiões; (iv) permitir que algumas áreas sejam compartilhadas com duas ou mais regiões; (v) dissolver uma região transferindo suas áreas para outras regiões; (vi) unir uma região com outra ou outras; e, (vii) obter mais dados e redefinir regiões.

Para mover ou remover áreas de uma região, utiliza-se uma medida de discordância. A opção principal considerada para revisar uma região é eliminar uma ou mais áreas que são extremamente discordantes em relação a outras áreas da região. A área eliminada de uma região é transferida para outra região (destinatária) que seja mais próxima da área eliminada no espaço de atributos multidimensional, desde que a transferência não afete negativamente a homogeneidade da região receptora (RAO; SRINIVAS, 2008).

3.1.2 Medida de discordância

A medida de discordância, proposta por HOSKING; WALLIS (1997), é útil para identificar áreas ou séries de dados históricos com erros grosseiros nos mesmos, ou aqueles que são grosseiramente discordantes com a região como um todo. Para estimar os valores de discordância para series em uma região, as séries são consideradas como pontos no espaço tridimensional de proporções de momentos-L da amostra (L-CV, L-Skewness, e L-Kurtosis). O centroide da região é considerado um ponto que representa a média de proporções de momentos-L de amostra das séries na região.

Qualquer ponto que esteja longe do centroide da região é marcado como discordante. A medida de discordância é descrita pela Equação 6.

$$D_i = \frac{1}{3} N_R (u_i - \bar{u})^T S^{-} 1 (u_i - \bar{u}) \tag{6}$$

onde i é uma série de dados históricos, N_R é o número de séries históricas presentes na região, u_i é calculado pela Equação 7, $\bar{u}$ é calculado pela Equação 8 e S é calculado pela Equação 9.

$$u_i = [t^i t_3^i t_4^i]^t \tag{7}$$

$$\bar{u} = \frac{\sum_{i=1}^{N_R} u_i}{N_R} \tag{8}$$

$$S = \sum_{i=1}^{N_R} (u_i - \bar{u})(u_i - \bar{u})^T \tag{9}$$

Segundo (HOSKING; WALLIS, 1997), para que uma série de dados históricos seja discordante em relação ao restante da região, dependerá do valor de N_R, ou seja, do número de séries históricas presentes na região. HOSKING; WALLIS (1997) definem que seja seguida a Tabela 1 e que sejam consideradas as séries discordantes quando o valor *D* da série for maior que o valor crítico.

Tabela 1 – Valores críticos para medida de discordância, segundo HOSKING; WALLIS (1997)

N_R	Valor Crítico para D_i
5	1.333
6	1.648
7	1.917
8	2.140
9	2.329
10	2.491
11	2.632
12	2.757
13	2.869
14	2.971
15	3.000

3.2 Seleção de atributos

Como exposto no Capítulo 2, a seleção de atributos é um passo importante para encontrar *clusters* mais significativos. No caso de clusterização de bacias hidrográfi-

cas, com o intuito da busca de regiões hidrologicamente homogêneas para a previsão de vazões máximas, é importante a busca de atributos que melhor representem as características de vazão em cursos d'água.

Como observado em estudos recentes (HADDAD; RAHMANB, 2012; RAO; SRINIVAS, 2008), os atributos relacionados às características fisiográficas, localização e climatologia da bacia hidrográfica têm sido bastante utilizados para clusterização de bacias hidrográficas.

A seguir, estão listados os principais atributos utilizados:

- **Área da bacia hidrográfica (A)**: valor que corresponde ao total da área, representado na unidade de quilômetros quadrados.
- **Coeficiente de compacidade (kc)**: um número adimensional formado pela relação do perímetro da bacia com o perímetro de um círculo de mesma área. Quanto mais irregular a bacia, maior será o valor do coeficiente de compacidade.
- **Densidade de drenagem (Dd)**: representada pela unidade de km/km^2, é calculada pela divisão do Comprimento total de todos os segmentos (L) pela área da bacia em quilômetros quadrados.
- **Comprimento do principal curso d'água (L)**: medição do comprimento do principal curso d'água presente na bacia, na unidade de quilômetros.
- **Declividade do principal curso d'água (Ls)**: medido na unidade de m/m, corresponde a declividade entre o início e o final do principal curso d'água.
- **Altitude média (E)**: medido em quilômetros, corresponde a altura média compreendida na área de drenagem.
- **Declividade média (S)**: representada na unidade de m/m, é definida calculando uma média entre as diferentes declividades presentes na área de drenagem.
- **Precipitação diária máxima anual (Pd)**: medida em milímetros, a précipitação é considerada máxima por ocorrência extrema, com duração, distribuição temporal e espacial crítica para uma área de drenagem.
- **Precipitação média anual (Pa)**: medida também em milímetros, corresponde à média de precipitação ao longo do período de um ano na área de drenagem.

As diferentes características de cada região do mundo podem apresentar atributos mais ou menos significativos. Para melhor escolha desses atributos, pode-se comparar sua correlação com variáveis relacionadas à inundações – por exemplo: inundação anual média e mediana anual; média e mediana anual por área; enchente anual média e mediana por precipitação anual média –, selecionando-se, assim, os atributos de maior correlação (RAO; SRINIVAS, 2008).

3.3 Regionalização por regiões de influência

A abordagem de Regionalização por regiões de influência (do inglês, *Region Of Influence* – ROI), proposta por BURN (1990), permite que cada site tenha sua própria região. O ROI de uma área de drenagem é formado pelas áreas cuja distância para a área de drenagem alvo não excede ao valor limiar escolhido. Para definir a distância, são utilizados os atributos no espaço multidimensional.

Na estimativa de uma curva de crescimento regional, cada área de drenagem pode ser ponderada de acordo com sua proximidade com a área de drenagem alvo. A seleção e a ponderação de atributos pode ser um problema, por haver o aumento do número de atributos disponíveis para a análise (RAO; SRINIVAS, 2008). BURN (1990) recomenda o uso de Mahalanobis, no lugar da medida euclidiana, como medida de distância para avaliar a semelhança entre as áreas de drenagem, porque demonstra de forma mais significativa a correlação entre os atributos de bacias hidrográficas e a regionalização.

A distância Mahalanobis leva em consideração a variância e covariância das variáveis, o que não é possível com a distância euclidiana. A abordagem ROI permite considerar a incerteza da estimativa devido à variabilidade da amostragem nas medidas que descrevem a sazonalidade das inundações de bacias hidrográficas.

3.4 Aplicação de métodos de clusterização à regionalização

Na observação de diferentes trabalhos (RAO; SRINIVAS, 2008; CORREA, 2014; NOTO; LOGGIA, 2009; KINGSTON et al., 2011; BESKOW et al., 2016), é notável a utilição de diversos métodos de clusterização para a regionalização de bacias hidrográficas. Nesta Seção, serão apresentados alguns desses trabalhos e um resumo de seus resultados.

Dentre os trabalhos observados, destacam-se os trabalhos de CORREA (2014) e de BESKOW et al. (2016), que implementam os algoritmos k-means, k-medoids, fuzzy c-means, kharmonic means e genetic k-means algorithm em conjunto com as métricas de distância de Minkowski, Euclidiana, Manhattan e Mahalanobis. Além disso, aplicam estes métodos a um conjunto de dados do estado do Rio Grande do Sul, Brasil. Os melhores resultados foram alcançados utilizando o algoritmo GKA, logo após resultados intermediários com o algoritmo FCM. Como critérios de validação, foram utilizados Índice de Soma dos Erros Quadráticos, Índice de Davies-Boudin, Índice de Calkinski-Harabasz e Índice PBM.

Outro trabalho relevante é o de RAO; SRINIVAS (2008), que aplica métodos de clusterização hierárquica (*simple linkage* e *complete linkage*), k-means, clusterização híbrida (utilizando clusterização hierárquica e k-means) em um conjunto de dados do estado de Indiana nos EUA. Como validação dos clusters, RAO; SRINIVAS (2008)

utilizam o *Cophenetic Correlation Coefficient*, índice de Davies-Boudin e *Silhouette Width*. Além deste, RAO; SRINIVAS (2008) também validam a qualidade das regiões formadas com base no teste de heterogeneidade de HOSKING; WALLIS (1997). Apresentam-se, no referido trabalho, melhores resultados com métodos híbridos, e, com resultados intermediários, o método k-means.

No trabalho de NOTO; LOGGIA (2009) foi utilizado o método hierárquico *Ward's Method*, o qual obteve melhores resultados em comparação a métodos de regionalização que não utilizam clusterização. Já no trabalho de KINGSTON et al. (2011) foram utilizados os métodos *Ward's Method*(método hierárquicos) e *k-means*(método não hierárquicos), obtendo melhores resultados na utilização do método *k-means*.

Observa-se também que a regionalização só é efetiva e utilizável quando atende aos níveis mínimos do teste de heterogeneidade, sendo esta a medida mais importante para a validação dos *clusters* (RAO; SRINIVAS, 2008; HOSKING; WALLIS, 1997).

4 METODOLOGIA E DESENVOLVIMENTO

Neste capitulo, será exposta a metodologia utilizada para o desenvolvimento do trabalho, o estudo de caso, as atividades e métodos utilizados para alcançar os objetivos propostos e os resultados obtidos. Pode-se dividir o processo em quatro partes: preparação e separação dos dados, aplicações de técnicas de clusterização, armazenamento dos resultados, processamento e tabulação dos resultados.

4.1 Preparação e separação dos dados

Como estudo de caso, foi escolhido o estado do Rio Grande do Sul, Brasil, do qual foram obtidos dados sobre as áreas de drenagem presentes no estado junto ao Laboratório de Simulação Hidrológica e Processamento de Dados (LSHPD), do curso de Engenharia Hídrica e do Programa de Pós-Graduação em Recursos Hídricos, da Universidade Federal de Pelotas. A região ao todo é divida em 103 áreas de drenagem, as quais são utilizadas no processo de regionalização.

Destas áreas de drenagem foram fornecidos os dados:

- Área da bacia hidrográfica (A)
- Coeficiente de compacidade adimensional (kc)
- Densidade de drenagem (Dd)
- Comprimento do principal curso d‘água (L)
- Declividade do principal curso d‘água (Ls)
- Altitude média (E)
- Declividade média (S)
- Precipitação diária máxima anual (Pd)
- Precipitação média anual (Pa)

CODIGO	A	Slope	L	ELEV_b	ELEV_r	Kc	DD	Panual	Pmaxdia	PAanual	PmdA
72400000	-0,25508	0,060702	-0,360578	1,467914	0,146782	-1,245017	0,081371	1,083896	-0,86234	-0,186733	-0,317889
72430000	-0,179876	0,140369	-0,234298	1,454284	0,023454	-1,180265	0,190967	1,136542	-0,761755	-0,10764	-0,235642
72580000	-0,498808	0,02356	-0,373723	1,257744	0,181006	-1,383786	-0,225479	-0,398329	-0,836522	-0,527775	-0,558168
72630000	0,227173	0,628161	0,382134	1,210506	-0,319759	-0,771014	0,294376	0,511428	-0,424751	0,26168	0,194035
72680000	0,428809	0,769325	0,629557	1,185297	-0,476331	-0,461	0,258237	0,654086	-0,426003	0,47384	0,394319
73480000	0,439072	0,270744	0,616571	1,190105	-0,451646	-0,15297	1,045399	0,765285	0,210637	0,491352	0,452064
74205000	-1,1597	0,619306	-1,208001	0,850397	0,796163	-1,489996	1,268279	0,687849	-0,081103	-1,122809	-1,158521
74210000	0,163952	0,048439	0,755353	0,846743	-1,107414	0,653923	-0,252967	1,056256	0,206276	0,233236	0,178334

Figura 5 – Amostragem dos valores dos atributos

Uma amostragem destes dados estão dispostos na Figura 5.

Também foram obtidos os dados das vazões máximas anuais no principal curso d'água de cada área de drenagem apresentada na Figura 4 (página 29). Valores esses que serão utilizados para fazer os testes de heterogeneidade e discordância de HOSKING; WALLIS (1997).

4.1.1 Seleção de Atributos

Para seleção dos atributos a serem utilizados, foi aplicada a metodologia proposta por RAO; SRINIVAS (2008) na qual se monta uma tabela de correlações dos atributos em relação às variáveis de inundação listadas a seguir:

- Inundação anual média (MAF)
- Mediana de inundação máxima anual (MEF)
- Inundação anual média por área (MAF/A)
- Mediana de inundação máxima anual por área (MEF/A)
- Inundação anual média por precipitação anual média (MAF/P)
- Mediana de inundação por precipitação anual média (MEF/P)

Esses dados foram também obtidos junto ao LSHPD e é demonstrada sua correlação com os atributos na Tabela 2.

RAO; SRINIVAS (2008) consideram que um atributo apresenta colinearidade elevada quando os coeficientes de correlação são superiores a 0,8. De acordo com a Tabela 2, a densidade de drenagem, elevação e declividade da bacia hidrográfica e a precipitação diária máxima anual têm a correlação mais fraca com as variáveis relacionadas às inundações, portanto não foram utilizadas na análise de agrupamento.

Tanto a área (A) quanto o comprimento do canal principal (L) têm forte correlação com as variáveis relacionadas à inundação. No entanto, também apresentam colinearidade elevada, uma vez que compartilham um coeficiente de correlação de 0,86. Isto posto, para garantir apenas o uso de variáveis independentes, foi selecionado o que

Tabela 2 – correlação entre os atributos das áreas de drenagem e as variáveis de inundação

	A	Kc	Dd	L	Ls	E	S	Pd	Pa
A	1								
Kc	0.57	1							
Dd	0.04	0.02	1						
L	0.86	0.68	0.07	1					
Ls	-0.29	-0.28	-0.30	-0.43	1				
E	-0.14	0.15	-0.08	0.06	0.15	1			
S	-0.18	0.17	-0.35	-0.11	0.60	0.50	1		
Pd	0.08	-0.20	0.10	0.02	-0.40	-0.62	-0.59	1	
Pa	-0.19	-0.15	-0.21	-0.06	-0.13	0.21	0.00	0.43	1
MAF	0.79	0.60	0.07	0.88	-0.32	0.13	-0.02	-0.10	-0.17
MEF	0.80	0.60	0.07	0.88	-0.32	0.12	-0.02	-0.10	-0.18
MAF/A	-0.38	-0.35	0.06	-0.41	0.27	0.10	0.15	-0.01	0.14
MEF/A	-0.38	-0.35	0.08	-0.42	0.29	0.09	0.18	-0.02	0.10
MAF/P	0.79	0.60	0.08	0.86	-0.31	0.12	-0.02	-0.12	-0.23
MEF/P	0.80	0.60	0.08	0.86	-0.31	0.11	-0.02	-0.12	-0.23

apresentou a maior correlação com as variáveis relacionadas à inundação (ou seja, o comprimento do canal principal).

O coeficiente de compacidade adimensional (kc) também foi selecionado, porque apresenta algum grau de correlação com as variáveis relacionadas à inundação sem mostrar colinearidade forte com qualquer outro atributo selecionado. Também foram selecionados os atributos Altitude média (E) e Declividade média (S) por apresentarem níveis razoáveis de correlação.

Para fins de experimentos, foram realizados os procedimentos de clusterização com diferentes atributos, comparando os resultados entre a seleção conforme proposta por RAO e a utilização de todos os atributos disponíveis. Mais detalhes sobre estas comparações serão apresentados no decorrer deste capítulo.

4.2 Metodologias para clusterização e formação de regiões

Nesta seção, serão abordadas as técnicas de clusterização utilizadas para a formação das regiões e tecnologias utilizadas.

4.2.1 Algoritmo K-Means e Variações na inicialização

Neste trabalho, o algoritmo mais explorado foi o k-means e suas variações. Na aplicação do método k-means como inicialização, foi utilizado o k-means++ e também a inicialização aleatória. Foram explorados diferentes quantidades de *clusters* iniciais e foi utilizado o método *Elbow* para verificar a melhor quantidade de *clusters*, além de

avaliar o resultado do ponto de vista hidrológico.

Observou-se que, ao aplicar o algoritmo k-means nos dados do estudo de caso, quando o início é dado de forma aleatória, por vezes apresenta diferentes resultados ao fim da clusterização. Em virtude da inicialização do k-means, que pode alterar significativamente os resultados (conforme apresentado no Capitulo 2), foram desenvolvidos dois métodos diferentes para alterar a inicialização do k-means, de forma a iniciar com centroides baseados em regiões com uma melhor qualidade do ponto de vista de homogeneidade hidrológica. Estes métodos serão descritos a seguir.

4.2.2 k-means com realocação de itens baseada no teste H

Neste método, proposto neste trabalho, é feita uma exploração simples do k-means alterando apenas um item de cada *cluster*, realocando para outro *cluster*, com o intuito de melhorar a clusterização. Este processo será dividido em três passos.

Como primeiro passo, executa-se o método k-means com uma inicialização do tipo k-means++, até o final de seu processo normal.

Após esta execução, a partir dos *clusters* resultantes, é feita a realocação de itens, que se pode definir como o segundo passo. Esta realocação para outro *cluster* é feita apenas em *clusters* que não forem hidrologicamente homogêneos, conforme o teste H (HOSKING; WALLIS, 1997), exposto no Capítulo 3. Nela - na realocação -, um item pode ser movido para outro *cluster* ou removido, caso não seja encontrado um *cluster* adequado para este.

Para definir qual item remover dos *clusters* ainda não homogêneos é utilizada a medida de discordância (HOSKING; WALLIS, 1997), exposta também no Capitulo 3. O item mais discordante será, então, movido ou removido do *cluster*. Quando da escolha do novo *cluster* para o qual este item será movido, é feita uma simulação adicionando este item em cada um dos demais clusters. A cada novo *cluster* formado, é feito o teste H; o *cluster* em que este item causar maiores melhoras ao teste H será o *cluster* a que este item será atribuído. Caso não seja observada melhora, este item é apenas removido do *cluster* original.

Após a execução do passo 2 em todos os *clusters* que não eram hidrologicamente homogêneos, inicia-se o passo 3, no qual se repete o método original do k-means, porém esta vez com a inicialização dos centroides com base nos centroides dos *clusters* formados com os itens realocados no passo 2.

E, por fim, os passos 2 e 3 são repetidos até que não haja mais mudanças nos *clusters* formados.

Este método pode ser melhor visualizado no algoritmo apresentado na Figura 6. Para simplificar a representação, são assumidos como métodos *k-means* (método k-means original), *k-meansCentroids* (método k-means original com a definição dos centroides), *maisDiscordante* (teste de discordância de HOSKING; WALLIS (1997),

buscando o item mais discordante), *testeH* (teste de heterogeneidade de HOSKING; WALLIS (1997)).

Algorithm 1:

```
Data: dados: Dados necessários para a clusterização
Result: clusters: Variável com os clusters formados ao final do processo
// passo 1
clusters = k-means(dados);
clustersAnteriores = [];
while clustersAnteriores <> clusters do
    clustersAnteriores = clusters
    // passo 2
    for pos = 0; pos < tamanho(clusters); pos++ do
        if testeH(clusters[pos]) >= 2 then
            maisDiscordante = maisDiscordante(clusters[pos]);
            posDiferencaH = Null;
            melhorDiferencaH = 0;
            for pos2 = 0; pos2 < tamanho(clusters); pos2++ do
                clusterSimulacao = clusters[pos2];
                clusterSimulacao.add(maisDiscordante);
                if (testeH(clusters[pos2]) - testeH(clusterSimulacao)) >
                  melhorDiferencaH then
                    posDiferencaH = pos2;
                end
            end
            if melhorDiferencaH > 0 then
                clusters[pos2] = clusterSimulacao;
            end
            clusters[pos].remove(maisDiscordante);
        end
    end
    // passo 3
    clusters = k-meansCentroids(clusters)
end
clusters
```

Figura 6 – K-means com realocação de itens baseada no teste H

Este método tem por objetivo realocar os itens buscando o cluster mais adequado, explorando uma deficiência do método k-means original – deficiência essa que tem por característica o fato de que quando ocorre uma inicialização ruim, acaba por não conseguir corrigir o erro ao longo da sua execução. Contudo, com o método aqui exposto, ocorrendo uma realocação dos itens, é possível corrigir esta deficiência, de modo que se obtenha um *cluster* mais adequado em cada item mais discordante encontrado apenas nos *clusters* não-hidrológicamente homogêneos. Este fluxo está representado na figura 7, exibida a seguir. A realocação dos itens entre os clusters faz uma

pequena alteração a cada iteração. Porém, com muitas várias iterações, alcançam-se resultados bem diferentes e bem interessantes se comparados com o resultado inicial do k-means. Este método também busca trazer uma nova perspectiva de execução do algoritmo k-means, incorporando uma função de avaliação em cada cluster e realocando os itens com base nesta função de avalição.

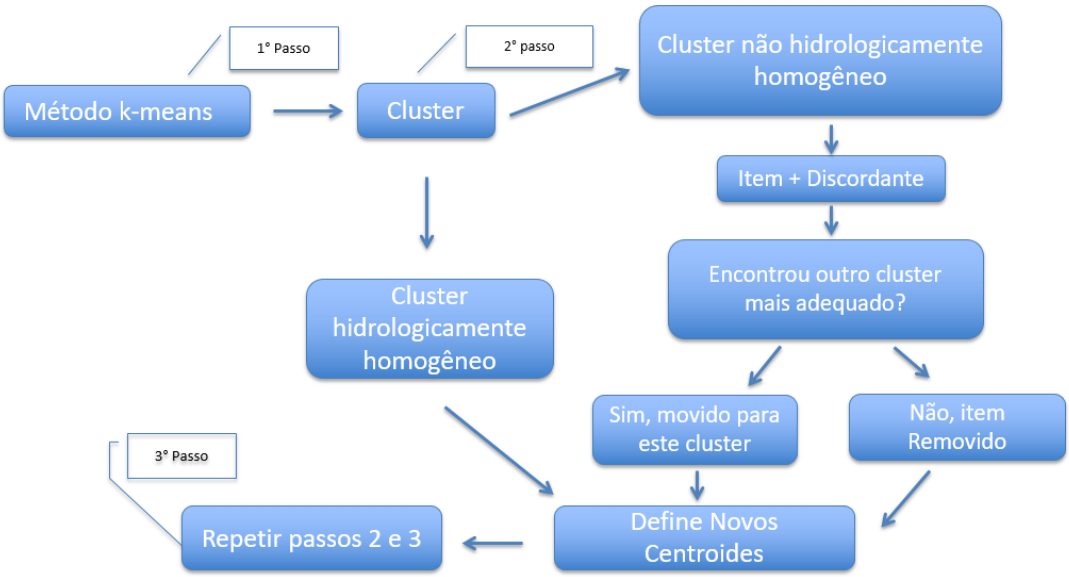

Figura 7 – Fluxograma representativo do método K-means com realocação de itens baseada no teste H

4.2.3 K-means com remoção de itens discordantes

Este método, proposto neste trabalho, realiza-se uma adaptação da inicialização do k-means utilizando a remoção de todos os itens discordantes do *cluster*. O processo é dividido em três passos, descritos a seguir.

No passo 1, é executado o método k-means com uma inicialização do tipo k-means++, até o final de seu processo normal.

No passo 2, é realizado, a partir dos *clusters* resultantes do primeiro passo, a remoção de todos os itens mais discordantes (do ponto de vista hidrológico) para cada *cluster*. Para a remoção dos itens discordantes, é utilizada a medida de discordância HOSKING; WALLIS (1997), exposta no capítulo 3. A remoção dos itens mais discordantes é feita apenas nos *clusters* que não atendem aos critérios de homogeneidade do teste H(HOSKING; WALLIS, 1997), também exposto no Capítulo 3, feita item a item mais discordante, e, depois, executando o teste H a cada remoção, até que atenda o critério de homogeneidade ($H < 2$) ou atenda ao critério de parada de quantidade mínima de itens na região. Esta quantidade mínima dependerá da aplicação a ser utilizada e dos dados trabalhados; para os fins deste trabalho, foi estipulado

que o critério de parada é de uma região de no mínimo 7 itens, por ser um tamanho de região suficiente para os cálculos utilizados no laboratório LSHPD.

Após a execução do passo 2 em todos os *clusters* que não eram hidrologicamente homogêneos, no passo 3 é repetido o método original do k-means, porém, esta vez, com a inicialização dos centroides com base nos centroides dos *clusters* formados e itens removidos no passo 2. Dessa forma, os passos 2 e 3 são repetidos até que não haja mais mudanças nos *clusters* formados em relação a iteração anterior.

Este método pode ser melhor visualizado no algoritmo apresentado na Figura 8. Para simplificar a representação, são assumidos como métodos *k-means* (método k-means original), *k-meansCentroids* (método k-means original com a definição dos centroides), *maisDiscordante* (teste de discordância de HOSKING; WALLIS (1997), buscando o item mais discordante), *testeH* (teste de heterogeneidade de HOSKING; WALLIS (1997)).

Algorithm 2:

```
   Data: dados: Dados necessários para a clusterização
   Data: tamanhoMinimo: Tamanho mínimo de itens em um cluster
   Result: clusters: Variável com os clusters formados ao final do processo
 1 // passo 1
 2 clusters = k-means(dados);
 3 clustersAnteriores = [];
 4 while clustersAnteriores <> clusters do
 5 |   clustersAnteriores = clusters
 6 |   // passo 2
 7 |   for pos = 0; pos < tamanho(clusters); pos++ do
 8 |   |   while testeH(clusters[pos]) >= 2 and tamanho(clusters[pos]) >
   |   |    tamanhoMinimo do
 9 |   |   |   maisDiscordante = maisDiscordante(clusters[pos]);
10 |   |   |   clusters[pos].remove(maisDiscordante);
11 |   |   end
12 |   end
13 |   // passo 3
14 |   clusters = k-meansCentroids(clusters)
15 end
```

Figura 8 – Método Método K-means com Remoção de itens discordantes de um cluster

Assim como no método anteriormente descrito, este método busca suprir uma das ineficiências na inicialização do k-means, que pode fazer uma formação dos conjuntos iniciais de forma equivocada e acabar por perpetuar este problema até o final de sua execução. Com a remoção dos itens discordantes, busca-se encontrar melhores resultados, de forma a remover o item mais discordante dos grupos que não alcançaram os requisitos mínimos de homogeneidade definidos por (HOSKING; WALLIS, 1997).

Outra representação para este método está disposta na figura 9, que expõe, através de um fluxograma, os passos executados pelo método.

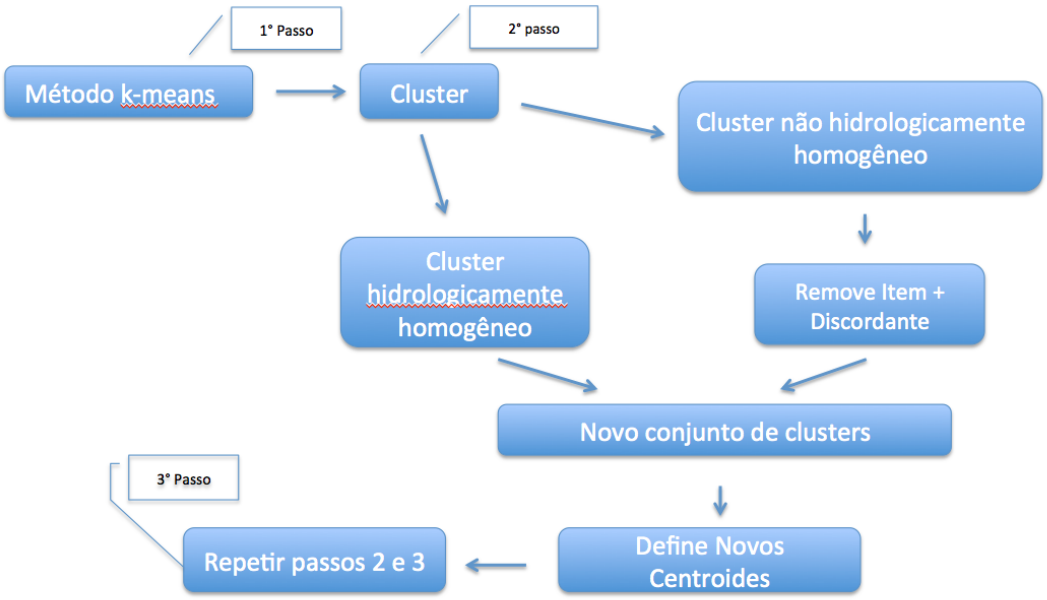

Figura 9 – Fluxograma representativo do método Método K-means com Remoção de itens discordantes

Além do método k-means nativo e a exploração de variações, foram também aplicados os métodos de clusterização *Aglomerative Clustering* e *Affinity Propagation*. Além destes, também foi explorada uma variação da técnica de regionalização por áreas de influência, que será explicada a seguir.

4.2.4 Clusterização combinada com ROI

Na regionalização por áreas de influência (BURN, 1990) é criada uma região para cada área de drenagem. Neste presente trabalho, propõe-se a aplicação deste conceito para a tentativa de uma regionalização de melhor qualidade; para tanto, é explorado um misto de técnicas de clusterização com a regionalização por áreas de influência. Uma característica importante na clusterização é que um item pertence apenas a um *cluster*, entretanto, observa-se que, para o problema de regionalização de bacias hidrográficas, para fins de cálculos e previsões estatísticas, não seria uma questão que uma área de drenagem pertença a mais de uma região.

Então, neste método proposto, pretende-se preencher as lacunas de uma clusterização convencional, com a regionalização por áreas de influência, permitindo que um item pertença a mais de um *cluster*, ou seja, que uma área de drenagem possa pertencer a mais de uma região.

O objetivo deste método é encontrar a maior quantidade de regiões hidrologicamente homogêneas. Espera-se alcançar esta finalidade combinando os resultados de

diferentes técnicas de clusterização em conjunto com o resultado da regionalização por áreas de influência. Este método será explicado pelos passos a seguir.

Como passo 1, são aplicados diferentes métodos de clusterização e são encontrados e armazenados os diferentes resultados.

No passo 2, são combinados os resultados sob duas perspectivas: na perspectiva **a**, é priorizado o quesito de menor quantidade de itens repetidos nas combinações, com o objetivo de construção de um mapa visual de regionalização; e, na perspectiva **b**, é priorizada a formação de maior quantidade de regiões com melhor qualidade de homogeneidade hidrológica com base no teste H (HOSKING; WALLIS, 1997).

No passo 3, após resultados da combinação do passo 2, são então preenchidas as lacunas com resultados da regionalização por áreas de influência, atendendo às perspectivas **a** e **b**.

Em todos os passos, executa-se um processo de remoção dos itens discordantes. conforme a medida de discordância de HOSKING; WALLIS (1997), e são consideradas apenas regiões com o módulo do valor do teste H (menor que 2), também de HOSKING; WALLIS (1997).

Este método apresenta grande potencial, uma vez que propõe uma busca de refinamento dos resultados anteriormente obtidos, e, por conseguinte, ao combiná-los com o método ROI, é possível melhorá-los significativamente, como se pode observar no capítulo 5. Essa melhora ocorre porque, na formação de *hard clusters* (EVERITT et al., 2011), não há ocorrências de um mesmo elemento em mais de um cluster. Entretando, com o método proposto, essa ocorrência é possível, de forma a buscar, para cada série de dados históricos – removidos para que a região atendesse os critérios de homogeneidade (HOSKING; WALLIS, 1997) –, a formação de uma nova região que atenda aos critérios de homogeneidade necessários, ou seja, quando o valor obtido no teste H é <= 2.

4.3 Armazenamento dos dados

O armazenamento de dados foi feito no formato JSON, organizado conforme o tipo de algoritmo de clusterização, os atributos utilizados, os resultados obtidos em cada passo da execução e os dados tabulados e refinados ao final do processo.

A modelagem dos dados ficou organizada conforme a Figura 10, que contém um diagrama de entidade e relacionamento (ER). No diagrama, observamos as entidades Cenário de Testes, Iteração, Cluster e ClusterItem. Cada Cenário de Testes tem *N* iterações, cada iteração corresponde a um resultado parcial ou final da execução do algoritmo. Cada Iteração tem *N clusters* e cada *cluster N* ClusterItens, onde cada um representa uma área de drenagem.

A estrutura proposta foi transposta para o formato JSON, onde cada chave estran-

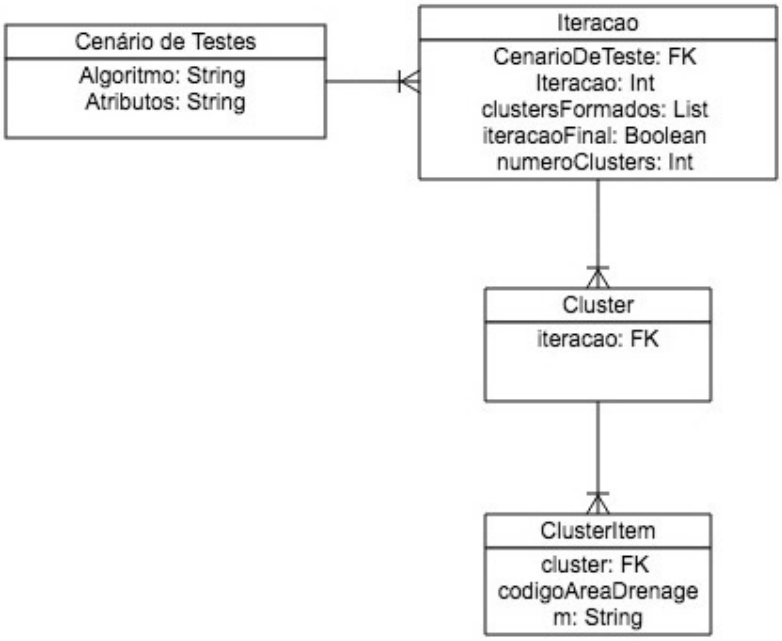

Figura 10 – Organização e armazenamento dos dados

geira se torna uma lista de itens dentro da entidade de relacionamento, a qual está ligada em uma relação $1:N$. Ou seja, cada objeto do Cenário de Testes passou a ter uma lista de iterações, tendo, cada Iteração, uma lista de *clusters* e cada Cluster tendo uma lista de ClusterItens.

4.4 Processamento e tabulação dos dados

Em um primeiro momento, foram explorados apenas os métodos de clusterização tradicionais e observados os resultados. Neste ponto, observou-se que era necessário fazer ajustes nas regiões formadas para atenderem aos critérios de homogeneidade de HOSKING; WALLIS (1997). Neste ponto, desenvolveu-se um algoritmo que faz a remoção de todos os itens discordantes do *cluster*, baseados na medida de discordância de HOSKING; WALLIS (1997), até que o *cluster* passe a ser hidrologicamente homogêneo.

Após, foram exploradas as combinações entre os algoritmos e, então, desenvolvidos os métodos apresentados neste trabalho. Em cada etapa de processamento, foi feita a observação das regiões formadas sob os seguintes aspectos:

- Quantidade de itens presentes no *cluster*, nas áreas de drenagem presentes na região formada originalmente.
- Quantidade de itens presentes no *cluster*, nas áreas de drenagem presentes na região formada após a remoção dos itens discordantes.

- Valor do teste H de HOSKING; WALLIS (1997) para cada região formada originalmente.
- Valor do teste H de HOSKING; WALLIS (1997) para cada região formada após a remoção dos itens discordantes.

Após este enfoque, cada cenário de testes foi qualificado em relação à:

- quantidade de áreas de drenagem presentes em regiões hidrologicamente homogêneas, onde $H < 1$.
- quantidade de áreas de drenagem presentes em regiões potencialmente homogêneas, onde $H > 1$ e $H < 2$.
- quantidade de áreas de drenagem descartadas, onde $H > 2$.

Outro fator importante para a qualidade de uma região formada é que ela tenha um número expressivo de elementos para garantir uma quantidade suficiente de dados, para que seja possível fazer os cálculos estatísticos. Essa quantidade pode variar conforme a aplicação, porém, é definido que, para a maioria das aplicações, a quantidade média de 500 anos de séries de dados é suficiente (HOSKING; WALLIS, 1997; RAO; SRINIVAS, 2008). Considerando que os dados disponíveis para o estudo de caso, tem em média 80 anos por cada região, a formação de mais de 15 regiões faz com que muitas delas não alcancem o total de 500 anos de séries de dados, não formando, consequentemente, regiões expressivas. Em razão desta necessidade, os testes elaborados com os algoritmos k-means e aglomerative clustering, que permitem a parametrização de número de *clusters*, utilizou o intervalo de 2 à 15 clusters.

No Capítulo seguinte estão apresentados os resultados obtidos em cada etapa deste trabalho, as comparações entre os resultados e as análises que levaram a busca e elaboração dos métodos propostos.

5 RESULTADOS

Neste Capítulo, são apresentados e comparados os resultados obtidos ao longo do trabalho, exibindo-os de forma a distinguir dois cenários de atributos, onde um utiliza atributos selecionados, e o outro utiliza todos os atributos. Para a seleção dos atributos, foram escolhidos aqueles com maior correlação com vazões máxima, conforme método apresentado no Capítulo 4, sendo eles: Altitude média (E), Declividade média (S), Comprimento do canal principal (L) e Coeficiente de compacidade adimensional (kc).

5.1 Resultados obtidos com o k-means

O primeiro método explorado neste trabalho foi o k-means, com as variações entre inicialização aleatória e pela heurística k-means++. Foram variados também os atributos utilizados para a formação dos *clusters* e a quantidade de *clusters* utilizados.

Na Tabela 3, estão os resultados do método k-means++ em relação ao número de *clusters*, representado pela variável *n_clusters*. Os valores presentes como resultados referem-se à quantidade de séries de dados que ficaram em regiões hidrologicamente homogêneas, ou seja, que o módulo de H é menor que 1; ou, então, que sejam potencialmente homogêneas, na qual o módulo de H seja menor ou igual a 2.

Observa-se que o melhor resultado em regiões com H menor que 1 foi a quantidade de 4 *clusters*, que apresentou 28 áreas de drenagem pertencendo à regiões hidrologicamente homogêneas, e 14 potencialmente homogêneas, totalizando em 42 Séries de dados históricos com valor $H <= 2$. Em outra perspectiva, temos bons resultados também com 7 *clusters*, onde temos 54 séries de dados com valor $H <= 2$.

Nestes resultados, foram separados os dados presentes em clusters iniciais, sem nenhuma remoção de séries de dados discordantes, e os dados após a remoção das séries discordantes com o intuito de melhorar o valor de H.

Tabela 3 – Quantidade de itens por valor H e por número de *clusters* – k-means++ com atributos selecionados

clusters		2	3	4	5	6	7	8	9	10	11	12	13	14
Cluster Inicial	H < 1	0	0	0	0	0	0	0	0	0	0	8	12	0
	1 <= H <= 2	0	0	0	0	0	6	6	5	5	13	6	11	20
	H > 2	106	106	106	106	106	100	100	101	101	93	92	83	86
Com Séries Removidas	H < 1	17	0	28	12	7	0	8	0	0	0	16	20	0
	1 <= H >= 2	0	12	14	37	30	54	6	20	20	22	6	22	20
	H > 2	26	14	7	7	14	21	42	53	53	53	61	42	65
Total	H <= 2	17	12	42	49	37	54	14	20	20	22	22	42	20

Já na Tabela 4, estão apenas os melhores resultados de cada um dos quatro cenários elencados a seguir: (I) inicialização k-means++ com atributos selecionados; (II) inicialização k-means++ com todos os atributos; (III) inicialização k-means aleatório com atributos selecionados; e, (IV) inicialização k-means aleatório com todos os atributos. Além disso, a Tabela mostra os dois melhores resultados em cada cenário, conforme a quantidade de *clusters*. Os cenários que apresentaram melhores resultados quanto à quantidade total de séries de dados com valor $H < 2$ foram os cenários I, com 7 *clusters*, e II com 6 *clusters*. Já quanto ao critério de maior quantidade de séries de dados com $H < 1$ foi o cenário III com 4 *clusters*.

Tabela 4 – Comparação dos melhores resultados obtidos com o k-means

Cenário	I	I	II	II	III	III	IV	IV
clusters	4	7	11	6	4	3	9	14
H $<$ 1	28	0	0	18	41	28	26	23
1 $<=$ H $<=$ 2	14	54	43	33	0	12	15	27
H $>$ 2	7	21	25	7	7	12	21	37
Total H $<=$ 2	42	54	43	51	41	40	41	50

5.2 Resultados obtidos com o *affinity propagation* e *aglomerative clustering*

Além do k-means, mais dois métodos de clusterização foram aplicados. Um destes foi o *affinity propagation* (AP), que trabalha com propagação de afinidades como descrito no Capítulo 2. Na Tabela 5, estão relacionados os resultados obtidos em dois cenários, um com os atributos selecionados e outro com todos os atributos disponíveis. Obteve-se melhor resultado com os atributos selecionados, e o algoritmo formou de 11 *clusters* neste cenário, uma vez que, diferente do k-means, no AP não é parametrizado o número de *clusters*.

Tabela 5 – Comparação dos melhores resultados obtidos com o *affinity propagation*

Cenário		I	II
clusters		11	14
Cluster Inicial	H $<$ 1	0	0
	1 $<=$ H $<=$ 2	16	21
	H $>$ 2	90	85
Com Séries Removidas	H $<$ 1	8	0
	1 $<=$ H $<=$ 2	40	37
	H $>$ 2	38	44
Total	H $<=$ 2	48	37

O outro método explorado foi o *Aglomerative Clustering* (AC), que se trata de um método hierárquico, também explicado no Capítulo 2. Assim como o k-means, o AC

permite a parametrização do número de *clusters*, por isso foi explorado, assim como na aplicação do k-means, verificou-se o comportamento na variação entre 2 e 14 *clusters*. Assim como os demais métodos, também foi explorada a variação entre os atributos selecionados e todos os atributos. Foi constatado, no que se referiu ao cenário I, no qual os atributos foram selecionados, houve melhores resultados, chegando a 52 séries de dados com $H <= 2$, sendo eles expostos na Tabela 6.

Tabela 6 – Comparação dos melhores resultados obtidos com o *aglomerative clustering*

Cenário		I	I	II	II
clusters		5	7	11	13
Cluster Inicial	H $<$ 1	0	0	0	0
	1 $<=$ H $<=$ 2	0	0	8	8
	H $>$ 2	106	106	98	98
Com Séries Removidas	H $<$ 1	13	0	0	0
	1 $<=$ H $<=$ 2	11	52	45	30
	H $>$ 2	21	15	35	58
Total	H $<=$ 2	24	52	45	30

5.3 Resultados obtidos com K-means e variações na inicialização

Conforme explicado no Capítulo 4, dois métodos para variações na inicialização do k-means são propostos neste trabalho, que foram nomeados como *k-means com realocação de itens* e *k-means com remoção de itens discordantes*. Para ambos os métodos, foram executados testes com variações de números de *clusters* entre 2 e 14, também variando entre atributos selecionados e entre todos os atributos.

Os melhores resultados para o método *k-means com realocação de itens* estão expostos na Tabela 7, divididos em dois cenários: I) um no qual são utilizados apenas atributos selecionados; e II) outro no qual se utiliza todos os atributos. Nos resultados presentes na Tabela 7, observa-se como melhor resultado para séries de dados hidrologicamente homogêneas a quantidade de 41, formando 2 *clusters*.

Na Tabela 8, estão os resultados para o método *k-means com remoção de itens discordantes*, método no qual todos os itens discordantes são removidos a cada iteração e utilizados como centroides para realimentar o k-means. Assim como o método anterior, este também foi dividido em dois cenários: I) cenário em que são utilizados atributos selecionados; e, II) cenário em que são utilizados todos os atributos.

Como melhores resultados, neste método foram encontrados, no cenário I, com 7 *clusters*, 15 séries de dados hidrologicamente homogêneas somado a 37 potencialmente homogêneas, totalizando em 52 séries de dados com $H <= 2$; e, também com 7 *clusters*, 55 séries de dados potencialmente homogêneas. Já, no cenário II, os

Tabela 7 – Comparação dos melhores resultados obtidos com o k-means utilizando realocação de itens

Cenário		I	I	II	II
clusters		2	5	10	4
Cluster Inicial	H $<$ 1	0	0	0	0
	1 $<=$ H $<=$ 2	0	0	12	0
	H $>$ 2	106	106	94	106
Com Séries Removidas	H $<$ 1	41	12	22	0
	1 $<=$ H $<=$ 2	0	37	28	51
	H $>$ 2	0	7	21	7
Total	H $<=$ 2	41	49	50	51

melhores resultados foram de 47 séries de dados com $H <= 2$ sendo destas 20 com $H < 1$; e, 46 séries de dados com $H <= 2$ sendo destas 11 com $H < 1$.

Tabela 8 – Comparação dos melhores resultados obtidos com k-means com remoção de itens discordantes

Cenário		I	I	II	II
clusters		7	7	6	5
Cluster Inicial	H $<$ 1	0	0	0	0
	1 $<=$ H $<=$ 2	0	0	0	0
	H $>$ 2	106	106	106	106
Com Séries Removidas	H $<$ 1	15	0	20	11
	1 $<=$ H $<=$ 2	37	55	27	35
	H $>$ 2	21	14	7	7
Total	H $>$ 2	52	55	47	46

5.4 Comparação entre os resultados de Clusterização

Com o intuito de melhor visualizar os resultados, foi feita uma comparação entre os mesmos dos diferentes métodos de clusterização e podem ser observados no gráfico apresentado na Figura 11. Nesta comparação, foram escolhidos os melhores resultados de cada método de clusterização observando dois critérios: a quantidade de séries de dados com valor de $H < 1$ e a quantidade de séries de dados com valor de $H < 2$. Os métodos comparados foram: k-means, *Affinity Propagation* (AP), *Aglomerative Clustering* (AC), k-means com realocação de itens (k-means RI) e k-means com remoção de itens discordantes (k-means RID).

Observou-se como melhor resultado, referente à quantidade de séries de dados hidrologicamente homogêneas, o método k-means RI com 41 séries de dados com valor de H menor que 1. Outrora, observando na perspectiva do total de séries potencialmente hidrologicamente homogêneas – ou seja, com o valor de H menor que 2 –, o melhor resultado foi obtido com o método k-means RID com 55 séries de dados

potencialmente homogêneas.

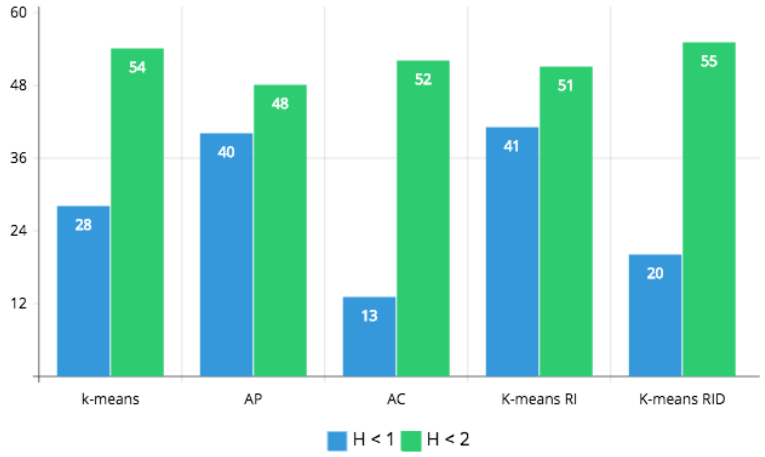

Figura 11 – Gráfico de comparação entre os resultados de clusterização

5.5 Resultados obtidos com clusterização e ROI

Como descrito no Capítulo 4, neste trabalho é proposta uma combinação dos resultados entre clusterização e regionalização por ROI. Neste método, selecionou-se o melhor resultado por clusterização e buscou-se o complemento dos resultados nas séries de dados que foram removidas por não contemplarem o critério de homogeneidade de HOSKING; WALLIS (1997).

Sendo assim, foi criada uma região de influência para cada uma destas séries de dados e, então, buscou-se a formação de regiões com a maior quantidade de itens, repetindo menos as áreas de drenagem. Os resultados selecionados foram os de maior quantidade de séries de dados com valor de $H < 1$ e de maior quantidade de séries de dados com valor de $H < 2$.

Na Tabela 9, estão apresentados os resultados da implementação deste método, com a perspectiva de maior quantidade de séries de dados com valor de $H < 1$. Diferente dos métodos de clusterização apresentados anteriormente, neste método há casos onde um item poderá estar em mais de um *cluster*, ou seja, uma série de dados presente em mais de uma região formada.

Para a obtenção destes dados, foram escolhidas as regiões formadas por ROI em que aparecesse a maior quantidade de séries de dados removidas do resultado de clusterização escolhido, porém, com a menor repetição de séries. O resultado de clusterização escolhido foi o método k-means RI, com 41 séries de dados com valor de H menor que 1.

Podemos observar o aumento para 67 séries únicas, com o valor de $H < 1$, obtidos com a formação de 11 regiões que são apresentadas nos mapas presentes nas Figuras 13 e 14. Nos mapas representando as regiões, os pontos em preto representam

Tabela 9 – Regiões formadas pelo melhor resultado obtido com clusterização e ROI na perspectiva de $H < 1$

Região	Series	H
1	24	0,8712466064
2	17	0,7219383312
3	14	0,630466818
4	17	0,2629357614
5	7	-0,07011467912
6	12	0,6846722673
7	7	-0,9481916304
8	21	0,8860988498
9	9	0,8023604323
10	10	0,9883880622
11	17	0,9785138913
Total de Séries com $H < 1$		67
Séries repetidas		46

as séries de dados históricos de vazões em cursos de água que forma aquela região. Cada um desses pontos corresponde à medição no principal curso d'água na área de drenagem, dados esses que são subsídios para o cálculo do teste H de HOSKING; WALLIS (1997).

Dentre as 67 séries de dados, 46 aparecem em ao menos duas regiões. Para a aplicação do método ROI, foram utilizados os atributos selecionados, visto que estes apresentaram melhores resultados em todas as clusterizações. Após formadas as regiões de influência, foram removidos os itens discordantes e formadas regiões hidrologicamente homogêneas.

Na Tabela 10, podemos observar o resultado com a perspectiva de maior quantidade de séries de dados com valor de $H < 2$. Sendo os resultados obtidos da mesma forma que a perspectiva anterior, o resultado de clusterização escolhido para compor a primeira parte de resultados foi o método k-means RID com 55 séries de dados com valor de H menor que 2.

Observa-se, neste resultado da Tabela 10, apenas 19 séries de dados com $H < 1$, porém 96 séries de dados com $H < 2$. Para tal, foi necessária a formação de 19 regiões, nas quais houve a repetição de 62 séries de dados em ao menos duas regiões.

As regiões formadas com ROI acabam por repetir elementos em regiões, o que pode trazer vantagens ou desvantagens conforme a aplicação. A vantagem se apresenta tendo em vista que o aspecto de encontrar a região com melhor valor de H para uma área de drenagem se facilita, já que há mais de uma opção de região para buscar uma área de drenagem. Entretanto, quando se trata de uma representação visual em um mapa, a sobreposição de séries de dados em diferentes regiões formadas pode

Tabela 10 – Regiões formadas pelo melhor resultado obtido com clusterização e ROI na perspectiva de $H < 2$

Região	Series	H
1	13	1,968225838
2	18	1,484267014
3	9	1,052020973
4	15	1,515356146
5	15	1,988720569
6	19	1,554820634
7	20	1,597016084
8	23	1,411799217
9	8	1,90952889
10	12	0,6846722673
11	13	1,722409643
12	17	1,487843587
13	7	-0,07011467912
14	7	1,764892583
15	8	1,468050407
16	11	1,872803879
17	14	1,548468078
18	17	1,728011737
19	19	1,884300638
Total de Séries com H < 1		19
Total de Séries com H < 2		96
Séries repetidas		62

dificultar uma boa representação.

Conforme a Figura 12, ao comparar os resultados obtidos neste método e nos outros métodos implementados neste trabalho, observamos um salto nos resultados de 38,6% para 63,2% de aproveitamento das regiões formadas, partindo do ponto de vista de regiões hidrologicamente homogêneas, ou seja, que tenham o valor de H menor que 1, dentre as 106 séries de dados disponíveis. Do ponto de vista de regiões potencialmente hidrologicamente homogêneas, houve um salto de 51,9% de aproveitamento para 90,5%.

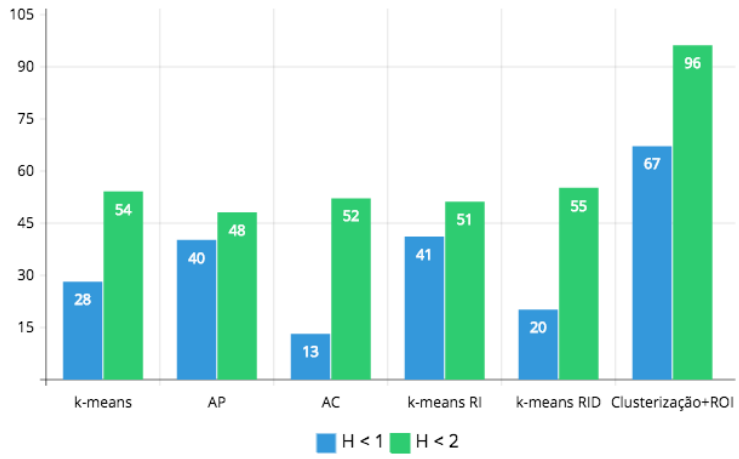

Figura 12 – Gráfico de comparação entre os resultados finais

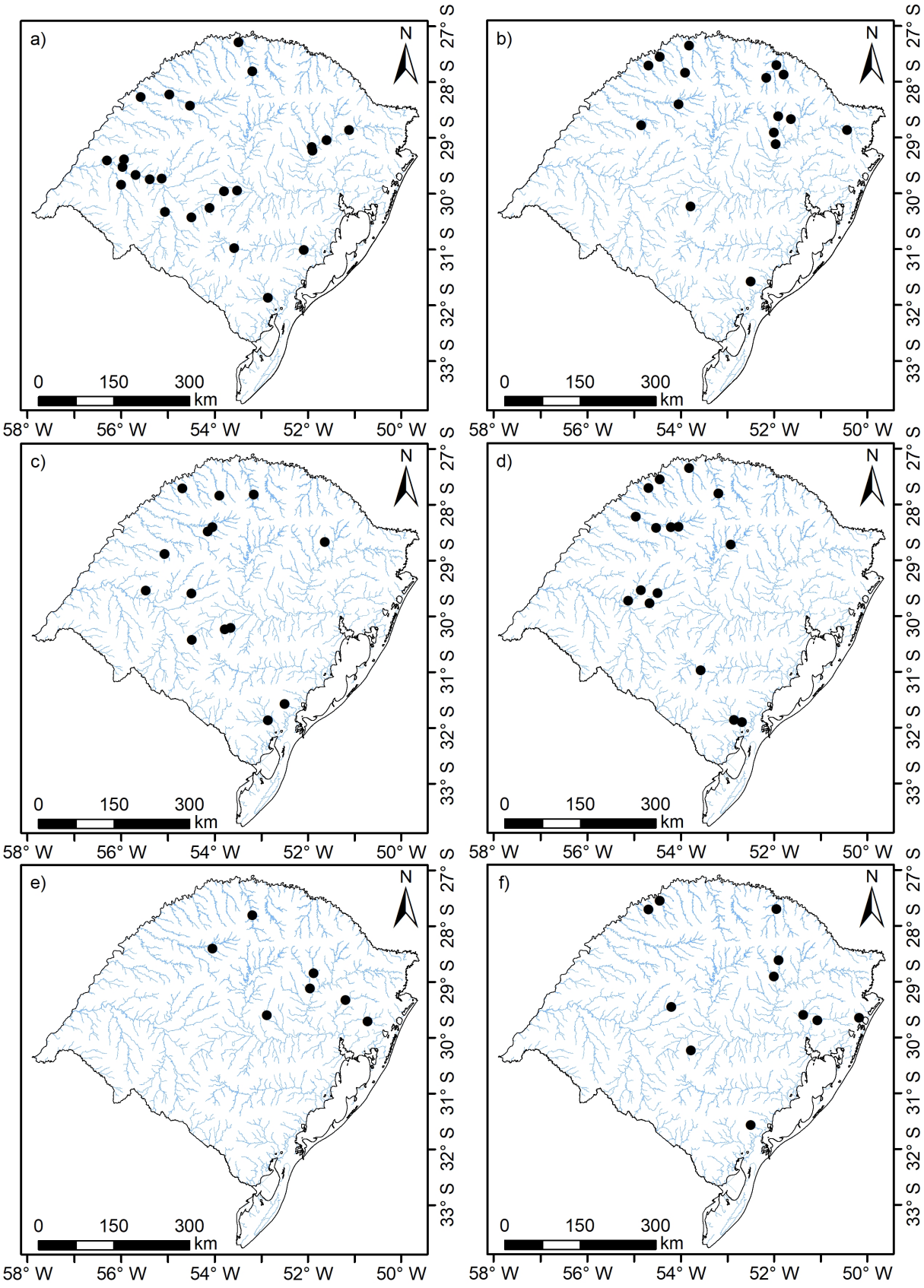

Figura 13 – Mapas com as regiões 1, 2, 3, 4, 5 e 6 formadas pela clusterização com ROI na perspectiva $H < 1$

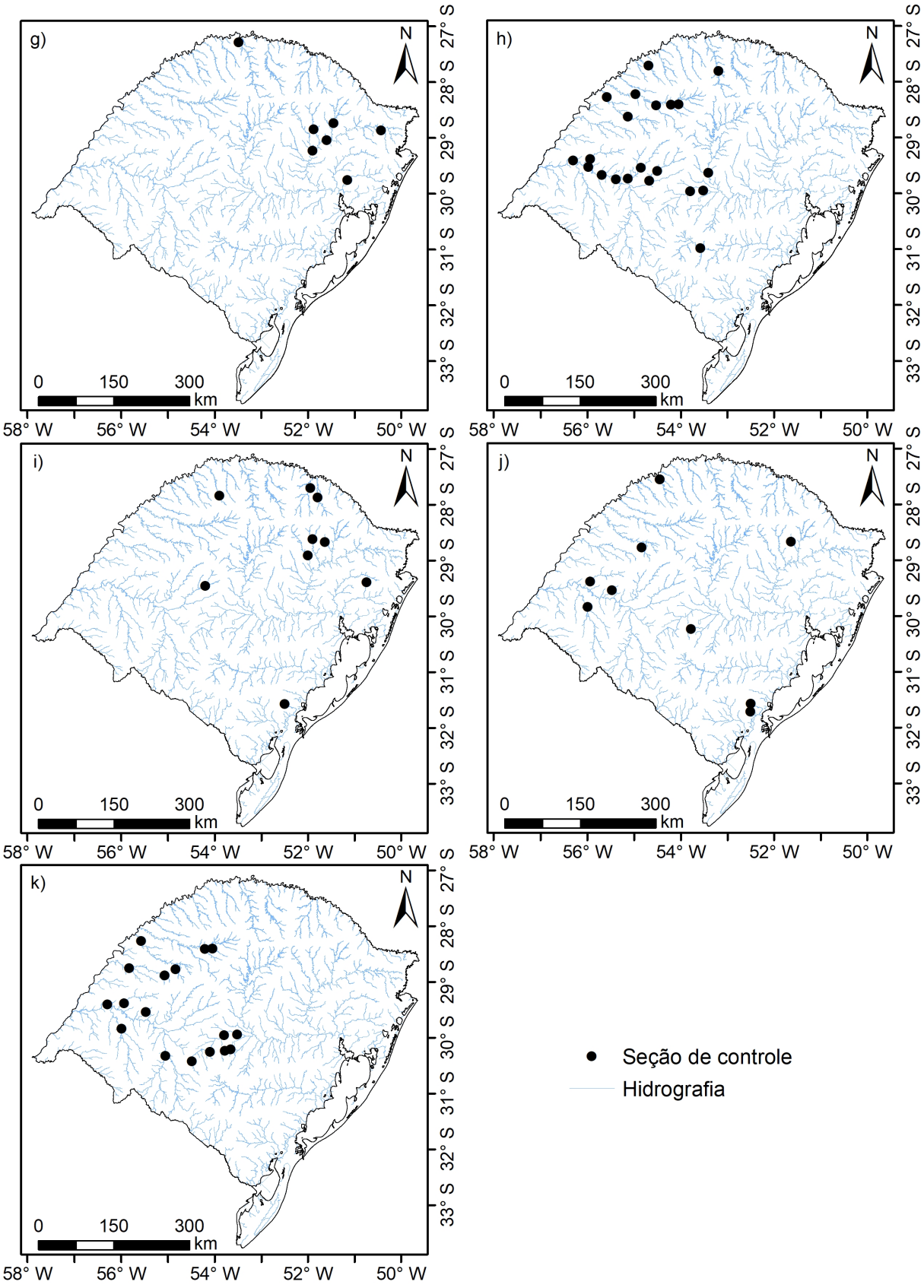

Figura 14 – Mapas com as regiões 7, 8, 9, 10 e 11 formadas pela clusterização com ROI na perspectiva $H < 1$

6 CONCLUSÃO

Neste trabalho, foram exploradas técnicas de clusterização para a regionalização de bacias hidrográficas. A formação de regiões tem vasta aplicação na área da hidrologia aplicada, pois é possível fazer previsões de comportamentos hidrológicos, como vazão ou estiagem em cursos d'água. O principal desafio na regionalização de bacias hidrográficas é a formação de regiões que sejam hidrologicamente homogêneas, critério definido por (HOSKING; WALLIS, 1997) pelo teste H. Para a solução deste problema, a aplicação de técnicas de clusterização tem demonstrado bons resultados.

Para tanto, executaram-se os métodos de clusterização *k-means*, *aglomerative clustering* e *affinity propagation* e também fez-se a proposição de três métodos para clusterização e formação de regiões hidrológicamente homogêneas, os quais são: *k-means* com realocação de itens, *k-means* com remoção de itens discordantes e clusterização com ROI. Os métodos envolvendo o *k-means* fazem uma variação na inicialização do *k-means* com base nos testes de heterogeneidade e discordância de HOSKING; WALLIS (1997). O método de clusterização com ROI propõe a utilização do resultado de um método de clusterização, complementado pela busca de outras regiões nas séries que não formarem regiões hidrologicamente homogêneas, complementando e melhorando o resultado anteriormente obtido.

Dentre os métodos trabalhados, os que obtiveram melhores resultados na clusterização foram (I) o *k-means* com realocação de itens para séries de dados com o valor de H menor que 1 e (II) *k-means* com remoção de itens discordantes para séries de dados com valor de H menor que 2, chegando a um percentual de aproveitamento das séries de 38,6% no primeiro método e 51,9% no segundo método.

Já, com o complemento das séries não aproveitadas, utilizando o método ROI, obtiveram-se os resultados de 63,2% de aproveitamento no critério em que o valor de H é menor que 1 e de 90,5% no critério em que o valor de H é menor que 2. Para estes resultados, foram utilizados dois cenários de atributos: o primeiro, o cenário que utiliza atributos selecionados conforme a correlação com variáveis relacionadas a inundação; e, segundo, cenário que utiliza todos os atributos disponíveis neste trabalho. Os resultados obtidos com o cenário I foram, em todas as aplicações, melhores

que os resultados partindo do cenário II.

Conclui-se que houve uma melhora com as variações do método *k-means* em relação aos outros métodos de clusterização testados, com diferença nos resultados em pouco mais de 1%. Já, utilizando o método de complementação com o método ROI, observou-se uma melhora muito significativa nos resultados; houve, porém, sobreposição entre as regiões formadas, repetindo séries de dados entre regiões, o que dificultou a representação em mapas, entretanto, de qualquer forma, notou-se aplicações estatísticas que podem trazer benefícios. Neste contexto, os métodos propostos atenderam o objetivo geral deste trabalho de propor métodos utilizando de IA que obtivessem melhores resultados que outros métodos existentes.

Este trabalho foi aplicado ao estudo de caso das bacias hidrográficas do estado do Rio Grande do Sul, Brasil, e acredita-se que melhores resultados poderão ser obtidos na aplicação destas técnicas em dados de áreas maiores, tais como de um país ou continente, o que poderia complementar de forma significativa para a validação dos resultados deste trabalho em um futuro. Outra perspectiva de trabalhos futuros é o da exploração de outras técnicas de IA, como Algoritmos Genéticos e Redes Neurais, para as quais se prevê aplicação em estudos de caso com uma maior quantidade de dados.

REFERÊNCIAS

ARTHUR, D.; VASSILVITSKII, S. k-means++: The Advantages of Careful Seeding. **SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Pages 1027-1035**, [S.l.], 2007.

BESKOW, S. et al. Artificial intelligence techniques coupled with seasonality measures for hydrological regionalization of Q90 under Brazilian conditions. **Journal of Hydrology, Volume 541, Part B, October 2016, Pages 1406-1419**, [S.l.], 2016.

BURN, D. H. An appraisal of the "region of influence" approach to flood frequency analysis. **Hydrological Sciences Journal, 35:2, 149-165, DOI: 10.1080/02626669009492415**, [S.l.], 1990.

CHA, S.-H. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. **Issue 4, Volume 1, 2007, Pages 300-307**, [S.l.], 2007.

CORREA, L. **Implementação de Análise de Técnicas de inteligência Artificial Aplicados a Clusterização em Recursos Hídricos**. 2014. 96p. Trabalho de Conclusão (Curso de Ciência da Computação) — Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, RS.

DALRYMPLE, T. Flood frequency analyses. **US Geol. Surv. Wat. Supply Pap. 1543-A, 1960**, [S.l.], 1960.

DASH, M.; LIU, H. Feature Selection for Classification. **Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997**, [S.l.], 1997.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. **Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231**, [S.l.], 1996.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. King's College London, UK: John Wiley and Sons, Ltd, 2011. 330p.

FREY, B. J.; DUECK, D. Clustering by passing messages between data points. **Science. 2007 Feb 16;315(5814):972-6. Epub 2007 Jan 11.**, [S.l.], 2007.

HADDAD, K.; RAHMANB, A. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. **Journal of Hydrology - Volumes 430–431, 2 April 2012, Pages 142-161**, [S.l.], 2012.

HOLLAND, J. H. **Adaptation in natural and artificial systems**: An introductory analysis with applications to biology, control, and artificial intelligence. MA, USA: MIT Press Cambridge, 1992. 228p.

HOSKING, J. R. M.; WALLIS, J. R. **Regional Frequency Analysis - An Approach Based on L-Moments**. Cambridge, UK: Cambridge University Press, 1997. 224p.

HUSSAIN, Z.; PASHA, G. R. Regional flood frequency analysis of the seven sites of Punjab, Pakistan, using L-moments. **Water Resour Manag 23(10):1917-1933. doi: 10.1007/s11269-008-9360-7**, [S.l.], 2009.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Englewood Cliffs, New Jersey, US: Prentice Hall, 1988. 320p.

JOHNSON, S. C. Hierarchical clustering schemes. **PSYCHOMETRIKA, Vol 32, No. 3, September, 1967**, [S.l.], 1967.

KAUFMAN, L.; ROUSSEEUW, P. J. Clustering by Means of Medoids. **Statistical Data Analysis Based on the L1–Norm and Related Methods, pp. 405–416**, [S.l.], 1987.

KINGSTON, D.; HANNAH, D. M.; LAWLER, D. M.; MCGREGOR, G. R. Regional classification. variability. and trends of northern North Atlantic river flow. **Hydrological Processes, v. 25, p. 1021-1033, 2011. doi: 10.1002/hyp.7655.**, [S.l.], 2011.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of Cluster in K-Means Clustering. **International Journal of Advance Research in Computer Science and Management Studies(IJARCSMS), ISSN: 2321-7782, Volume 1, Issue 6, November 2013, p 90-95**, [S.l.], 2013.

KRISHNA, K.; MURTY, M. N. Genetic K-means algorithm. **Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, v. 29, n. 3, p. 433-439**, [S.l.], 1999.

LANCE, G. N.; WILLIAMS, W. T. Computer programs for hierarchical polythetic classification ("similarity analyses"). **The Computer Journal, Volume 9, Issue 1, 1 May 1966, Pages 60–64**, [S.l.], 1966.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281–297**, [S.l.], 1967.

MAESSCHALCK, R. D.; JOUAN-RIMBAUD, D.; MASSART, D. The Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems, Volume 50, Issue 1, 4 January 2000, Pages 1-18**, [S.l.], 2000.

MILLER, H.; HAN, J. **Geographic Data Mining and Knowledge Discovery**: an overview. Bristol, PA, USA: Taylor & Francis, Inc., 2001. 338p.

NOTO, L. V.; LOGGIA, G. L. Use of L-moments approach for regional flood frequency analysis in Sicily, Italy. **Water Resources Management, v. 23, n. 11, p. 2207–2229**, [S.l.], 2009.

RAO, A. R.; SRINIVAS, V. **Regionalization of Watersheds - An Approach Based on Cluster Analysis**. Purdue University, West Lafayette, IN, USA and Department of Civil Engineering, Indian Institute of Science (IISc), Bangalore, India: Springer Science+Business Media B.V., 2008. 241p.

SCHIAVETTI, A.; CAMARGO, A. **CONCEITOS DE BACIAS HIDROGRÁFICAS - TEORIAS E APLICAÇÕES**. Ilhéus, Bahia, Brasil: EDITUS - UESC, 2002. 293p.

SMYTH, P. Clustering using Monte Carlo Cross-Validation. **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp 126-133**, [S.l.], 1996.

Proposta de métodos de clusterização de dados com validação por testes de heterogeneidade e discordância aplicados à regionalização de bacias hidrográficas – Leroi Floriano de Oliveira

UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Dissertação

Proposta de métodos de clusterização de dados com validação por testes de heterogeneidade e discordância aplicados à regionalização de bacias hidrográficas

LEROI FLORIANO DE OLIVEIRA

Pelotas, 2018