UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

USO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE ÁREAS HIDROLOGICAMENTE HOMOGÊNEAS

ROBERTA BRITO NUNES DINIZ

Dissertação apresentada ao Programa de Pós-Graduação em Informática da UFPB como requisito parcial para obtenção do título de Mestre em Informática.

Orientadora: Profa. Dra. Valéria Gonçalves Soares

Segundo Orientador: Prof. Dr. Lucídio dos Anjos Formiga Cabral

João Pessoa

Junho de 2009.

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de Roberta Brito Nunes Diniz, candidata ao Título de Mestre em Informática na Área de Sistemas de Computação, realizada em 26 de junho de 2009.

Aos vinte e seis dias do mês de junho do ano dois mil e nove, às quatorze horas, no Auditório do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar a candidata ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa "*Computação Distribuída*", a Sra. Roberta Brito Nunes Diniz. A comissão examinadora foi composta pelos professores doutores: Valéria Gonçalves Soares (DI-UFPB), Orientador e Presidente da Banca Examinadora, Lucídio dos Anjos Formiga Cabral (DI-UFPB), como segundo orientador, Gustavo Henrique Matos Bezerra Motta (DI-UFPB) como examinador interno e Anne Magály de Paula Canuto (UFRN), como examinadora externa. Dando início aos trabalhos, a Profª. Valéria Gonçalves Soares, cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra à candidata para que a mesma fizesse, oralmente, a exposição do trabalho de dissertação intitulado "*USO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE ÁREAS HIDROLOGICAMENTE HOMOGÊNEAS*". Concluída a exposição, a candidata foi argüida pela Banca Examinadora que emitiu o seguinte parecer: "*aprovada*". Assim sendo, deve a Universidade Federal da Paraíba expedir o respectivo diploma de Mestre em Informática na forma da lei e, para constar, eu, professor José Antônio Gomes de Lima, membro do Colegiado deste Programa, representando a coordenação do PPGI, lavrei a presente ata que vai assinada por mim mesmo e pelos membros da Banca Examinadora. João Pessoa, 26 de junho de 2009.

José Antônio Gomes de Lima

Profª. Dra. Valéria Gonçalves Soares
Primeiro Orientador (DI-UFPB)

Prof. Dr. Lucídio dos Anjos Formiga Cabral
Segundo Orientador (DI-UFPB)

Prof. Dr. Gustavo Henrique Matos B. Motta
Examinador Interno (DI-UFPB)

Profª. Dra. Anne Magály de Paula Canuto
Examinador Externo (UFRN)

Ao meu esposo Laudízio e a meus filhos Ananda, Matheus e Arthur,

Aos meus pais, Agenor e Oneide, a quem devo muito,

Dedico esta conquista.

AGRADECIMENTOS

À Deus, por ter me sustentado cada dia e pelas ricas bênçãos recebidas das suas poderosas mãos.

Aos meus pais, Agenor e Oneide, por terem nos oferecido uma boa educação, muitas vezes fazendo sacrifícios, pelo incentivo e por serem um exemplo de persistência diante das dificuldades. Aos meus irmãos pelo carinho e pela força transmitidos em todos os momentos.

Ao meu esposo Laudízio, pelo incentivo, amor, paciência e apoio incondicional que foram fundamentais no cumprimento desta etapa da minha vida, e aos meus filhos, Ananda, Matheus e Arthur incentivadores incansáveis da minha luta, renovando a cada dia as minhas forças e me fazendo ver que vale a pena lutar.

À professora Valéria Gonçalves, minha orientadora, por acreditar na minha capacidade, pela amizade e pelo conhecimento prestado durante todo o período da pós-graduação.

Ao professor Lucídio Formiga Cabral, pela dedicação na co-orientação, que, com sua experiência e visão prática dos assuntos, possibilitou o êxito deste trabalho.

À professora Anne Magáli de Paula Canuto, pela sua disponibilidade, interesse, atenção e colaboração na fase final da minha dissertação.

À Maria Alice e Isabel pela amizade e pelas palavras de encorajamento que tanto me ajudaram nessa jornada.

Ao professor Antônio Marcos Moreira pela orientação nas análises estatísticas realizadas.

À Wastenayda Patrício pela disponibilidade em elaborar os mapas do Estado da Paraíba.

A todos, que de uma forma ou de outra, estiveram presentes e me auxiliaram na elaboração deste trabalho.

RESUMO

A escassez de dados fluviométricos e a má qualidade dos dados existentes sobre os cursos d'água do Nordeste do Brasil têm obrigado os especialistas em hidrologia a buscar novos caminhos, ampliando assim os conhecimentos e metodologias sobre como desenvolver a região com base em suas limitações ambientais. Uma dessas metodologias consiste na utilização de técnicas de regionalização hidrológica, que possibilitam a transferência de dados e informações entre bacias com características similares.

Nesse contexto, este trabalho visa identificar regiões hidrologicamente homogêneas no Estado da Paraíba, utilizando Mineração de Dados, através da técnica de Clusterização, possibilitando assim a identificação de padrões que permitam a transposição de dados de uma região para outra. Foram utilizados algoritmos com métodos baseados em partição, métodos hierárquicos e métodos baseados em redes neurais.

Foram empregados dados de 41 bacias hidrográficas localizadas no Estado da Paraíba. Para todas as bacias foram determinadas 32 características físicas e climatológicas, sendo seis características das medidas lineares das bacias, três de forma, nove da rede de drenagem, sete do relevo, duas da capacidade de escoamento, três das classes de solos e duas da climatologia.

Foram compostos quatro cenários para a execução dos algoritmos dos quais três foram obtidos através da técnica de seleção de atributos. Para avaliar os grupos obtidos pelos algoritmos três índices de validação foram aplicados, a saber índice de Silhouette, índice Davies-Bouldin e índice Dunn. Os resultados da validação estatística mostraram que o algoritmo Ward se destacou na clusterização das 41 bacias hidrográficas, identificando 6 regiões hidrologicamente homogêneas no Estado da Paraíba.

Os resultados obtidos mostraram a viabilidade do uso de técnicas de mineração de dados em estudos de regionalização hidrológica e aplicações práticas em engenharia de recursos hídricos no âmbito do Estado da Paraíba.

Palavras-chave: mineração de dados; clusterização; regionalização hidrológica.

ABSTRACT

The lack of fluviometric data and the bad quality of existing data related to watercourses of Brazilian northeast have obligated hydrology specialists to search for new solutions, improving the knowledge and methodologies to build the region up according to its environment limitations. Through the hydrologic regionalization techniques it's possible to transfer data and information among similar watersheds.

In this context, the purpose of this work is to identify hydrologically similar regions in the State of Paraiba using Clustering - a kind of data mining technique - to find patterns that allow data transposition from one region to other. Algorithms were used with methods based on partition, hierarchical methods, and methods based on neural networks.

It is used data of 41 watersheds located in the State of Paraíba. For all the watersheds, 32 physical and climatological characteristic were determined, being six characteristics of the linear measures of the watersheds, three of shape, nine of the drainage net, seven of the relief, two of the drainage capacity, three of the classes of soils and two of the climatology.

Four sceneries were composed for the execution of the algorithms of which three were obtained through the technique of selection of attributes. To evaluate the groups obtained by the algorithms three validation indexes were applied, namely index of Silhouette, index Davies-Bouldin and index Dunn. The results of the statistical validation showed that the algorithm Ward stood out in the clustering of the 41 watersheds, identifying 6 homogeneous regions in the State of Paraíba.

The obtained results showed the viability of the use of techniques of data mining in studies of hydrologic regionalization and practical applications in engineering of water resources in the ambit of the State of Paraíba.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CAPÍTULO I

INTRODUÇÃO

Uma rede de monitoramento hidrológico eficaz pressupõe a instalação de equipamentos de coleta, transmissão, tratamento, armazenamento e divulgação de dados hidrológicos de uma determinada região.

A escassez de dados para realizar estudos hidrológicos é uma realidade no nosso país e a implantação e operação de estações de medições em campo exige elevados investimentos financeiros nem sempre disponíveis, dada a dimensão continental do Brasil. As comunidades acadêmica e profissional da área de hidrologia têm buscado, através de pesquisas, ampliar os conhecimentos e metodologias para suprir essa deficiência. Uma dessas metodologias consiste na utilização da técnica de regionalização hidrológica que possibilita a transferência de dados e informações entre bacias de regiões diferentes, mas de características similares [Porto et al. 2004].

O desenvolvimento de métodos eficientes de regionalização hidrológica se torna cada vez mais necessário no Brasil, devido à implantação dos sistemas estaduais e federais de gerenciamento de recursos hídricos e da escassez de dados obtidos através de medições em campo. Entre tais sistemas, destacam-se os subsistemas de outorga e cobrança pelo uso da água bruta, o licenciamento ambiental e a elaboração dos planos diretores de bacias, que necessitam estimar curvas de permanência de vazões, na maioria das vezes sem dados disponíveis [Diniz 2008].

Os resultados da regionalização também são importantes para projetos tais como: dimensionamento de obras hidráulicas, dimensionamento do volume de reservatórios, estimativas de vazões de cheias, entre outros [Peralta 2003].

Do ponto de vista econômico uma boa regionalização permite a redução do número de estações de medição na medida em que a informação produzida em uma região pode ser usada em outras com características similares.

De acordo com a Agência Nacional de Águas – ANA [ANA 2007], a rede básica de monitoramento brasileira, apesar do crescente aumento verificado nos últimos anos, ainda é incipiente e bastante concentrada em algumas poucas regiões do País. Em números gerais, ela possui 9.324 estações fluviométricas, mas apenas 57% em operação, e 13.531 estações pluviométricas, mas com apenas 60% operando. Isto pode sinalizar falta de planejamento, insuficiência de recursos financeiros para implantação, manutenção e operação e, por conseguinte, escassez ou má qualidade dos dados.

No estado da Paraíba, de acordo com o Sistema HidroWeb - Sistema de Informações Hidrológicas - da ANA [ANA 2009], de um total de 72 estações fluviométricas apenas 14

disponibilizam informações confiáveis, num flagrante vazio de informações básicas em um Estado cuja disponibilidade hídrica é a menor da Federação [Brasil 2006].

1.1 MOTIVAÇÃO

Com a queda no custo de armazenamento dos dados e a rápida automatização das empresas tem se verificado um crescimento em larga escala da quantidade de dados armazenados em meios magnéticos. Informações úteis e interessantes podem ser obtidas a partir desses dados e isso tem chamado a atenção de empresas e instituições acadêmicas. Muitas vezes, estas informações não estão disponíveis devido à falta de ferramentas apropriadas para sua extração, pois devido à grande quantidade de dados existentes, esse processo se torna impossível de ser realizado por analistas através de métodos manuais tradicionais.

Uma nova área da Tecnologia da Informação surge como alternativa para atender, entre outras, essa necessidade: a Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases – KDD*). Esse processo permite analisar e utilizar de maneira útil os dados disponíveis para obter a informação desejada, a partir do uso de ferramentas de extração de conhecimento.

A mineração de dados, uma das etapas da Descoberta de Conhecimento em Base de Dados, tem avançado nestes últimos anos e tem despertado grande interesse junto às comunidades científica e industrial. Ela é considerada, atualmente, o ponto mais alto na busca de conhecimento para tomada de decisões [Cortês 2002]. Esta etapa é responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão [Silva 2004].

Várias técnicas de mineração de dados aumentam as possibilidades de se desenvolver métodos mais adequados para os estudos de regionalização hidrológica, fornecendo ao especialista do domínio, ferramentas que possibilitem uma maior segurança no processo de tomada de decisão.

Estudos desenvolvidos por pesquisadores mostraram que a análise de agrupamento (Clusterização) tem sido usada com sucesso na definição de regiões hidrologicamente homogêneas [Porto 2004]. A clusterização busca dividir um conjunto de objetos não rotulados em grupos, chamados de partição ou *cluster*, de maneira que os objetos pertencentes a um mesmo grupo apresentam alta similaridade entre si e baixa similaridade em relação aos objetos dos demais grupos. Esta técnica vem sendo utilizada em tarefas de exploração de dados e extrações de padrões [Metz 2006].

Motivado pelo que foi apresentado anteriormente e aliado à realidade que se tem na região semi-árida paraibana, quanto aos dados hidrológicos, nesta dissertação será aplicada a técnica de análise de agrupamento que não só permite a transferência de informação de uma área para outra,

mas também o preenchimento de falhas e a recuperação de dados de tal forma que os projetos de engenharia possam ser melhor qualificados com otimização de custos de implantação, operação e manutenção. Uma outra motivação para este estudo é o fato de não haver, na literatura, nenhum trabalho em que esse tema seja abordado no âmbito do Estado da Paraíba.

1.2 **OBJETIVOS**

1.2.1 **Objetivo Geral**

Apresentar uma proposta de metodologia com base na utilização da análise de agrupamento (Clusterização) para identificação de áreas hidrologicamente homogêneas no Estado da Paraíba. A metodologia será aplicada sobre os dados de 41 bacias hidrográficas do Estado da Paraíba.

1.2.2 **Objetivo Específico**

- Levantamento das características físicas e climatológicas das bacias hidrográficas;
- Seleção das características que melhor descrevem cada bacia;
- Aplicar algoritmos de Clusterização;
- Determinar o grau de significância dos resultados obtidos pelos algoritmos de clusterização aplicando a validação estatística, e selecionar o método mais adequado aos estudos de regionalização hidrológica no âmbito do Estado da Paraíba;
- Examinar os grupos encontrados pelo algoritmo para tentar descobrir significados que estejam relacionados ao domínio da aplicação.

1.3 **ESTRUTURA DO TRABALHO**

Neste capítulo foi apresentado o contexto em que se insere este trabalho, bem como a motivação e o objetivo do mesmo. O restante da dissertação está organizada da maneira descrita a seguir.

No capítulo 2 é descrito, de modo geral, o processo de Descoberta de Conhecimento em Base de Dados, abordando todas as suas etapas. Como a técnica de Clusterização será utilizada neste trabalho, também é apresentada neste capítulo uma definição da técnica bem como algoritmos que representam as abordagens particional, hierárquica e baseada em modelos que serão utilizados no processo de agrupamento.

O capítulo 3 descreve a metodologia aplicada nessa pesquisa. É feita uma descrição do conjunto de dados utilizado, apresentando as características físicas e climatológicas das bacias

hidrográficas. São descritas também as atividades realizadas em cada uma das etapas do processo de Descoberta de Conhecimento em Base de Dados.

O capítulo 4 apresenta os resultados obtidos e a partir da análise dos mesmos, é selecionado o método mais adequado na identificação de regiões hidrologicamente homogêneas no Estado da Paraíba.

O capítulo 5 apresenta a conclusão do trabalho e as perspectivas para novas pesquisas.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

2.1 DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS

Segundo Fayyad et al. [Fayyad 1996], *Knowledge Discovery in Databases* (KDD) ou descoberta de conhecimento em bases de dados é o processo não trivial de identificação, a partir de dados, de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis.

O termo não trivial significa que envolve algum mecanismo de busca ou inferência, e não qualquer processamento direto de uma quantidade pré-definida de dados. Nessa definição, um conjunto de dados representa fatos enquanto que os padrões podem ser interpretados como uma expressão, em alguma linguagem, capaz de descrever um subconjunto de dados ou um modelo aplicável a este subconjunto. Os padrões descobertos devem ser válidos diante de novos dados com algum grau de certeza. Estes padrões podem ser considerados conhecimento dependendo de sua natureza. Os padrões devem ser novos, compreensíveis e úteis, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para tomada de decisão [Romão 2002].

O processo de descoberta de conhecimento em bases de dados tem como objetivo a utilização de mecanismos automáticos de extração de conhecimento que auxilie na tomada de decisões. A Figura 2.1 ilustra o processo de extração de conhecimento a partir de dados.

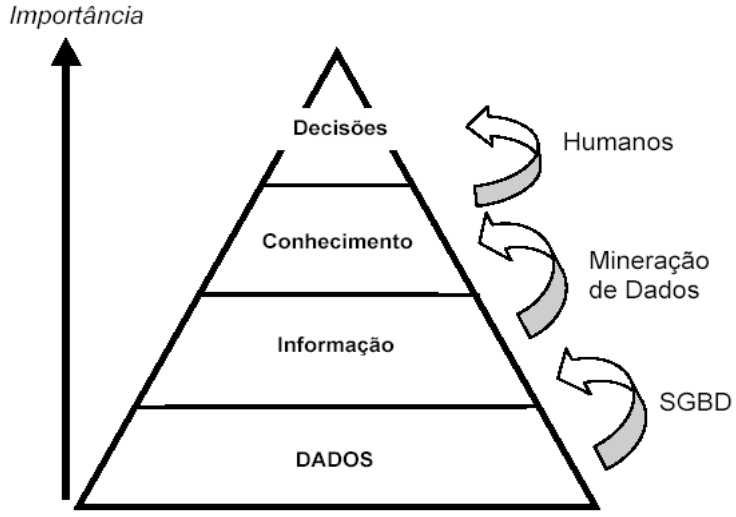


Figura 2.1: **Obtenção de conhecimento para tomada de decisões [Romão 2002].**

Na base do triângulo estão os *dados*, os quais tomam o maior volume da memória do computador, e oferecem pouca utilidade estratégica na hora de se tomar decisões. A partir dos

dados é possível obter muita *informação* através de aplicativos desenvolvidos para fins específicos ou através das ferramentas dos Sistemas Gerenciadores de Banco de Dados (SGBD) que exigem conhecimento das mesmas por parte do analista para se obter o máximo proveito da montanha de dados disponíveis e em crescimento. A partir das informações ou dos próprios dados é possível extrair um tipo de informação mais completa, o *conhecimento*, normalmente mais resumido e em menor quantidade, mas de maior inteligibilidade para se tomar decisões. Finalmente, no topo do triângulo da Figura 2.1 [Romão 2002], aparecem as *decisões* realizadas pelo homem com base no conhecimento obtido pelas ferramentas de Mineração de Dados (MD) e que serão apresentadas mais adiante.

O processo de descoberta de conhecimento em bases de dados evoluiu, e continua evoluindo, através da interseção de vários campos de pesquisa: aprendizado de máquina, reconhecimento de padrões, bancos de dados, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados, e computação de alto-desempenho [Fayyad 1996]. O KDD focaliza o processo global de descoberta de conhecimento através dos dados e é composto de várias etapas, conforme a Figura 2.2, as quais serão descritas a seguir.

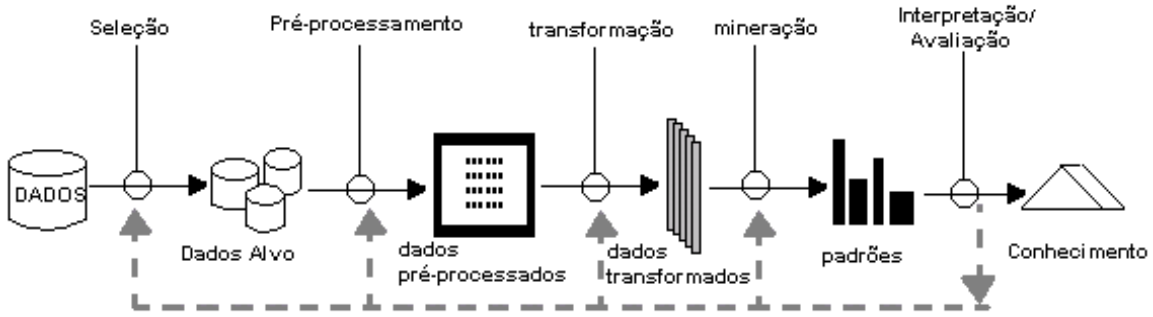


Figura 2.2 – **Etapas que compõem o processo de KDD [Fayyad 1996].**

O processo de descoberta de conhecimento em bases de dados é iterativo podendo conter repetições entre qualquer dos passos, buscando sempre aprimorar os resultados a cada iteração. A Mineração de Dados é a etapa-chave no processo de descoberta de conhecimento em bases de dados [Fayyad 1996]. Nela acontecem consultas complexas à base de dados e também são desvendados relacionamentos até então escondidos no banco de dados, no intuito de permitir ao analista inferir regras ou comprovar hipóteses. Esta etapa é vista como um processo iterativo e focado na interação entre usuários, especialistas do domínio e responsáveis pela aplicação.

2.1.1 **Etapas do processo de KDD**

2.1.1.1 **Seleção**

Nesta etapa ocorre a seleção dos dados que serão utilizados em todo o processo. Os dados não necessariamente estão em um banco de dados, eles podem ser extraídos de planilhas eletrônicas, de formulários de pesquisas, tabelas, mapas entre outras fontes.

2.1.1.2 **Pré-processamento**

É nesta etapa que os dados são preparados para a modelagem. Problemas de redundância, inconsistência e ausência de valores devem ser resolvidos. Essa etapa possui fundamental relevância no processo de descoberta de conhecimento, é no pré-processamento que se garante a qualidade dos dados analisados, tratando os possíveis problemas encontrados e, consequentemente, assegurando maior fidelidade aos resultados obtidos pelos algoritmos de mineração.

Um dos processos mais importantes executados na etapa de pré-processamento é a escolha dos atributos que serão submetidos aos algoritmos de mineração. Esse processo consiste em identificar quais atributos são redundantes e irrelevantes no conjunto dos dados. Segundo Hall [Hall 1999], este processo reduz a dimensionalidade dos dados e permite que os algoritmos de mineração operem de forma rápida e eficiente.

2.1.1.3 **Transformação dos dados**

Etapa onde os dados são transformados num formato apropriado para serem submetidos aos algoritmos de mineração. Algumas transformações podem ser aplicadas aos dados, tais como: normalização, discretização e conversão de atributos.

2.1.1.4 **Mineração**

Etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões que interessam. Esta etapa será detalhada mais adiante.

2.1.1.5 **Avaliação ou Pós-processamento**

É a fase onde são identificados os padrões interessantes de acordo com algum critério do usuário. O especialista em mineração de dados juntamente com o especialista do domínio da aplicação avaliam os resultados obtidos e validam os padrões encontrados, para uma posterior utilização prática dos mesmos e/ou definem novas alternativas de investigação nos dados.

2.2 **MINERAÇÃO DE DADOS**

O termo Mineração de Dados é definido como sendo aplicação de algoritmos específicos para extração de padrões nos dados [Fayyad 1996]. A extração do conhecimento a partir dos dados armazenados é uma ferramenta importante em um processo decisório. Esse conhecimento descoberto deve ser compreensível e útil para os usuários finais, pois servirá de suporte no processo de tomada de decisão. O papel da mineração de dados é prover as ferramentas que melhoram a compreensão e inspiram a criatividade baseado em observações nos dados [BL 2004]. O ambiente ideal para mineração de dados é provido de pessoas cuja habilidade em processamento e mineração de dados só é ultrapassada pelo entendimento de como funciona a área afim e suas perspectivas para o futuro. O grupo de mineração de dados inclui peritos em banco de dados, programadores, estatísticos, analistas de sistemas, e analistas de negócios, todos trabalhando em conjunto para assegurar que as decisões empresariais sejam baseadas em informações precisas. Este time de pessoas tem a habilidade de propagar tudo o que eles possam aprender de uma determinada área da organização [BL 2004].

A tecnologia de mineração de dados tem sido usada em diversas áreas com a finalidade de descrever características do passado, assim como predizer tendências para o futuro. Dentre as que têm explorado o uso desta tecnologia encontramos as áreas de marketing, vendas, finanças, energia, saúde, recursos hídricos, genética e biologia.

a) Marketing

Database marketing é um segmento emergente que vem modificando a forma de encarar e fazer a divulgação dos produtos de uma empresa. Quando aliado às técnicas de mineração de dados amplia suas potencialidades abrindo novas e diferentes formas de avaliar e melhorar a relação entre o cliente e o faturamento da empresa [Thomé 2002].

b) Vendas

No setor de vendas a aplicação provavelmente de maior interesse seja a de identificar produtos que possam ser colocados em uma mesma cesta ou pacote. Isto envolve a garimpagem por associação entre produtos, que pode revelar afinidades ou aversões nunca imaginadas e como conseqüência, sugerir estratégias para maximizar o lucro [Thomé 2002].

c) Finanças

As aplicações incluem a análise da avaliação para concessão de crédito a clientes; segmentação de contas a receber; análise de desempenho de investimentos financeiros como ações, bônus e fundos mútuos; avaliação de opções financeiras e detecção de fraudes [Cortês 2002].

d) Energia

Previsão de consumo e previsão de falhas em sistemas de transmissão ou de distribuição são as duas aplicações mais comuns, embora muitas outras tenham sido pesquisadas e difundidas na literatura [Thomé 2002].

e) Saúde

As aplicações incluem a análise da eficácia de certos tratamentos; otimização de processos dentro de um hospital; relacionamento de dados sobre o estado de saúde do paciente com a qualificação médica; e análise de efeitos colaterais de drogas [Cortês 2002].

f) Recursos Hídricos

As técnicas de mineração de dados nesta área estão sendo utilizadas para: balizar as escolhas de previsões de vazões naturais em bases estocásticas, diminuindo assim a probabilidade de erros [Cataldi 2007]; e determinar regiões hidrologicamente homogêneas [Júnior 2005].

g) Genética

As bases de dados genéticos apresentam, na sua grande maioria, um grande número de atributos (genes), inviabilizando a análise humana sem algum suporte tecnológico. Dentre muitas outras aplicações na área da genética, as técnicas de mineração de dados têm contribuído para redução de dimensionalidade em bases de expressão gênica o que permite uma maior confiabilidade nos resultados obtidos [Borges 2006].

h) Biologia

A classificação de padrões de dados ecológicos é um exemplo de aplicação na área de biologia onde técnicas de mineração de dados têm atuado com sucesso, como por exemplo, classificação das espécies de peixes em categorias tróficas [Francisco 2004].

2.3 **IMPORTÂNCIA DO USO DA MINERAÇÃO DE DADOS**

O grande volume de dados armazenados nas empresas e instituições acadêmicas vem sendo muito valorizado e analisado, pois muitas informações realmente novas e interessantes estão contidas nessas bases de dados, como perfis de clientes no uso de cartão de crédito (que podem ser usados para combater fraudes), padrões de pacientes que desenvolveram doenças (que podem ser úteis na tentativa de formular hipóteses e antecipar tratamentos), perfis de compra de clientes (para usar em futuras promoções).

A competitividade existente no mercado tem levado as empresas a buscarem cada vez mais subsídios para darem suporte ao processo de tomada de decisão, pois o uso da informação pode gerar conhecimento que ajude na análise de padrões históricos para conseguir uma previsão dos fatos futuros.

As bases de dados são a memória das empresas, mas memória sem inteligência tem pouco uso [BL 2004]. Segundo Berry e Linoff [BL 2004] o contexto ideal para a tecnologia de mineração de dados é uma organização que aprecie o valor da informação.

A tecnologia de mineração de dados visa explorar essas bases de dados para obter, de forma automática, valiosas informações que poderão causar diferenciais efetivos nos negócios. Para isso, a mineração de dados fornece um conjunto de atividades específico para alcançar os objetivos das empresas no que tange à análise dos seus dados.

Essas atividades podem ser classificadas em dois tipos [BL 2004]:

Mineração de Dados Supervisionada: usada quando se tem o conhecimento do que se procura exatamente, o objetivo é usar os dados disponíveis para construir um modelo que descreva uma variável particular de interesse em relação ao resto dos dados;

Mineração de Dados Não Supervisionada: usada quando o objetivo for encontrar padrões nos dados e posteriormente determinar se esses padrões encontrados são importantes ou não, o objetivo é estabelecer algum relacionamento entre todas as variáveis disponíveis nos dados.

A partir destas duas classificações, temos as seguintes tarefas de Mineração de Dados:

- Supervisionada
    - Classificação: consiste em examinar as características de um registro da base de dados, atribuindo a ele uma classe pré-definida com características que definirão o objeto. A partir disto, cria-se um modelo para ser aplicado e enquadrar os objetos em alguma das classes criadas.
    - Estimação: consiste na análise de determinados dados, definindo-se um valor para alguma variável a ser consultada a partir de situações semelhantes já conhecidas. Assim, define-se uma probabilidade de determinada circunstância ocorrer.
    - Predição ou Previsão: consiste na tentativa de classificar um objeto de acordo com seus comportamentos passados. O histórico do objeto em questão é utilizado para a construção de um modelo que explica o seu comportamento atual.

- Não Supervisionada

    - Associação: serve para definir quais objetos têm associações mútuas, agrupando-os em categorias. Uma das aplicações mais usuais desta técnica é a venda *cross-selling*, na qual são identificados produtos considerados adequados para compra em conjunto com o produto em questão.
    - Clusterização ou Agrupamento: consiste em segmentar um determinado grupo em alguns grupos menores, nos quais os objetos são agrupados de acordo com suas características comuns, visando auxiliar a definição e entendimento de suas necessidades, para posteriormente tomar as decisões adequadas.

2.4 **CLUSTERIZAÇÃO**

A tarefa de clusterização, também chamada de Agrupamento, tem sido utilizada em tarefas de exploração de dados e extrações de padrões. A clusterização busca dividir um conjunto de objetos não rotulados em grupos, chamados de partição, de maneira que os objetos pertencentes a um mesmo grupo apresentam alta similaridade entre si e baixa similaridade em relação aos objetos dos demais grupos.

A clusterização está normalmente associada com a análise exploratória (formulação de hipóteses e tomada de decisão), pois envolve problemas em que há pouca informação a priori acerca dos dados, e poucas hipóteses podem ser sustentadas. Portanto a clusterização pode fornecer novas hipóteses a respeito dos inter-relacionamentos dos dados e de sua estrutura intrínseca [MZ 2002].

A qualidade dos resultados obtidos por meio da clusterização é totalmente dependente da escolha de parâmetros como as medidas de similaridade e dos métodos de agrupamento utilizados.

Em geral, o processo de clusterização requer que o usuário determine qual o número de grupos a ser considerado. Com base neste número, os registros de dados são então separados nos grupos de forma que registros similares fiquem em um mesmo grupo e registros diferentes em grupos distintos. Uma vez, tendo esses grupos, é possível fazer uma análise dos elementos que compõem cada um deles, identificando as características comuns aos seus elementos e, desta forma, podendo criar rótulos que representem cada grupo [GP 2005].

Segundo Han e Kamber [HK 2000], as técnicas mais utilizadas para agrupar dados podem ser divididas em cinco categorias:

1. Métodos baseados em partição: baseia-se na construção de uma partição de um banco de dados em k grupos representados pelo valor médio dos objetos do grupo (*k_means*) ou por um objeto representativo do grupo que esteja localizado perto do centro do mesmo (*k-medoid*).
2. Métodos hieráquicos: esse método cria uma decomposição hierárquica de um dado conjunto de objetos e pode ser classificado como sendo aglomerativo (*bottom-up*) ou divisivo (*top-down*) de acordo com a maneira como a decomposição hierárquica é realizada.
3. Métodos baseados em densidade: a idéia geral desse método consiste no crescimento contínuo de um dado grupo até que a densidade (número de objetos) na vizinhança exceda um determinado limiar (*threshold*).
4. Métodos baseados em grade ou retícula (*grid based*): esse método quantifica o espaço do objeto em um número finito de células que formam uma estrutura de grade, onde todas as operações de agrupamento (*clustering*) são realizadas.
5. Métodos baseados em modelos: cria um modelo hipotético para cada grupo e a idéia geral é encontrar os objetos que mais se adaptem a esse modelo.

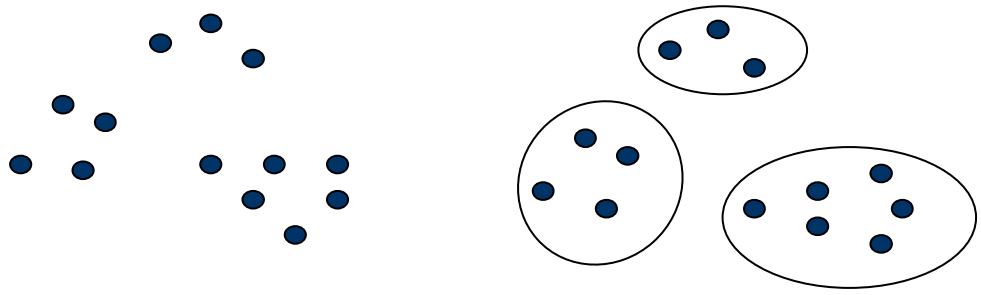
A seguir, descreveremos os métodos que serão utilizados no presente trabalho.

2.4.1 **Métodos baseados em partição**

O objetivo dessa abordagem é dividir a base de dados em *k* grupos, onde o usuário escolhe o número *k.* A divisão dos dados em grupos se dá sem sobreposição, ou seja, cada objeto está em exatamente um grupo, como pode ser observado na Figura 2.3. Inicialmente são escolhidos *k* objetos como sendo os centros dos *k* grupos. Os objetos são então divididos entre os *k* grupos de acordo com o cálculo da distância (medida de similaridade) adotada, de forma que cada objeto fique no grupo que forneça o menor valor de distância entre o objeto e o centro do mesmo. Segundo Metz

[Metz 2006], um problema associado ao *clustering* particional é a necessidade de informar com antecedência o número de grupo desejado. Além disso, apresenta alta variância, pois a seleção dos exemplos representantes afeta, significativamente, o resultado da clusterização e pode fazer com que a solução vá em direção a um máximo local da função de avaliação. Para minimizar esse efeito negativo, usualmente os algoritmos são executados diversas vezes com exemplos iniciais diferentes, e, então, a melhor solução é atribuída ao resultado do processo de *clustering*.

Os algoritmos ***K-Means*** e ***K-Medoids*** são os representantes mais utilizados dessa abordagem [Bogorny 2003]. Descreveremos a seguir o algoritmo *K-Means*, utilizado no presente trabalho.



Pontos Originais Agrupamento Particional

Figura 2.3 – **Agrupamento Particional**

2.4.1.1 **Algoritmo *K-Means***

O algoritmo *K-Means* é largamente utilizado na tarefa de clusterização e busca encontrar o melhor particionamento dos *n* objetos em *k* grupos. Ele exige a definição prévia do número de grupos (*k*) e os centros dos grupos, chamados de centróides, são inicialmente escolhidos de forma aleatória. Cada ponto é associado ao grupo com o centróide mais próximo. A cada iteração, o centróide de cada grupo é recalculado, sendo seu valor obtido pela média dos pontos do grupo. A distância Euclidiana, definida pela Equação 2.1, é a mais comumente utilizada para calcular a similaridade entre os objetos embora outros critérios possam ser empregados também.

$$dist(E_i, E_j) = \sqrt{\sum_{l=1}^{M} (x_{il} - x_{jl})^2} \qquad \text{Eq. 2.1}$$

Os passos básicos do algoritmo *K-Means* são:

1. escolha do valor de *k* e seleção de *k* objetos para serem centros iniciais dos *k* grupos.
2. cada objeto é associado a um grupo, para o qual a dissimilaridade entre o objeto e o centro do grupo é menor que as demais.
3. os centros dos grupos são recalculados, redefinindo cada um, em função dos atributos de todos os objetos pertencentes ao grupo.

4. retorna ao passo 2 até que os centros dos grupos se estabilizem.

A cada iteração, os objetos são agrupados em função do centro do grupo mais próximo e, por conseqüência, os centros dos grupos são reavaliados (passo 3). Isso provoca no espaço, o deslocamento dos centros médios, conforme pode ser observado nas Figuras 2.4 e 2.5. O algoritmo é interrompido quando as médias não mais são deslocadas, ou há uma insignificante realocação de objetos entre os grupos. Boa parte dos grupos já converge nos primeiros passos do algoritmo, ficando somente uma pequena quantidade que ainda se modificam.

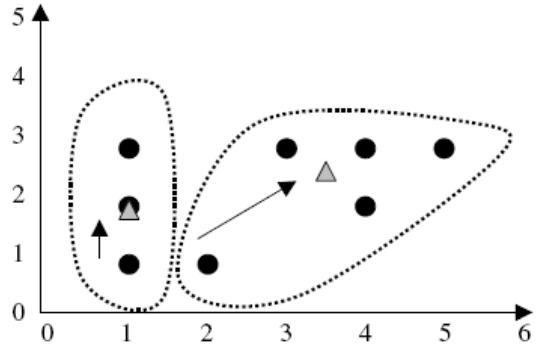


Figura 2.4 – ***Clusters*** **e contróides após o passo 3 da 1º iteração [Larose 2005]**

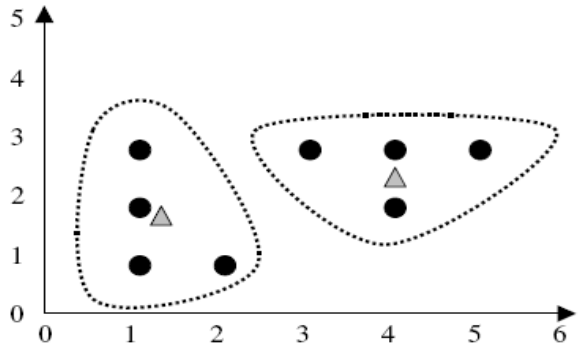


Figura 2.5 – **Grupos e contróides após o passo 3 da 2º iteração [Larose 2005]**

O objetivo do algoritmo *K-Means* é minimizar a distância dos pontos que estão dentro de um grupo. A função soma dos erros quadrados, que é dada pela Equação 2.2, é normalmente utilizada como um dos critérios de convergência dos grupos formados. Este critério tenta fazer com que os *k* grupos resultantes sejam tão compactos e tão separados quanto possível.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, m_i)^2 \qquad \text{Eq. 2.2}$$

Onde x é um ponto de dado do grupo C_i e m_i é o centróide do grupo C_i.

2.4.2 **Métodos Hierárquicos**

Segundo Metz [Metz 2006], nessa abordagem o resultado obtido não é constituído apenas de uma partição do conjunto de dados inicial, mas sim de uma hierarquia que descreve um particionamento diferente a cada nível analisado. Um agrupamento hierárquico agrupa os dados de modo que se dois exemplos são agrupados em algum momento, nas próximas iterações eles continuam fazendo parte do mesmo grupo, mesmo se forem agrupados em outros grupos mais gerais, caracterizando assim uma hierarquia de grupos, conforme a Figura 2.6. O agrupamento hieráquico pode ser visualizado como um dendograma, que é um diagrama tipo árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos, conforme ilustrado na Figura 2.7. Essa técnica permite analisar os grupos em diferentes níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos. Nessa abordagem, não é necessário assumir qualquer número particular de grupos, qualquer número desejado de grupos pode ser obtido cortando o dendograma no nível apropriado.

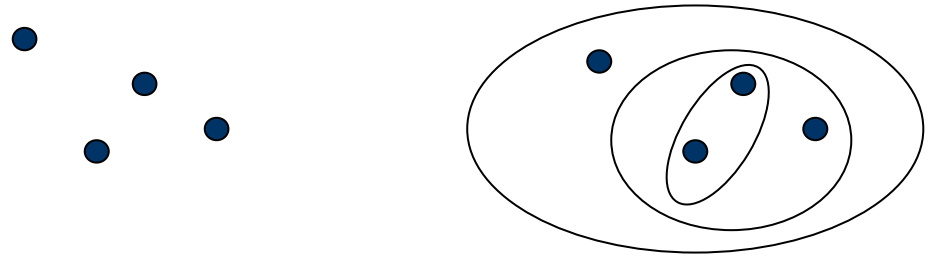


Figura 2.6 – **Agrupamento Hierárquico**

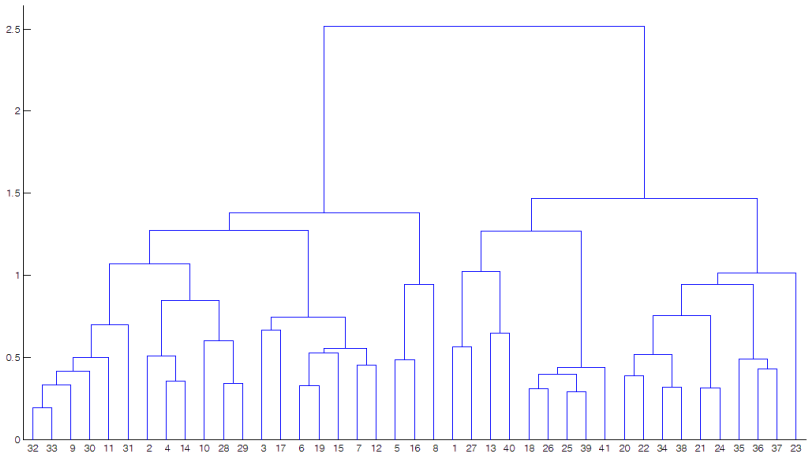


Figura 2.7 - **Dendograma**

Os algoritmos de *clustering* hierárquico utilizam duas estratégias na sua implementação:

1. Aglomerativa (*botton-up*) ;
2. Divisiva (*top-down*).

Na estratégia aglomerativa parte-se das folhas para a raiz, portanto inicialmente cada exemplo é considerado um grupo unitário. Em cada etapa, calcula-se a distância entre cada par de grupos. Essas distâncias são, geralmente, armazenadas em uma matriz de similaridade. Então, são escolhidos dois grupos com a distância mínima, juntando-os em seguida. Atualiza-se a matriz de similaridade. Esse processo continua até que todos os objetos estejam em um único grupo (o nível mais alto da hierarquia), ou até que uma condição de término ocorra (por exemplo, o número de grupos desejado tenha sido alcançado) [GP 2005].

Na estratégia divisiva parte-se da raiz para as folhas. Inverte-se o processo por começar com todos os objetos em um único grupo. Em cada etapa, um grupo é escolhido e dividido em dois grupos menores. Esse processo continua até que se tenha *n* grupos ou até que uma condição de término aconteça [GP 2005].

Segundo Metz [Metz 2006], na literatura, trabalhos relacionados ao *clustering* hierárquico, geralmente, referenciam a abordagem aglomerativa, não mostrando muito interesse pelos métodos divisivos. Mesmo os pacotes de *software* que implementam algoritmos de *clustering* dificilmente incluem algoritmos divisivos. O principal aspecto que influencia esse fato é a complexidade computacional desses métodos. Algoritmos aglomerativos são quadráticos em relação ao número de exemplos do conjunto de entrada, i.e., $O(N^2)$, mas ainda assim muitas vezes aplicáveis. A complexidade dos algoritmos divisivos, por outro lado, cresce exponencialmente em relação ao tamanho do conjunto de entrada, o que torna proibitiva sua aplicação sobre conjunto de dados com pouco mais que algumas centenas de exemplos.

Diferentes abordagens para definir a distância entre grupos distinguem os diferentes algoritmos. Os algoritmos frequentemente referenciados como clássicos na literatura de *clustering* hierárquico são: ***Single-Linkage, Complete-Linkage, Average-Linkage*** *e* ***Ward*** [Metz 2006]. Apresentaremos a seguir uma descrição dos algoritmos *Single-Linkage, Complete-Linkage e Ward*, utilizados no presente trabalho.

2.4.2.1 **Algoritmo *Single-Linkage***

É um dos algoritmos de *clustering* hierárquico aglomerativo mais simples. Esse método utiliza a técnica do vizinho mais próximo (*Nearest Neighbor Technique*), na qual a distância entre dois grupos é determinada pela distância do par de exemplos mais próximo, sendo cada exemplo

pertencente a um desses grupos, conforme mostra a Figura 2.8. A distância Euclidiana é usualmente utilizada para obter a matriz das distâncias dos elementos de dados a serem agrupados.

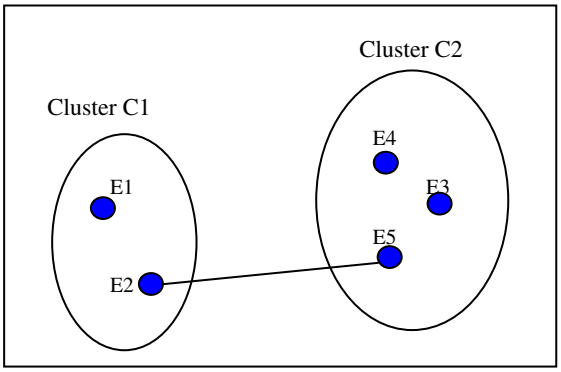


Figura 2.8 – **Algoritmo *Single-Linkage* [Metz 2006]**

Inicialmente cada objeto forma um grupo e a similaridade entre dois grupos é medida pela distância do par de dados mais próximos. O processo de juntar grupos se repete até que todos os objetos são, eventualmente, reunidos para formar um único grupo.

O algoritmo *Single-Linkage* tende a formar grupos longos se comparado com os grupos formados por outros métodos de agrupamento [Meyer 2002].

Segue abaixo o algoritmo *Single-Linkage*:

1. criar *n* grupos, cada grupo contendo um objeto;
2. calcular a matriz de distâncias (Euclidiana);
3. combinar os grupos com menor distância;
4. atualizar matriz de distâncias;
5. Repetir os passos 3 a 4 até que se tenha apenas um grupo que inclua todos os objetos .

2.4.2.2 **Algoritmo *Complete-Linkage***

Esse método utiliza a técnica conhecida como *Farthest Neighbor* ou vizinho mais distante. Ao contrário do algoritmo *Single-Linkage*, esse algoritmo determina a distância entre dois grupos de acordo com a maior distância entre um par de exemplos, sendo cada exemplo pertencente a um grupo distinto, conforme a Figura 2.9. Com isso, tem maior propensão a identificar grupos menos alongados [Metz 2006].

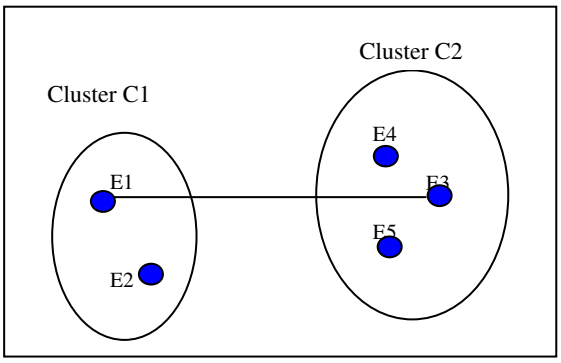


Figura 2.9 – **Algoritmo *Complete-Linkage* [Metz 2006]**

Os passos do algoritmo *Complete-Linkage* podem ser descritos, como segue:

1. criar *n* grupos, cada grupo contendo um objeto;
2. calcular a matriz de distâncias (Euclidiana);
3. combinar os grupos com maior distância;
4. atualizar matriz de distâncias;
5. Repetir os passos 3 a 4 até que se tenha apenas um grupo que inclua todos os objetos.

2.4.2.3 **Algoritmo *Ward***

O método de *Ward* é um método de agrupamento onde os grupos de dados são formados em etapas e são sistematicamente reduzidos (*n,n-1,n-2,...*), considerando a união de todos os *n(n-1)/2* possíveis pares e selecionando a união que tem um valor máximo para a função objetivo.

Segundo Ward [Ward 1963], em geral uma função objetivo pode ser qualquer relação funcional que o especialista selecione para refletir o que se deseja nos agrupamentos. A natureza do problema e o critério para o valor mínimo da função objetivo, com certeza, ditam a escolha e interpretação da função objetivo. Ela tem como objetivo orientar a escolha da melhor junção dos grupos. A função soma dos erros quadrados é geralmente utilizada para este fim.

O processo de agrupamento começa com *m* membros, cada membro formando um conjunto, apesar de conter apenas um elemento. O primeiro passo do agrupamento é escolher dois desses *m* subconjuntos os quais, quando unidos, reduzirá de um o número de subconjuntos, com menos prejuízo do valor ótimo da função objetivo (minimizar a perda de informação). Este procedimento pode continuar até o número de grupos desejado ser atingido ou o valor da função objetivo atingir um valor aceitável e se desejado, até todos os membros do conjunto original estarem em um grupo só. Os grupos resultantes são mutuamente exclusivos.

O algoritmo *Ward* pode ser apresentado em linhas gerais, como segue:

1. criar *m* grupos, cada grupo contendo um vetor componente da base de dados ;
2. considerar as *m* x *m* combinações e calcular o erro interno para cada combinação;
3. escolher a combinação que obteve o menor erro interno;
4. o número de grupos é reduzido de 1 (*m=m-1*) e o processamento retorna ao passo 2 até que se tenha apenas um grupo que inclua todos os objetos ou qualquer outro critério de parada seja atingido.

2.4.3 Métodos baseados em modelos

Os métodos baseados em modelos criam um modelo hipotético para cada grupo desejado e procuram ajustar os dados da melhor maneira ao modelo criado. Os algoritmos baseados neste método são capazes de descobrir os grupos por meio de ações de densidade que refletem a distribuição espacial dos objetos [Espenchitt 2008].

Os métodos baseados em modelos seguem três principais abordagens: a abordagem estatística, o aprendizado de máquina e a abordagem de redes neurais.

Na abordagem de redes neurais, encontramos os mapas auto-organizáveis, também conhecidos como mapas SOM – *Self Organizing Map* ou Redes Neurais de *Kohonen*, que serão descritas a seguir.

2.4.3.1 **Rede Neural de *Kohonen***

A rede neural de *Kohonen* é um tipo de rede neural artificial baseada em aprendizado competitivo e não supervisionado, capaz de mapear um conjunto de dados de entrada, em um conjunto finito de neurônios organizados numa grade regular de baixa dimensão, geralmente unidimensional ou bidimensional. Dimensões maiores são possíveis, mas geralmente não são usadas devido à dificuldade de visualização dessas configurações [Vesanto 2000]. A organização dos neurônios fornece um mapa topográfico dos dados de entrada no qual as localizações espaciais (coordenadas) dos neurônios na grade são indicativas das características estatísticas intrínsecas aos dados de entrada [Haykin 2004].

A rede neural de *Kohonen* é formada por apenas duas camadas: a de entrada e a de saída, conforme Figura 2.10. Cada neurônio está conectado a todas as entradas da rede e o mapeamento topológico produzido pela rede é baseado no aprendizado competitivo.

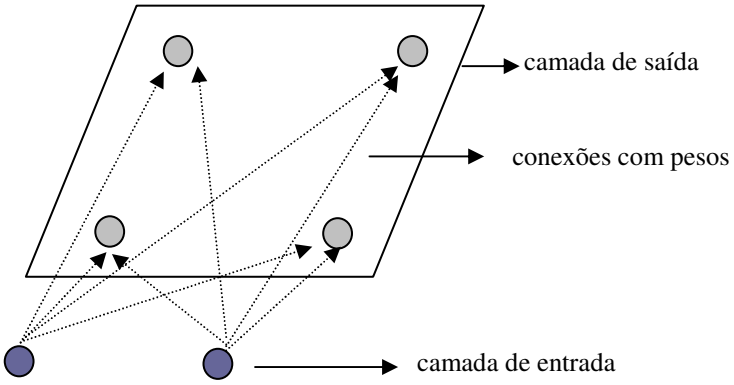
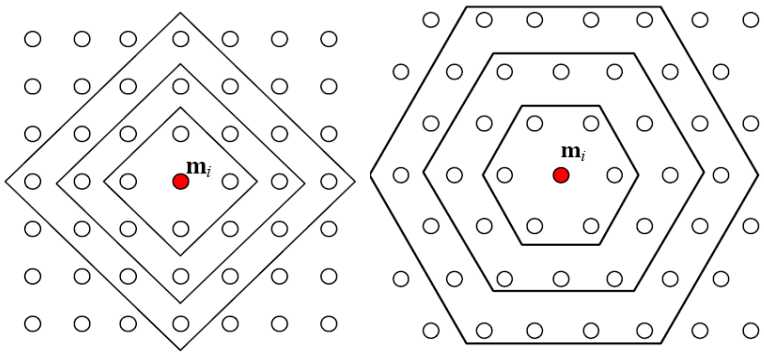


Figura 2.10 – **Rede Neural de *Kohonen***

No arranjo bidimensional, a grade apresenta uma topologia particular, que pode ser retangular, hexagonal e outras, conforme Figura 2.11.



(a) (b)

Figura 2.11 – **Diferentes configurações de arranjo para a rede de *Kohonen*. Em (a) vê-se uma topologia retangular e em (b) uma topologia hexagonal [Zuchini 2003].**

Treinamento da rede neural de *Kohonen*

Segundo Haykin [Haykin 2004], o algoritmo responsável pela formação do mapa auto-organizável começa primeiramente inicializando os pesos sinápticos de grade. Isto pode ser feito atribuindo-lhe valores pequenos tomados de um gerador de números aleatórios. Fazendo dessa forma, nenhuma organização prévia é imposta ao mapa de características. Uma vez que a grade tenha sido apropriadamente inicializada, há três processos essenciais envolvidos na formação do mapa auto-organizável, como resumido aqui:

1. Competição – Para cada padrão de entrada, os neurônios da grade calculam seus respectivos valores de uma função discriminante. Essa função discriminante fornece

a base para a competição entre os neurônios. O neurônio particular com o maior valor da função discriminante é declarado vencedor da competição.

2. Cooperação – O neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados, fornecendo assim a base para a cooperação entre os neurônios vizinhos.
3. Adaptação Sináptica – Este último mecanismo permite que os neurônios excitados aumentem seus valores individuais da função discriminante em relação ao padrão de entrada através de ajustes adequados aplicados a seus pesos sinápticos. Os ajustes feitos são tais que a resposta do neurônio vencedor à aplicação subseqüente de um padrão de entrada similar é melhorada. A Figura 2.12 ilustra essa etapa.

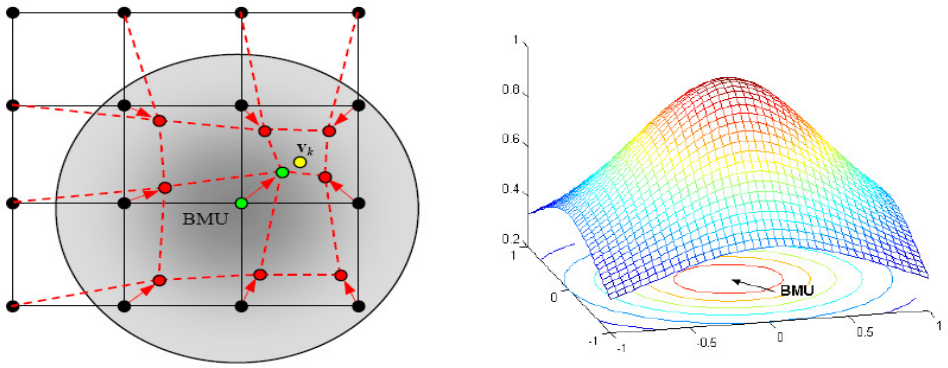


Figura 2.12 – **Ilustração da adaptação dos pesos de uma rede de *Kohonen* [Zuchini 2003].**

O algoritmo básico da rede de *Kohonen* pode ser resumido nos seguintes passos:

1. Inicialização – Escolher a topologia para a rede e definir o número de iterações.
   - o conjunto de pesos iniciais w_{ij} é gerado aleatoriamente com valores pequenos;
   - atribuir para o valor de vizinhança (σ) um valor suficientemente grande para incluir todas as unidades;
   - adotar valor para a taxa de aprendizagem (α) e para as constantes τ_1 e τ_2.

2. Apresentação dos exemplos de entrada
   - competição – nesta fase os exemplos de entrada são apresentados e calcula-se a resposta dos neurônios de saída. O neurônio que apresentar a menor distância euclidiana, calculada pela Equação 2.3, é o vencedor da competição.

$$d_{xw} = \sqrt{\sum_{j=1}^{n} \left(x_j - w_{ij}\right)^2} \qquad \text{Eq. 2.3}$$

3. Cálculo do valor da vizinhança topológica

   - cooperação – uma vez selecionado o neurônio vencedor calcula-se o valor da vizinhança topológica dada pela Equação 2.4:

$$h_i^v(n) = \exp\left[-\frac{(d_i^v)^2}{2\sigma^2(n)}\right] \qquad \text{Eq. 2.4}$$

   onde: d_i^v é a distância entre o neurônio vencedor 'v' e o neurônio vizinho excitado.

   σ(n) é o valor da largura da função de vizinhança (raio) na iteração n.

   Várias funções, listadas a seguir, satisfazem as exigências para a vizinhança topológica h_i^v [Francisco 2004]:

   a) $h_i^v(n) = (\sigma(n) - d_i^v)$ Eq. 2.5

   b) $h_i^v(n) = \exp\left[-\frac{(d_i^v)^2}{2\sigma^2(n)}\right]$ Eq. 2.6

   c) $h_i^v(n) = \exp\left[-\frac{(d_i^v)^2}{2\sigma^2(n)}\right](\sigma(n) - d_i^v)$ Eq. 2.7

   d) $h_i^v(n) = \max\{0, 1 - (\sigma(n) - d_i^v)^2\}$ Eq. 2.8

   Sendo que a mais utilizada é a função de vizinhança gaussiana dada pela Equação 2.6.

4. Ajuste dos pesos (adaptação sináptica)

   - os pesos do neurônio vencedor e de todas as unidades dentro da região de vizinhança da unidade vencedora têm seus pesos ajustados na direção do vetor de exemplos de entrada.

$$w_{ij}(novo) = w_{ij}(antigo) + \alpha h_i^v(x_j - w_{ij}(antigo)) \qquad \text{Eq. 2.9}$$

5. Cálculo do novo valor para σ(n) (largura da função) e para α(n) (taxa de aprendizagem).

$$\sigma(n) = \sigma(0)\exp\left[-\frac{n}{\tau_1}\right] \qquad \text{Eq. 2.10}$$

onde: σ(n) é o valor da largura da função na iteração n, σ(0) é um valor inicial para a largura da função e τ_1 é uma constante de tempo que diminui a medida que o número de iterações aumenta.

$$\alpha(n) = \alpha(0)\exp\left[-\frac{n}{\tau_2}\right] \qquad \text{Eq. 2.11}$$

onde: α(n) é a taxa de aprendizagem para a iteração n, α(0) é um valor inicial para a taxa de aprendizagem e τ_2 é uma constante de tempo que depende do número de iterações.

6. Testar a condição de parada.

7. Incrementar o número da iteração. Retornar ao passo 2, repetir até a condição de parada seja satisfeita.

Segundo Zuchini [Zuchini 2003], os parâmetros que regulam a rede de *Kohonen* são muitos, mas podem ser agrupados basicamente em dois conjuntos: aqueles que definem a estrutura do mapa (suas dimensões, vizinhança e formato do arranjo, raio e tipo da função de vizinhança) e aqueles que controlam o treinamento propriamente dito (se incremental, com a respectiva taxa de aprendizado α(t); em lote, com a função de decrescimento do raio de vizinhança σ(t); o número de épocas de treinamento). Podem também ser considerados parâmetros adicionais como o treinamento em duas fases e a normalização de dados de entrada, comum em atividades de mineração de dados.

Interpretação dos Mapas produzidos pela rede SOM

Para avaliar os possíveis agrupamentos, o método mais comumente empregado é a matriz de distâncias unificadas, conhecida como *matriz-U*, que permite realizar a discriminação dos agrupamentos, a partir de uma medida do grau de similaridade entre os pesos de neurônios adjacentes na rede.

Tomando a *matriz-U* como uma superfície de nível, pode-se avaliar visualmente a existência de vales (que surgem onde os vetores de pesos dos neurônios são mais próximos entre si) separados por elevações (onde os vetores de pesos dos neurônios encontram-se mais distantes), conforme ilustra a Figura 2.13. Um vale é associado com a ocorrência de um agrupamento e quanto mais alta

uma elevação separando dois vales, tanto mais distintos são estes agrupamentos no espaço de dados [Zuchini 2003].

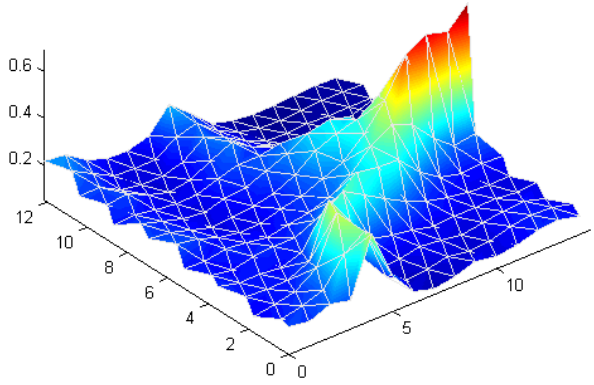


Figura 2.13 – ***Matriz-U* em 3D [Zuchini 2003]**

Outra forma de representar a matriz-U graficamente é apresentada na Figura 2.14. A matriz-U apresenta através das unidades coloridas do mapa, as distâncias entre os agrupamentos. O seu tamanho é quase duas vezes maior que o mapa original, pois existem hexágonos adicionais entre todos os pares de unidades vizinhas. As cores variam de acordo com uma escala de distâncias, como pode ser observado na Figura 2.14, indo do azul escuro ao vermelho, onde a cor azul representa os agrupamentos e as cores mais claras até o vermelho representam a separação dos agrupamentos.

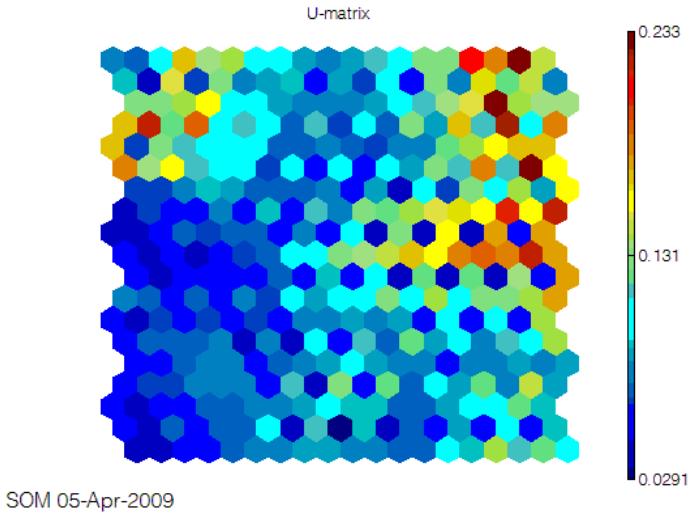


Figura 2.14 – ***Matriz-U***

Outras formas de visualização dos agrupamentos são apresentadas a seguir. A Figuras 2.15 apresenta o mapa com o rotulamento dos objetos agrupados e a Figura 2.16 apresenta o mapa com a subdivisão dos objetos em grupos, identificados cada um por uma cor diferente.

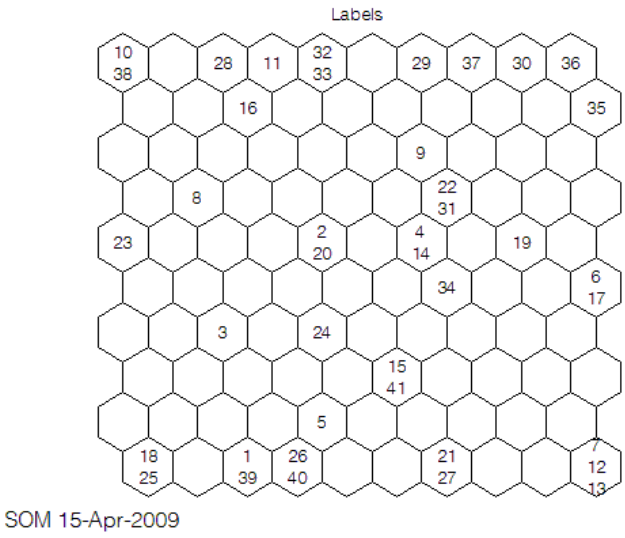


Figura 2.15 – **Mapa com o rotulamento dos objetos agrupados**

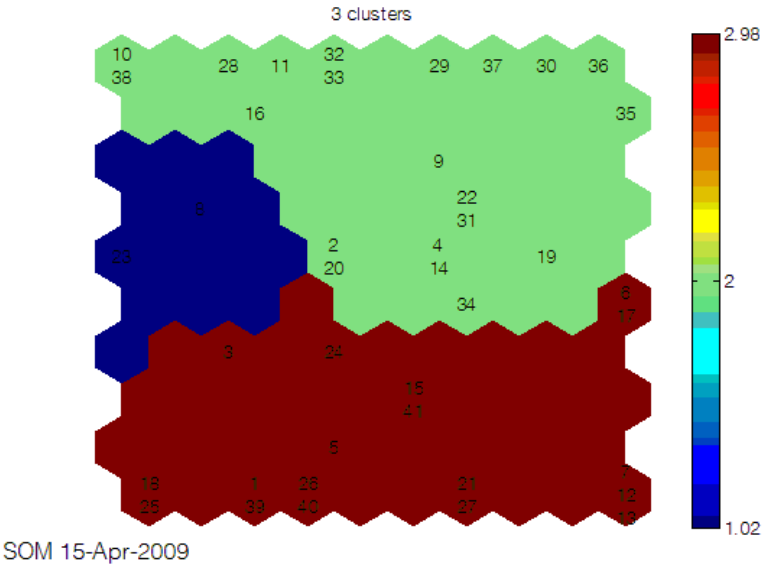


Figura 2.16 – **Mapa com a subdivisão dos objetos em grupos**

Avaliação dos resultados

Segundo Zuchini [Zuchini 2003], para se avaliar a qualidade de um mapa, normalmente são utilizadas duas métricas: a primeira é o *Erro Médio de Quantização (Quantization Error – QE)*, que corresponde à média das distâncias entre cada vetor de dados v_n e o correspondente vetor de pesos m_c do neurônio vencedor BMU. O índice QE é calculado pela Equação :

$$QE = \frac{1}{N}\sum_{n=1}^{N}\left\|m_c - v_n\right\| \qquad \text{Eq. 2.12}$$

A segunda medida é o *Erro Topográfico(Topographic Error – TE)*, que quantifica a capacidade do mapa de representar a topologia dos dados de entrada. Esta medida apresenta a proporção de todos os vetores de dados, para os quais, o primeiro e o segundo neurônio vencedor não são unidades adjacentes. Para cada objeto v_n são calculados seu BMU m_c e o segundo BMU m_d, onde $u(v_n) = 1$ caso m_c e m_d não sejam adjacentes.

$$TE = \frac{1}{N}\sum_{n=1}^{N}u(v_n) \qquad \text{Eq. 2.13}$$

2.5 **CONCLUSÃO**

Neste capítulo foi apresentado o conceito do processo de Descoberta do Conhecimento em Base de Dados. Dentro dessa conceituação foram apresentadas: as etapas do processo de KDD, dando ênfase a etapa de Mineração de Dados, e a tarefa de Clusterização, tarefa de maior interesse para o desenvolvimento deste trabalho. Foram descritos também os algoritmos clássicos utilizados para execução da tarefa de Clusterização. No próximo Capítulo será apresentada uma proposta de metodologia que utiliza a Clusterização para identificar de regiões hidrologicamente homogêneas no Estado da Paraíba.

CAPÍTULO III

METODOLOGIA

A metodologia do trabalho é ilustrada na Figura 3.1. Os processos apresentados estão contidos nas principais etapas do processo de descoberta do conhecimento: pré-processamento, mineração de dados e pós-processamento. A primeira etapa compreende desde a correção de dados (redundância, inconsistência, ausência de valores) até o ajuste da formatação dos mesmos para o algoritmo de mineração de dados a ser utilizado. Além disso, métodos de seleção de atributos são empregados com o objetivo de selecionar apenas os atributos que melhor descrevem e discriminam o conjunto de dados e suas estruturas latentes, o que conseqüentemente reduz a dimensionalidade dos dados, melhora a eficiência dos algoritmos em relação ao tempo de execução, e, em alguns casos, pode melhorar os resultados obtidos [Metz, 2006].

Na segunda etapa do trabalho serão executados e avaliados alguns algoritmos de clusterização com o objetivo de identificar qual deles adequa-se melhor aos estudos de regionalização hidrológica. A terceira etapa determina o grau de significância dos resultados obtidos pelo algoritmo de clusterização.

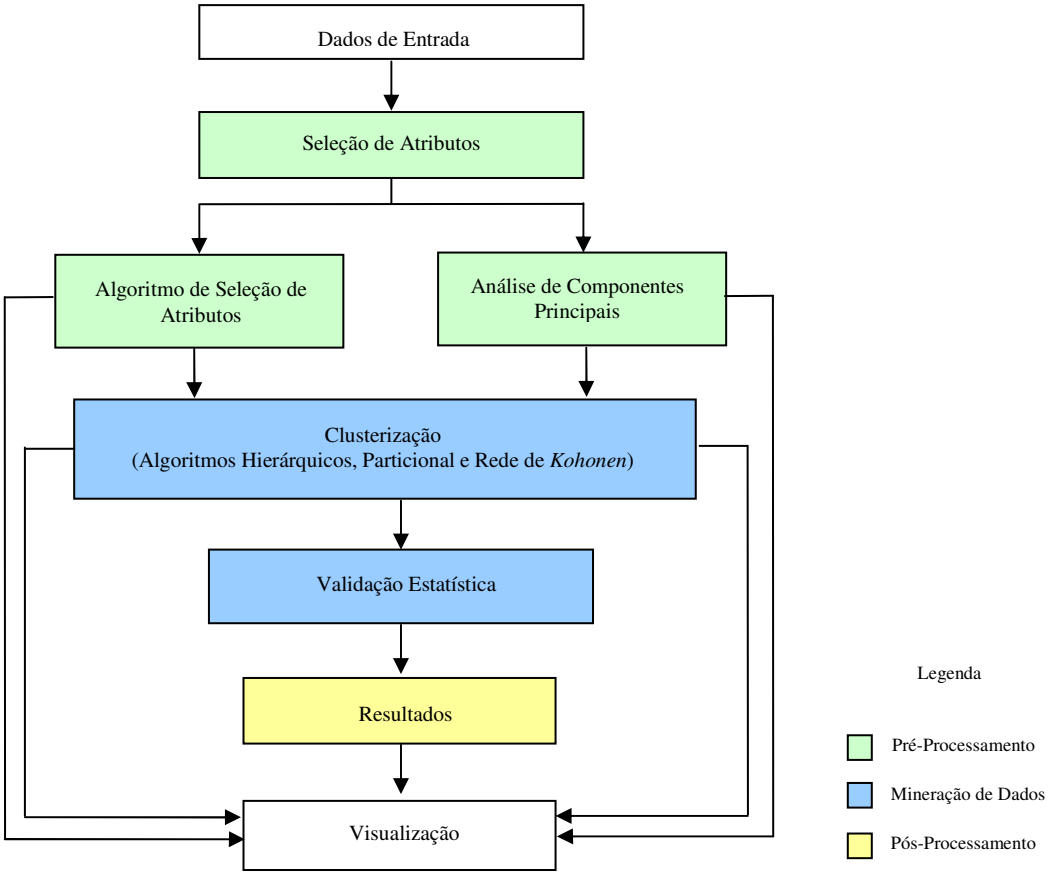


Figura 3.1 – **Esquema da Metodologia Proposta**

Para a execução dos experimentos foi utilizado o *software MATLAB* versão 7.0. O *MATLAB* (*Matriz Laboratory*) é uma linguagem de programação e um ambiente computacional interativo de alto desempenho para computação técnica. Este integra computação, visualização e programação num ambiente amigável, onde os problemas e soluções são formulados em notação matemática familiar. O *MATLAB* possui uma série de soluções para aplicações específicas, designadas *Toolboxes*. Áreas para as quais estão disponíveis *Toolboxes* incluem: processamento de sinais e de imagens, aquisição de dados, redes neurais, *clustering*, lógica difusa, e muitas outras [Celeste 2005].

Na validação estatística foi utilizado o programa Machaon CVE, (*Clusteringand Validation Environment*), que avalia a qualidade dos agrupamentos obtidos através de diferentes índices estatísticos [Bolshakova 2009].

3.1 DESCRIÇÃO DO CONJUNTO DE DADOS

As características físicas e climatológicas de uma bacia são elementos importantes para descrever o seu comportamento hidrológico [Diniz 2008]. As características que serão utilizadas no presente trabalho foram obtidas através de mapa digitalizado e atualizado da hidrografia do Estado da Paraíba e do Modelo Digital de Elevação (MDE), gerado para todo o Estado através do software *ArcView* GIS do ESRI (*Environmental Systems Research Institute*). O Modelo Digital de Elevação (MDE) foi obtido a partir de pontos cotados, georeferenciados e com os valores de altitudes associados, oriundos da SRTM (Missão Topografia de Rastreamento *Shuttle* - projeto internacional gerenciado pela agência nacional de Inteligência Geo-Espacial (NGA) e pela NASA, iniciado em 2000), cobrindo todo o estado com uma malha de 90 m x 90 m [Rabus et al. 2003]. Nas Tabelas 3.1 a 3.7 estão as 32 (trinta e duas) características obtidas, agrupadas por categoria, que identificam de forma completa cada bacia hidrográfica e influenciam grandemente no regime hidrológico da mesma.

A planilha com os valores das 32 características para cada bacia hidrográfica foi cedida pela AESA – Agência Executiva de Gestão das Águas do Estado da Paraíba, conforme Anexo D.

A Figura 3.2 apresenta o mapa do Estado da Paraíba com a localização das 41 bacias hidrográficas, selecionadas para o estudo.

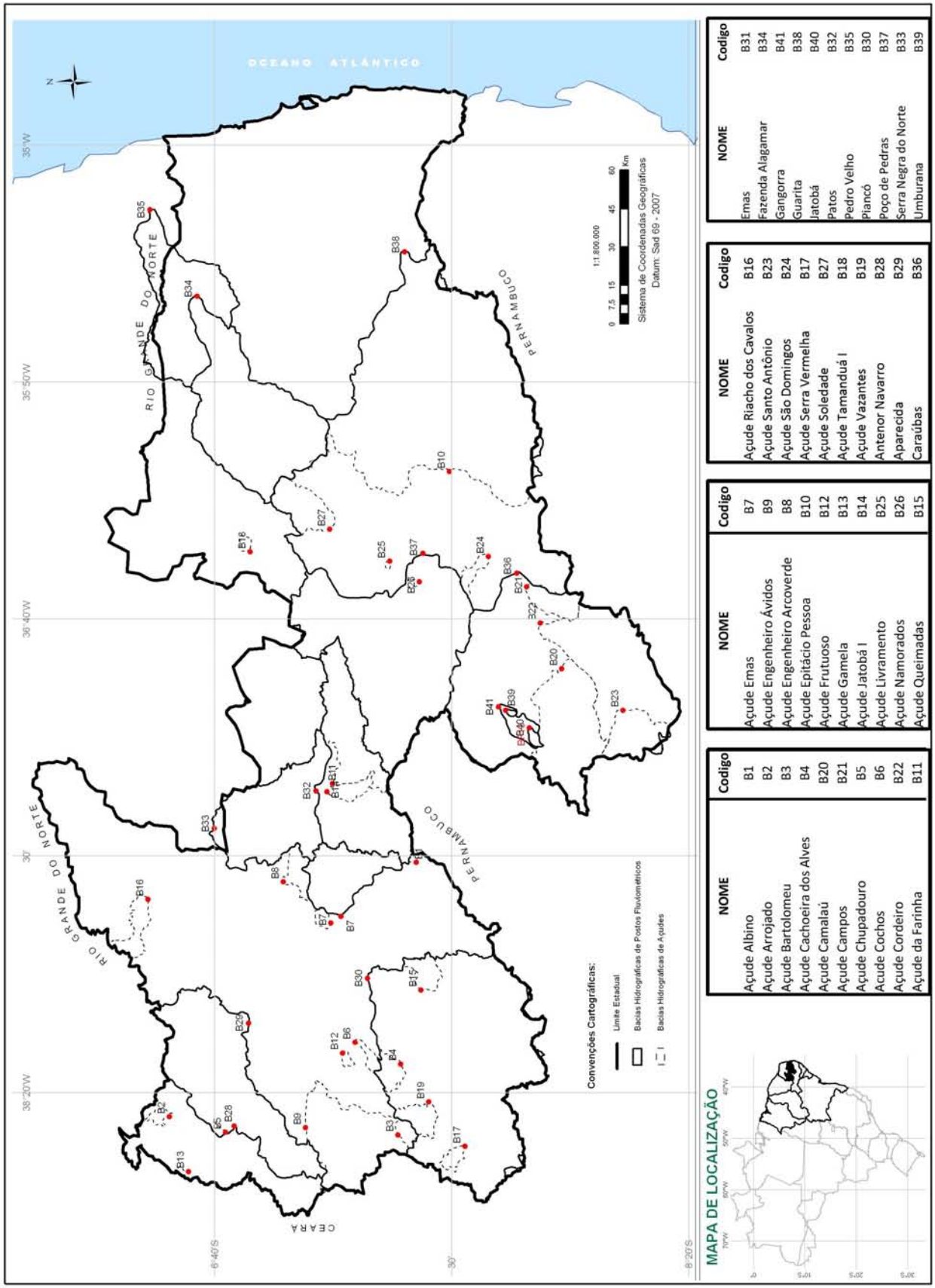


Figura 3.2 – **Localização das bacias hidrográficas selecionadas**

Tabela 3.1: **Características das medidas lineares da bacia**

Parâmetro	Descrição	Unidade
Área (A)	É a área plana de uma bacia (projeção horizontal) inclusa entre seus divisores topográficos.	Km^2
Perímetro (P_r)	Perímetro da bacia hidrográfica	Km
Linha de fundo (L)	Distância medida ao longo do curso de água principal desde a seção de referência até o divisor da bacia.	Km
Comprimento do curso d'água (L_t)	Distância medida ao longo do curso de água principal desde a seção de referência até o final do curso d'água.	Km
Comprimento da rede de drenagem (L_d)	Somatório dos comprimentos de todos os talvegues da bacia.	Km
Largura média (L_m)	É obtida quando se divide a área pelo comprimento da bacia. Quanto mais L_m se aproxima de L, tanto mais a bacia é propícia a ocorrência de cheias instantâneas	Km

Tabela 3.2: **Características da forma da bacia**

Parâmetro	Descrição	Unidade
Índice de compacidade (K_c)	É a relação entre o perímetro da bacia e a circunferência de um círculo de área igual à da bacia. A tendência à enchente de uma bacia será tanto maior quanto mais próximo da unidade for este coeficiente, pois maior será a possibilidade de toda a área estar contribuindo de uma só vez. Um coeficiente igual à unidade corresponderia a uma bacia circular.	Adm
Índice de circularidade (K_e)	Relação entre a área de um círculo que tem o perímetro igual ao da bacia e a área da bacia. A circularidade tende para a unidade à medida que a bacia se aproxima da forma circular e diminui à medida que a forma torna alongada.	Adm
Fator de forma (K_f)	Relação entre a largura média e a linha de fundo da bacia. O fator de forma também dá alguma indicação sobre a tendência a inundações.	Adm

Tabela 3.3: **Características da rede de drenagem**

Parâmetro	Descrição	Unidade
Ordem dos cursos de água (O_r)	Número de segmentos de cursos de água.	Adm
Índice de bifurcação (R_b)	Relação entre o número de talvegues de dada ordem, pelo número de talvegues de ordem imediatamente superior.	Adm
Índice dos comprimentos (R_L)	Média geométrica das relações entre os cumprimentos médios dos talvegues de duas classes consecutivas.	Adm
Índice das áreas (R_a)	Média geométrica das relações entre as áreas médias dos talvegues de duas classes consecutivas.	Adm
Densidade de drenagem (D_d)	Quociente entre o comprimento total da rede de drenagem e a área total da bacia hidrográfica. Densidades de drenagens baixas são observadas normalmente em solos mais resistentes à erosão ou muito permeável e onde o relevo é suave. Valores altos podem ser esperados em bacias cujos solos são facilmente erodidos ou relativamente impermeáveis, as declividades são altas e as coberturas de vegetação são ralas.	Km/Km^2
Coeficiente de Torrencialidade (C_t)	Relação entre o número de cursos de água e a área da bacia.	n/km^2
Índice de rugosidade (IR)	Produto da diferença máxima de altitude dentro de uma bacia pela sua densidade de drenagem.	Adm
Extensão média do escoamento superficial (L_e)	Distância média que a água de chuva teria que escoar sobre os terrenos da bacia (em linha reta) do ponto onde ocorreu sua queda até o curso d'água mais próximo.	Km
Sinuosidade do curso d´água (SIN)	Relação entre o comprimento do rio principal e a sua projeção horizontal, que seria o seu comprimento reto medido a partir do fundo do seu leito.	Adm

Tabela 3.4: **Características do relevo da bacia**

Parâmetro	Descrição	Unidade
Declividade máxima (I_{max}) do rio	Indica a inclinação máxima do rio principal da bacia.	m/m
Elevação média da bacia (C_{med})	Calculada em função da área entre as curvas de nível, altitude média entre duas curvas de nível consecutivas e a área total da bacia.	m/m
Lado Maior (L_r)	Lado maior do retângulo equivalente. O retângulo equivalente consiste de um retângulo de mesma área e mesmo perímetro da bacia. O objetivo do retângulo equivalente é de comparar melhor a influência das características físicas da bacia sobre o escoamento.	Km
Lado Menor (l_r)	Lado menor do retângulo equivalente.	Km
Índice de declividade média da bacia (I_p)	A declividade relaciona-se com a velocidade em que se dá o escoamento superficial, afetando, portanto, o tempo que leva a água da chuva para concentrar-se nos leitos fluviais que constituem a rede de drenagem das bacias.	Adm
Índice de declividade global (IG)	É a diferença entre as altitudes correspondentes às percentagens de 5% e 95% dividida pelo comprimento do retângulo equivalente.	m/m
Desnível específico (DS)	É função da forma do contorno da bacia. Caracteriza o relevo da bacia (muito suave,suave,ondulado,forte, muito forte)	m/m

Tabela 3.5: **Características da capacidade de escoamento da bacia hidrográfica**

Parâmetro	Descrição	Unidade
L_{600}	Lâmina Média Anual escoada em uma bacia hidrográfica localizada na zona climática Sertão com precipitação anual média de 600 mm.	mm
AE/A	Área do Espelho (AE), percentagem da área da bacia hidrográfica coberta por espelho de água.	Adm

Tabela 3.6: **Características climatológicas da bacia hidrográfica**

Parâmetro	Descrição	Unidade
Precipitação média da bacia (P)	Chuva média precipitada na bacia, determinada por interpolação entre vários postos da região.	mm
Evapo-transpiração média da bacia	Soma entre a evaporação das superfícies e a transpiração da vegetação.	mm

Tabela 3.7: **Características dos solos da bacia hidrográfica**

Parâmetro	Descrição	Unidade
Percentual do solo tipo 1 (SOLO1)	Solos com alta capacidade de escoamento (solos rasos, rochosos, baixa permeabilidade)	Adm
Percentual do solo tipo 1 (SOLO2)	Solos com média capacidade de escoamento (solos mistos)	Adm
Percentual do solo tipo 1 (SOLO3)	Solos com baixa capacidade de escoamento (solos arenosos, profundos, permeabilidade alta)	Adm

3.2 ETAPAS DO PROCESSO DE DESCOBERTA DO CONHECIMENTO

3.2.1 PRÉ-PROCESSAMENTO

Nesta etapa torna-se necessária a aplicação de métodos para tratamento, limpeza e redução dos dados. A seguir, serão apresentados os dois métodos que foram executados com a finalidade de preparar os dados a serem submetidos aos algoritmos de mineração de dados.

3.2.1.1 Normalização dos atributos

Os atributos que foram coletados apresentavam escalas de valores diferentes. Essa diferença de escala pode influenciar de forma tendenciosa os resultados dos algoritmos de mineração de dados. Desta forma, foi aplicado o método de normalização linear que consiste em considerar os valores, mínimo e máximo de cada atributo no ajuste da escala. Os valores dos atributos foram mapeados em um intervalo fechado de 0 até 1, definidos pela Equação 3.1 .

$$A' = (A - Min)/(Max - Min) \qquad \text{Eq. 3.1}$$

onde:

A' = valor normalizado

A = valor do atributo a ser normalizado

Min = valor mínimo do atributo a ser normalizado

Max = valor máximo do atributo a ser normalizado

3.2.1.2 Seleção dos atributos

Com o objetivo de selecionar apenas os atributos que melhor descrevem e discriminam o conjunto de dados e suas estruturas latentes, utilizamos os seguintes métodos:

1. Algoritmo de seleção de atributos não supervisionado
    - Necessidade de dispor de uma ferramenta alternativa capaz de melhorar os resultados obtidos através da análise de componentes principais;
    - Permitir o uso dos valores reais das características físicas e climatológicas e não de uma combinação matemática das mesmas.

2. Análise de Componentes Principais

- Largamente utilizado em aplicações de clusterização apresentadas em trabalhos técnicos e científicos publicados na bibliografia especializada [Demirel 2007] [Menezes 2007] [Júnior 2006] [Llanillo 2006];

3.2.1.2.1 **Algoritmo de seleção de atributos não supervisionado**

O algoritmo escolhido para fazer a seleção dos atributos foi proposto por Mitra et al. [Mitra 2002]. Ele envolve dois passos, a saber, o particionamento do conjunto original de atributos em alguns subconjuntos distintos ou *clusters*, e a seleção de um atributo representativo de cada *cluster* que fará parte do conjunto reduzido de atributos. O particionamento dos atributos é feito baseado no princípio *k*-NN (*k nearest-neighbors*) usando como medida de similaridade, o índice de compressão máxima da informação (λ_2).

O índice de compressão máxima da informação é definido como sendo o menor auto-valor da matriz de covariância das variáveis *x* e *y*, e é obtido pela Equação 3.2.

$$2\lambda_2(x,y) = \mathrm{var}(x) + \mathrm{var}(y) - \sqrt{(\mathrm{var}(x)+\mathrm{var}(y))^2 - 4\,\mathrm{var}(x)\,\mathrm{var}(y)\left(1-\rho(x,y)^2\right)} \qquad \text{Eq. 3.2}$$

Onde var( ) é a variância do atributo e ρ é a covariância entre dois atributos.
O valor de λ_2 é zero quando os atributos são linearmente dependentes. À medida que a dependência entre os atributos diminui o valor de λ_2 aumenta. λ_2 não é nada mais que o auto-valor na direção normal da componente principal do par de atributos (*x*,*y*).

Segundo Mitra et al. [Mitra 2002], primeiramente são encontrados os *k* atributos mais próximos para cada atributo. Entre eles, é selecionado o atributo que tem o subconjunto mais compacto (determinado por sua distância em relação ao vizinho mais distante) e seus *k* vizinhos são descartados. O processo é repetido para os atributos restantes até todos serem selecionados ou descartados. Na primeira iteração, o valor para o erro (ϵ) é setado igual a distância do atributo selecionado ao seu vizinho mais distante. Nas iterações subseqüentes, é checado se o valor do índice de compressão máxima (λ_2), correspondente ao subconjunto de atributos, é maior que ϵ. Se sim, o valor de *k* é decrementado. Mitra et al [Mitra 2002] descreve o algoritmo nos seguintes passos:

Passo 1: Escolher um valor inicial de $k \leq D - 1$. Inicializar o conjunto reduzido de atributos R com o conjunto original de atributos $R \leftarrow O$.

O – Conjunto original de atributos; D – tamanho do conjunto original de atributos; r_i^k - representa a dissimilaridade entre o atributo F_i e seus *k* atributos mais próximos em R.

Passo 2: Para cada atributo $F_i \in R$, calcular r_i^k .

Passo 3: Encontrar o atributo F_i para o qual r_i^k é mínimo. Reter este atributo em R e discartar os *k* atributos mais próximos de $F_{i'}$.

(Nota: $F_{i'}$ representa o atributo para qual removendo os *k* vizinhos mais próximos irá causar o menor erro entre todos os atributos em R). Fazer $\epsilon = r_{i'}^k$.

Passo 4: Se *k* > cardinalidade (R) – 1: *k* = cardinalidade (R) – 1.

Passo 5: Se *k* = 1: Ir para o passo 8.

Passo 6: Enquanto $r_{i'}^k > \epsilon$ faça:

(a) $k = k - 1$.

$r_{i'}^k = \inf_{Fi \in R} r_i^k$.

(b) Se *k* = 1: Ir para o passo 8.

Fim.

Passo 7: Ir para o passo 2.

Passo 8: Retorna o conjunto reduzido de atributos.

3.2.1.2.2 **Análise de Componentes Principais**

Dado a elevada quantidade de características e o fato de várias delas terem sido derivadas entre si, a seleção das mais importantes visa descartar aquelas não relevantes e/ou redundantes uma vez que estas podem reduzir a precisão e a compreensibilidade das hipóteses induzidas pelos algoritmos de aprendizado não supervisionado a serem posteriormente aplicados.

Em geral os objetivos da análise de componentes principais são a redução da dimensionalidade e a interpretação do problema. Os dados das 32 características adotadas neste trabalho em cada uma das 41 bacias são em algum grau correlacionados entre si, indicando que alguma informação contida em uma dada característica está também contida em alguma outra das 31 características.

A análise de componentes principais de um conjunto de dados X pode ser entendida de duas maneiras:

a) algebricamente, consistindo em encontrar outras variáveis (componentes principais, C) que são combinações lineares dos dados X ou da matriz de correlações R, cujas principais propriedades são apresentar menor dimensionalidade e serem não correlacionadas, o que não acontece com as variáveis originais;

b) geometricamente, os componentes principais representam uma seleção de novo sistema de eixos coordenados, obtidos pela rotação do sistema original.

As operações envolvidas na análise de componentes principais são feitas sobre a matriz de correlação ou a matriz de variância e covariâncias, oriundos do conjunto de dados iniciais, basicamente determinando-se seus autovalores e autovetores. Quando esses dados são apresentados em unidades diferentes deve-se trabalhar com a matriz de correlações, ou seja, com os dados padronizados, como é o caso das características das bacias. Em notação matricial, tem-se:

$$\mathbf{C} = \mathbf{X} \cdot \mathbf{A}$$

Onde: **C** é a matriz dos componentes principais;

X é a matriz dos dados;

A é a matriz dos coeficientes que definem a transformação linear.

A interpretação dos resultados obtidos nesse tipo de análise é geralmente feita sobre o quanto poucos componentes principais podem explicar a variação total dos dados e, sobre a matriz de coeficientes, cuja magnitude mede a importância da variável correspondente na determinação do componente principal. Existem algumas regras genéricas para decidir quantos componentes principais devem ser escolhidos para estudos posteriores:

a) Selecionar apenas os componentes principais que juntos explicam uma percentagem arbitrária da variância total de 70%, 80% ou 90%.
b) Para componentes principais sobre a matriz de correlações (dados padronizados) selecionar aqueles cujos autovalores forem maior que 1.
c) Traçar um gráfico com os autovalores ordenados ou a percentagem de variância explicada por cada componente principal e escolher o número correspondente a ordem onde a variação do segmento de reta no gráfico passa a ser pequena.

No processo de análise das componentes principais, a que foram submetidos os dados das 41 bacias hidrográficas, foi realizado o teste KMO (*Kaiser-Meyer-Olkin Measure of Sampling Adequacy*) que examina o ajuste de dados, tomando as variáveis simultaneamente, e provê uma informação sintética sobre os dados. É gerada uma matriz antiimagem de correlação, que é uma matriz das correlações parciais (correlação de uma variável contra outra, controlados os efeitos de todas as outras consideradas no modelo). Nessa matriz, a diagonal mede a adequação amostral para cada variável e o julgamento faz-se pelos mesmos valores críticos utilizados para o teste KMO [Faria 2006].

O valor do KMO próximo a 1 indica a existência de correlações parciais muito pequenas e consequentemente uma perfeita adequação dos dados para a Análise Fatorial. Os valores críticos para o KMO são os seguintes:

1) Valores na casa dos 0,90 : adequação ótima dos dados à análise fatorial.
2) Valores na casa dos 0,80: adequação boa dos dados à análise fatorial.
3) Valores na casa dos 0,70 : adequação razoável dos dados à análise fatorial.
4) Valores na casa dos 0,60 : adequação medíocre dos dados à análise fatorial.
5) Valores na casa dos 0,50 ou menores : adequação imprópria dos dados à análise fatorial.

Ao término desta análise, após rotação nos pesos originais pelo método Varimax, foram obtidos a matriz de pesos e os respectivos escores dos componentes. O método Varimax maximiza a variância dos quadrados dos pesos em cada dimensão nova, espalhando os quadrados dos pesos para os extremos de sua faixa de domínio.

Com os resultados obtidos na etapa de seleção de atributos e com conjunto completo de atributos das bacias hidrográficas foram compostos quatro cenários para a execução dos algoritmos de mineração de dados. Os resultados dos métodos acima citados e a composição dos quatro cenários estão descritos no Capítulo IV.

3.2.2 **MINERAÇÃO DE DADOS**

3.2.2.1 **Aplicação dos algoritmos**

3.2.2.1.1 **Algoritmos Hierárquicos**

A Figura 3.3 mostra o fluxo de execução dos algoritmos *Single-Linkage*, *Complete-Linkage* e Ward. Para o cálculo da matriz de similaridade foi utilizada a medida de distância Euclidiana.

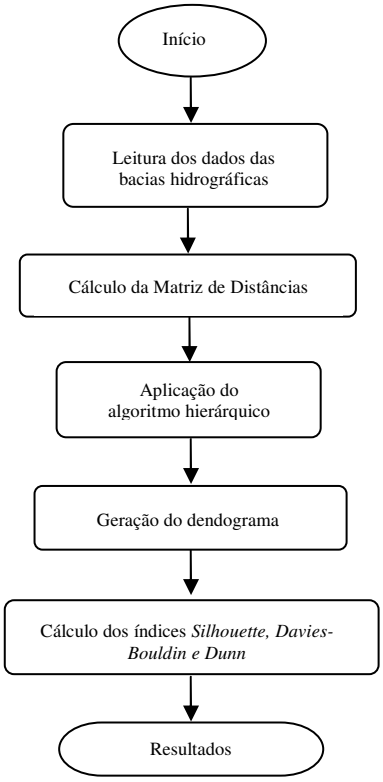


Figura 3.3 – **Fluxo de execução dos algoritmos hierárquicos**

A partir dos dendogramas gerados foram escolhidas quatro configurações para serem analisadas, a divisão das bacias hidrográficas em 2, 3, 4 e 6 grupos. Com as informações obtidas de cada configuração foram calculados os índices de validação.

3.2.2.1.2 **Algoritmo Particional**

A Figura 3.4 mostra o fluxo de execução do algoritmo *K-Means*. Para a execução do algoritmo *K-Means* também foi utilizada como medida de similaridade a distância Euclidiana com a escolha do centróide inicial feita de forma aleatória. Foram realizadas 10 (dez) iterações para o número de grupos (*k*) igual a 2,3,4 e 6. Ao final de cada iteração é fornecida, para cada grupo, a soma das distâncias de cada objeto do grupo ao seu centróide e o total da soma dessas distâncias. Com o agrupamento gerado por cada iteração foram calculados os índices de validação e obtidas a média e o desvio padrão dos valores de cada índice.

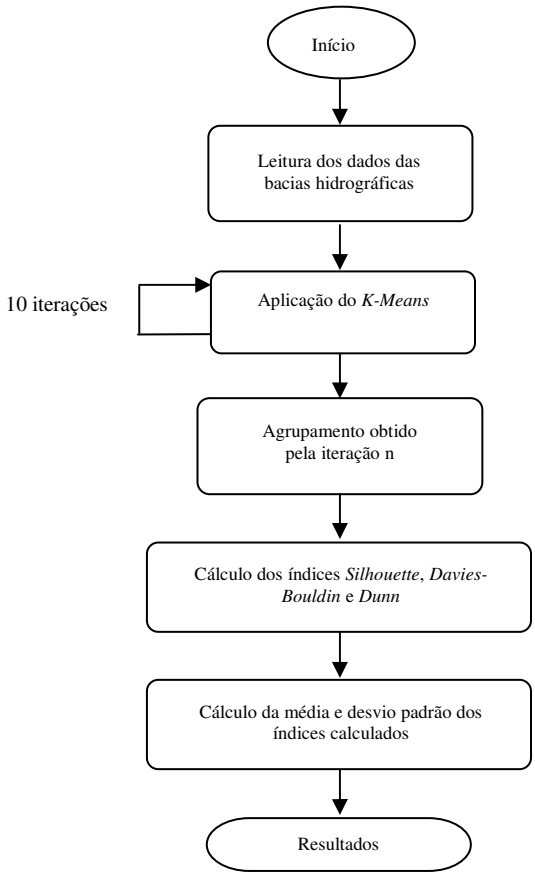


Figura 3.4 – **Fluxo de execução do algoritmo *K-Means***

3.2.2.1.3 **Rede Neural de *Kohonen***

Os parâmetros que regulam uma rede neural de *Kohonen* são muitos, portanto antes de aplicar o algoritmo de treinamento é importante a escolha da configuração da rede a ser utilizada. Neste trabalho a escolha da configuração se deu em duas etapas, a primeira para escolher a dimensão da rede e a segunda, para definir a melhor configuração com relação aos parâmetros de topologia, função de vizinhança e número de iterações das fases de ordenação e convergência do processo adaptativo dos pesos sinápticos da rede. Os resultados destas duas etapas estão detalhados no Capítulo IV. A Figura 3.5 apresenta o fluxo utilizado para a execução da rede neural de *Kohonen*.

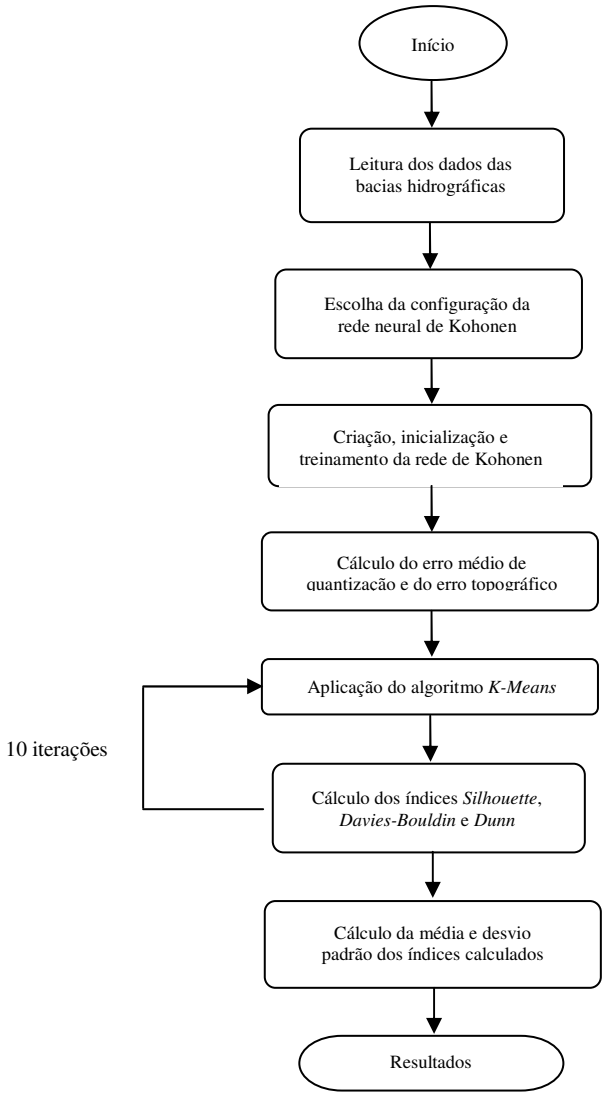


Figura 3.5 – **Fluxo de execução da rede neural de *Kohonen***

Após a escolha da configuração da rede, a mesma foi criada, inicializada e treinada. Com resultado do treinamento aplicou-se o algoritmo *K-Means* para obter os agrupamentos para cada valor de *k*. Foram realizadas 10 iterações para cada número de grupos (*k*) igual a 2, 3, 4 e 6. Com o agrupamento gerado por cada iteração foram calculados os índices de validação e obtidas a média e o desvio padrão dos valores de cada índice.

3.2.2.2 **Escolha dos Parâmetros**

Uma dificuldade freqüente nesta etapa é a escolha da parametrização adequada a um algoritmo. Apesar de o processo ser não supervisionado, na maioria dos algoritmos o usuário tem grande influência sobre o resultado final através da escolha de parâmetros iniciais.

Esse problema pode facilmente aumentar o número de iterações do processo de mineração de dados, na medida em que diversos algoritmos com diferentes parametrizações podem conduzir a diferentes situações na busca de resultados interessantes [GP 2005].

Uma estratégia que pode ser adotada é a checagem de índices de validação que podem ser utilizados para definir ajustes na parametrização dos algoritmos de mineração de dados, validando os resultados obtidos pelos mesmos. A seguir serão descritos índices e medidas que podem ser utilizados com esta finalidade.

3.2.2.2.1 **Validação dos agrupamentos**

Técnicas de validação podem ser utilizadas para avaliar a qualidade dos agrupamentos obtidos. Essas técnicas permitem comparar diversos algoritmos de agrupamento, comparar duas partições, determinar o valor mais apropriado de parâmetros do algoritmo, entre outros. Segundo Souto [Souto 2009], medidas numéricas que são aplicadas para avaliar os vários aspectos da validação de agrupamentos são classificadas em três grupos :

- Índices Externos: Usados para avaliar o agrupamento gerado de acordo com uma estrutura pré-especificada, imposta ao conjunto de dados.
  Ex: Índice *Rand* ajustado (*adjusted Rand*) [Faceli et. al 2005] e índice de *Jaccard* [Faceli et. al 2005]
- Índices Internos: Usados para medir a qualidade de um agrupamento com base apenas nos dados originais (instâncias ou matriz de similaridade).
  Ex: Índice *Davies-Bouldin* [DB 1979], Índice *Dunn* [RS 2006], Silhuetas [RS 2006] e outros
- Índices Relativos: Usados para comparar diversos agrupamentos para decidir qual deles é o melhor em algum aspecto. Em geral, pode ser qualquer um dos índices acima definidos.

Neste trabalho foram utilizados os seguintes índices internos: índice de correlação cofenético, Largura de *Silhouette*, índice *Davies-Bouldin* e índice *Dunn*. Os índices Largura de *Silhouette*, índice *Davies-Bouldin* e índice *Dunn* foram obtidos através do programa *Machaon* CVE [Bolshakova 2009]. Foi utilizado também o índice externo *Rand* ajustado. Os índices citados anteriormente são descritos a seguir.

Coeficiente de Correlação Cofenético

O coeficiente de correlação cofenético é uma medida de validade para algoritmos de agrupamento hierárquico. Ele é usado para medir quão bem a estrutura hierárquica do dendograma

representa os relacionamentos existentes nos dados de entrada [RS 2006]. Segundo Metz [Metz 2006], o cálculo do coeficiente de correlação cofenético resulta em um valor entre 0 (zero) e 1 (um). Segundo esse índice, um dendograma reflete as estruturas embutidas nos dados quando esse valor se aproxima de 1 (um). Esse coeficiente de correlação é definido pela Equação 3.3.

$$= \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}c_{ij} - \mu_p\mu_c}{\sqrt{\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_p^2\right]\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_c^2\right]}} \qquad \text{Eq. 3.3}$$

Onde μ_p e μ_c são as médias da matriz de similaridade e matriz cofenética respectivamente, enquanto d_{ij} e c_{ij} são respectivamente os elementos da matriz de similaridade original do conjunto de dados e da matriz de similaridade cofenética obtida a partir do dendograma construído pelo algoritmo de agrupamento hierárquico [RS 2006][Metz 2006]. A matriz cofenética é um tipo de matriz de distância com a mesma dimensão da matriz de similaridade e, os elementos são representados pelo nível de dissimilaridade apresentado pelo dendograma (eixo vertical), em função do grupo a que pertence cada elemento.

Largura de *Silhouette*

Segundo Rao e Srinivas [RS 2006], a largura de *Silhouette* para um vetor de característica é uma medida de quão semelhante aquele vetor de característica é para os vetores de características do seu próprio agrupamento comparado com os vetores de características dos outros agrupamentos. A largura de *Silhouette s(i)* para o vetor de característica *i*th em um cluster *k* é definido pela Equação 3.4.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad \text{Eq. 3.4}$$

Na Equação 3.4, a(i) é a média da distância do vetor de característica *i*th para todos os outros vetores de característica no agrupamento *k*; b(i) é a distância média mínima do vetor de característica *i*th para todos os outros vetores de características em outro cluster j (j=1,...,*k*; j <> *k*). Desta fórmula segue que $-1 \leq s(i) \leq 1$.

Se s(i) está perto de 1, nós podemos deduzir que o vetor de característica ith foi atribuído a um agrupamento apropriado. De outro lado, quando s(i) está perto de -1, nós podemos concluir que o ith vetor de característica foi mal classificado. Quando s(i) é aproximadamente zero, isto indica que o ith vetor de característica está igualmente longe dos dois agrupamentos.

Além da largura de *Silhouette* de cada vetor de características, ainda podemos calcular a largura de *Silhouette* de cada grupo formado, como também a média global da largura de *Silhouette*, que é a média das larguras de *Silhouette* para todos os vetores de característica da base de dados. Este é um modo de escolher o melhor valor de *k* (número de grupos) e selecionar o melhor grupo formado.

Esse índice foi calculado tanto para os agrupamentos gerados pela aplicação dos algoritmos hierárquicos como para os agrupamentos gerados pela aplicação do algoritmo particional *K-Means*.

Índice *Davies-Bouldin*

Segundo Rao e Srinivas [DB 1979], o índice *Davies-Bouldin* é bastante conhecido pela sua habilidade de identificar grupos compactos e bem separados. O índice de *Davies-Bouldin* é uma função da razão da soma das distâncias intergrupo para a distância entre grupos. O menor valor da Equação 3.5 indica o melhor agrupamento.

$$DB = \frac{1}{n}\sum_{i-1}^{n} \max_{i \neq j}\left\{\frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)}\right\} \qquad \text{Eq. 3.5}$$

Onde *n* é o número de grupos, S_n é a distância entre os objetos dos grupos *i* e *j* e $S(Q_i,Q_j)$ é a distância entre os centróides dos respectivos grupos. Assim, quando temos grupos compactos e distantes dos demais grupos, o valor da Equação 3.5 é pequeno. Consequentemente, o melhor agrupamento será indicado pelo índice *Davies-Bouldin* com o menor valor.

Índice *Dunn*

O índice *Dunn*, assim como o índice de *Davies-Bouldin*, tem a habilidade de identificar grupos compactos e bem separados [RS 2006]. Este índice é obtido através da Equação 3.6.

$$D = \min_{1 \leq i \leq n}\left\{\min_{1 \leq j \leq n, i \neq j}\left\{\frac{d(c_i, c_j)}{\max_{1 \leq k \leq n}\{d'(c_k)\}}\right\}\right\} \qquad \text{Eq. 3.6}$$

Onde $d(c_i, c_j)$ é a distância entre os grupos c_i e c_j (distância intergrupo), $d'(c_k)$ é a distância intragrupo do grupo c_k e *n* é o número de grupos. O principal objetivo da medida é maximizar as distâncias intergrupo e minimizar as distâncias intragrupos. Portanto, o valor de *k* que maximiza D é considerado o número ótimo de grupos.

R*and* ajustado

O rand ajustado é um índice externo que avalia o agrupamento gerado baseado em uma estrutura pré-especificada (conjunto de dados) [Souto 2009]. Este índice é obtido através da Equação 3.7.

$$CR = \frac{\sum_{i=1}^{CA}\sum_{j=1}^{CB}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{CA}\binom{n_i}{2}\sum_{j=1}^{CB}\binom{n_j}{2}\right]/\binom{n}{2}}{\left[\sum_{i=1}^{CA}\binom{n_i}{2} + \sum_{j=1}^{CB}\binom{n_j}{2}\right]/2 - \left[\sum_{i=1}^{CA}\binom{n_i}{2}\sum_{j=1}^{CB}\binom{n_j}{2}\right]/\binom{n}{2}} \qquad \text{Eq. 3.7}$$

Onde n_{ij} representa o número de objetos comuns aos grupos C_i de A e C_j de B, n_i é o número de objetos no grupo C_i de A, n_j é o número de objetos no grupo C_j de B, CA e CB são os números de grupos nas partições A e B respectivamente.

O índice externo *Rand* ajustado pode assumir valores entre -1 e 1, 1 indicando uma concordância perfeita entre partições, e valores próximos a 0 ou negativos correspondendo a concordâncias encontradas ao acaso [Souto 2009].

3.2.3 **RESULTADOS**

No processo de mineração de dados, a visualização dos resultados é um instrumento valioso que permite ao usuário confirmar hipóteses, analisar o comportamento dos dados, sugerir novas idéias.

Os resultados dos agrupamentos obtidos pelos algoritmos hierárquicos foram apresentados através de dendogramas, que permitem visualizar além da sequência de agrupamentos, a similaridade com que os grupos são formados.

Os resultados obtidos pela rede de *Kohonen* foram visualizados através da rede com o rotulamento das bacias hidrográficas.

Os resultados da validação estatística dos algoritmos foram sumarizados em Tabelas e gráficos, que permitiram uma melhor comparação dos resultados obtidos por cada algoritmo executado.

3.3 **TRABALHOS RELACIONADOS**

É notória a preocupação com a pouca quantidade de dados tão necessários aos estudos dos mais variados problemas na área da ciência hidrológica. Na literatura encontramos várias pesquisas, no Brasil e no restante do mundo, no sentido de suprir essa deficiência. A seguir apresentaremos algumas dessas pesquisas que, como no presente trabalho, utilizam métodos de mineração de dados para atingirem os seus objetivos.

- Ramachandra Rao e Srinivas [RS 2006] investigaram o potencial do método de *cluster* híbrido na regionalização de bacias para análise de freqüência de enchentes. Eles procuraram unir as vantagens dos algoritmos hierárquicos e particionais, já que estudos que utilizaram os algoritmos em separado não obtiveram sucesso. Três algoritmos de *cluster* híbrido, formados pela combinação de algoritmos hierárquicos aglomerativo e algoritmo particional, foram utilizados com o objetivo de definir regiões homogêneas no estado de Indiana, USA. Os algoritmos hierárquicos utilizados foram o *Single-Linkage*, *Complete-Linkage* e *Ward* e o algoritmo particional, o *K-means*. Dados de cheias e atributos das bacias hidrográficas passaram por uma análise de correlação para garantir a independências dos dados escolhidos. Quatro índices de validação de *cluster*, a saber, coeficiente de correlação cophenético, índice de Silhoutte, índice de *Dunn* e índice de *Davies-Bouldin* foram testados para determinar a eficácia na identificação da partição ótima fornecido pelos algoritmos de *clustering*. Os resultados obtidos demonstraram que o desempenho geral do modelo híbrido, na otimização da função objetivo, foi melhor que as dos algoritmos de *clustering* hierárquico e particional usados separadamente. Dos três modelos híbridos apresentados, a combinação dos algoritmos *Ward* e *K-means* resultou numa boa estimativa de grupos de bacias sendo recomendado para regionalizar bacias hidrográficas. Foram identificadas seis regiões homogêneas aceitáveis.

- Demirel et al. [Demirel 2007] aplicou o algoritmo *K-means* para analisar o problema de seca nos rios da Turquia. O objetivo do estudo foi identificar regiões homogêneas de forma que efeitos hidrológicos pudessem ser comparados nessas subregiões, permitindo a transferência de informações de uma área para outra. A análise de componentes principais foi aplicada

nos dados de 80 estações fluviométricas que cobrem 23 bacias hidrográficas, referentes a um período de 31 anos. O resultado apresentou uma variação na densidade dos grupos formados, isto devido a uma distribuição desigual de estações fluviométricas representativas em cada bacia hidrográfica. A pesquisa concluiu que a análise de componentes principais em uma base de dados relativamente pequena (80x31) não é recomendada como etapa anterior à aplicação do algoritmo *K-Means*, e que trabalhos adicionais são necessários para que os resultados possam ser utilizados na validação de modelos de predição de fluxo intermitentes em curto prazo.

- Júnior et al. [Júnior 2006] realizou estudos para determinar regiões homogêneas quanto à distribuição de freqüência de chuvas no leste do Estado de Minas Gerais. Foram utilizados métodos estatísticos de análise multivariada (componentes principais rotacionados) e os algoritmos hierárquicos *Single-Linkage, Complete-Linkage* e *Ward*. Os dados foram coletados de 163 estações pluviométricas geograficamente distribuídas com informações, referentes à um período de 30 anos. Os estudos concluíram que: a análise de componentes principais rotacionados serviu apenas para se realizar uma análise preliminar da distribuição espacial das regiões hidroclimaticamente homogêneas, sendo questionável a avaliação da divergência em análise gráfica e o estabelecimento de grupos de similaridade de maneira subjetiva, com base na simples inspeção visual da dispersão; o método *Ward* apresentou a melhor representação espacial das regiões hidroclimaticamente homogêneas caracterizadas pelos grupos, em relação aos demais métodos empregados.

- Porto et al. [Porto 2004] aplicou a técnica de clusterização para identificar sub-bacias com características físicas similares no estado do Ceará. Foi utilizada a abordagem hierárquica utilizando o método *Ward*. O critério utilizado na seleção de variáveis teve como base a importância das características geomorfológicas da bacia hidrográfica na definição do comportamento hidrológico da mesma e a facilidade de obter tais informações através de mapas. As variáveis selecionadas foram: área de drenagem, comprimento do rio principal, densidade de drenagem e declividade do rio principal. As 30 sub-bacias estudadas foram agrupadas em 3 grupos distintos. O coeficiente de variação de todas as variáveis analisadas foi calculado para fins de uma análise comparativa e uma avaliação da variabilidade em cada grupo. As variáveis, área de drenagem e densidade de drenagem, se mostraram como as mais importantes na definição dos grupos. Os resultados demonstraram que a técnica de clusterização pôde ser de grande aplicabilidade na identificação de regiões hidrologicamente homogêneas subsidiando assim estudos de regionalização de dados hidrológicos.

3.4 **CONCLUSÃO**

Neste capítulo foram apresentadas todas as etapas da metodologia proposta, desde a descrição da base de dados até a etapa de visualização dos resultados.

Foram apresentadas também algumas pesquisas, não só no Brasil mas também no restante do mundo, que buscam soluções para amenizar a problemática da escassez de dados na área da ciência hidrológica.

No próximo capítulo serão apresentados os resultados obtidos em cada uma das etapas acima mencionadas.

CAPÍTULO IV

RESULTADOS

Este capítulo visa determinar o grau de significância dos resultados obtidos pelos algoritmos de clusterização, a saber, *Single-Linkage, Complete-Linkage*, *Ward*, *K-Means* e Rede de *Kohonen*.

Foram obtidos resultados da execução dos algoritmos acima citados, sobre o conjunto completo dos trinta e dois atributos das bacias hidrográficas e três subconjuntos de atributos gerados através do algoritmo de seleção de atributos e sobre os componentes principais, ambos gerados a partir do conjunto original de atributos. Esses quatro conjuntos de atributos formaram os cenários que serão descritos na próxima seção.

4.1 RESULTADO DA SELEÇÃO DE ATRIBUTOS SOBRE A BASE DE DADOS

4.1.1 Seleção de Atributos

Como citado no Capítulo III, foi utilizado o algoritmo proposto por Mitra [Mitra et al., 2002] para a tarefa de seleção de atributos e a Tabela 4.1 mostra a execução do algoritmo para os vários valores de k, onde k representa o número de vizinhos mais próximos a serem considerados, já que o particionamento dos atributos realizado por esse algoritmo é baseado no algoritmo K-NN .

Como o papel hidrológico de uma bacia hidrográfica é grandemente influenciado pelas características físicas da mesma, o critério de escolha do valor de *k* foi que os atributos escolhidos na iteração caracterizassem ao máximo a bacia hidrográfica.

De acordo com a Tabela 4.1, podemos ver que, para alguns valores de *k,* os atributos selecionados se repetiam. Até o valor de $k = 19$ os atributos selecionados representavam duas ou três categorias de características das bacias. Nas iterações com o valor de *k* menor ou igual 18 a representatividade aumenta com a inclusão de atributos de outras categorias de característica. Por essa razão, os valores de *k* escolhidos foram 16 e 18. O valor de $k = 16$ foi escolhido em virtude de nessa iteração serem selecionados atributos que representam uma maior diversidade de características das bacias hidrográficas, ficando as características de *forma, rede de drenagem* e *relevo* com várias representações. O valor de $k = 18$ foi escolhido pelo mesmo motivo, com o diferencial de que o número de atributos, representando as características, ficou mais resumido e as características do tipo *linear, forma e relevo* ficaram com apenas uma representação.

Tabela 4.1 – **Atributos selecionados para cada valor de *k***

k	Características da Bacia Hidrográfica						
	Linear	Forma	Rede de Drenagem	Relevo	Escoamento	Climatológica	Solo
31		K_f					
30	A			IG			
29	A			IG	L_{600}		
28	A	K_f			L_{600}		
27	A			I_p			
26	A			I_p			
25	A			I_p			
24	A		R_a	I_p			
23	A		R_a	I_{max}			
22	A		R_a	I_{max}			
21	A		R_a	I_{max}			
20	A		R_a	I_{max}			
19	A		R_a,IR	I_{max}			
18	A	K_e	R_a,IR	I_{max}			
17	L_d	K_c	O_r,R_b,R_l,R_a				
16	A	K_c,K_e,K_f	O_r,R_b,R_l,C_t	I_p,DS	L_{600}		
15	A	K_c,K_e,K_f	O_r,R_b,R_a,D_d,IR,SIN	C_{med}			
14	A	K_c,K_e,K_f	O_r,R_b,R_a,D_d,IR,SIN	C_{med} , I_{max}			
13	A	K_c,K_e,K_f	O_r,R_b,R_a,D_d,C_t,IR, SIN	C_{med} , I_{max}			
12	A	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	C_{med} , I_{max}			
11	A	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	C_{med} , I_{max}		P	
10	A	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	C_{med} , I_{max}	L_{600}	P	
09	A	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	C_{med} , I_{max}, IG	L_{600}	P	
08	A,L,L_t	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	C_{med} , I_{max}, I_p , IG	L_{600}	P	
07	P_r ,L,L_t	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	I_{max},L_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3
06	A	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	I_{max},L_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3
05	A	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	I_{max}, C_{med}, L_r , l_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3
04	P_r ,L_d	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	I_{max}, C_{med}, L_r , l_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3
03	A,P_r ,L_m	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,SIN	I_{max}, C_{med}, L_r , l_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3
02	A,P_r ,L_m	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,L_e ,SIN	I_{max}, C_{med}, L_r , l_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3
01	A,P_r, L, L_t , L_d L_m	K_c,K_e,K_f	O_r,R_b,R_l,R_a D_d,C_t, IR,L_e ,SIN	I_{max}, C_{med}, L_r , l_r , I_p , IG,DS	L_{600} , AE/A	P,E	SOLO1, SOLO2, SOLO3

Com os resultados obtidos pela execução do algoritmo acima citado, foram criados dois cenários, com a seguinte composição:

1. Cenário 1
   a. Atributos selecionados na iteração k = 16 (Tabela 4.1)

      A – área, K_c – Índice de compacidade, K_e – Índice de circularidade, K_f – Fator de Forma, O_r – Ordem do curso d'água, R_b – Índice de bifurcação, R_l – Índice dos Comprimentos, C_t – Coeficiente de torrencialidade, I_p – Índice de declividade média da bacia, Desnível específico, L_{600} – Aptidão média de escoamento

   b. Número de bacias igual a 41

2. Cenário 2
   a. Atributos selecionados na iteração k = 18 (Tabela 4.1)

      A – área, K_e – Índice de circularidade, R_a – Índice das áreas, IR – Índice de Rugosidade, I_{max} – Declividade máxima do rio

   b. Número de bacias igual a 41

4.1.2 **Componentes Principais**

Os 32 (trinta e dois) atributos das bacias hidrográficas foram submetidos à análise de componentes principais onde foi realizado o teste KMO (Medida de Adequação da Amostra), já descrito no Capítulo III. A cada rodada foi feito um exame do ajuste de cada atributo, através da matriz anti-imagem de correlação, e o que apresentava valor inferior a 0,5 foi sendo eliminado. Ao final do processo, restaram 23 atributos e foi obtido um KMO = 0,777, demonstrando assim uma adequação razoável dos dados à analise fatorial. Os atributos resultantes do processo descrito acima estão dispostos na Tabela 4.2.

Tabela 4.2 **– Atributos selecionados**

Área de drenagem	A
Perímetro	Pr
Linha de fundo	L
Comprimento do curso d'água	Lt
Comprimento da rede de drenagem	Ld
Largura média	Lm
Índice de compacidade	Kc
Índice de circularidade	Ke
Ordem do curso d'água	Or
Densidade de drenagem	Dd
Índice de rugosidade	IR
Extensão média do escoamento superficial	Le
Declividade máxima do rio	Imax
Elevação média da bacia	Cmed
Lado maior	Lr
Lado menor	lr
Índice de declividade média da bacia	Ip
Índice de declividade global	IG
Desnível específico	DS
Precipitação média da bacia	P
Aptidão média ao escoamento	L600
Percentual do solo tipo 1	SOLO1
Área do espelho d'água/área de drenagem	AE/A

Em uma análise fatorial considerando 23 variáveis, poderíamos ter até 23 componentes principais, correspondendo às variáveis originais. Os 23 componentes principais foram calculados e, após a análise dos autovalores, optamos por selecionar os que apresentaram autovalores maiores que 1 (Critério de *Kaiser*) e que explicam juntos 84,50% dos atributos originais. No Anexo A.1 e A.2 são mostrados a Scree Plot, que é um procedimento gráfico utilizado para determinar o número dos componentes principais a ser utilizado e, a tabela com os autovalores de cada componente principal, respectivamente.

A Tabela 4.3 apresenta os fatores de peso dos 5 componentes principais escolhidos e quantifica o total da variância do modelo, explicado por cada componente.

Tabela 4.3 - **Fatores de peso dos componentes principais**

Bacia Hidrográfica	**Escores dos componentes principais**				
	C1	**C2**	**C3**	**C4**	**C5**
BA1	-0,531	-0,576	-1,277	-0,758	-2,254
BA2	-0,646	0,945	-0,636	-0,113	1,270
BA3	-0,349	-0,795	0,099	-0,228	1,787
BA4	-0,531	0,324	-0,373	0,293	1,530
BA5	-0,182	-1,001	-0,420	-1,569	1,051
BA6	-0,650	0,262	-0,161	-0,151	0,468
BA7	-0,400	-1,855	-0,391	2,481	-1,017
BA8	-0,489	1,208	0,255	-1,044	-0,078
BA9	0,023	0,162	-0,461	1,158	0,364
BA10	3,102	0,202	-0,191	-1,068	-0,367
BA11	-0,445	2,099	-0,794	0,921	0,100
BA12	-0,459	-1,206	0,129	1,941	0,263
BA13	-0,581	-0,412	2,875	1,628	-0,706
BA14	-0,548	0,571	-0,711	0,241	1,190
BA15	-0,517	0,072	-0,298	-0,331	-0,813
BA16	-0,674	1,941	-0,627	-0,279	0,053
BA17	-0,507	-0,167	-0,143	0,065	1,536
BA18	-0,369	-1,287	-0,633	-0,613	-0,774
BA19	-0,520	0,048	-0,275	0,967	1,207
BA20	-0,084	0,809	-0,586	-0,902	-1,069
BA21	-0,567	-0,277	1,828	-0,254	-0,879
BA22	0,068	0,281	-0,840	1,953	-1,011
BA23	-0,533	-0,319	0,399	-0,894	0,607
BA24	-0,823	1,084	-0,804	-0,775	-1,693
BA25	-0,322	-2,217	-0,395	-1,595	0,907
BA26	-0,430	-1,087	-0,609	-0,887	0,278
BA27	-0,456	-0,711	3,465	-0,530	-0,047
BA28	-0,070	2,157	1,615	-0,498	0,492
BA29	1,037	0,614	0,295	-0,857	0,671
BA30	1,485	-0,665	-0,414	1,175	2,044
BA31	-0,250	0,311	-0,228	0,344	-1,641
BA32	0,048	0,668	-0,299	0,435	0,134
BA33	0,596	1,281	-0,484	1,033	0,481
BA34	0,383	-0,176	0,695	0,365	-0,499
BA35	0,807	0,127	0,601	1,449	-0,566
BA36	1,404	-0,479	-0,639	0,084	-0,895
BA37	0,646	0,684	1,958	-1,132	-0,199
BA38	4,040	-0,356	-0,093	-0,126	-0,297
BA39	-0,480	-1,517	-0,476	-0,742	0,187
BA40	-0,619	-0,815	-0,448	-0,557	-0,874
BA41	-0,606	0,067	-0,506	-0,631	-0,941
Autovalor	**10,866**	**3,570**	**2,430**	**1,441**	**1,131**
% variança total	**47,242**	**15,521**	**10,566**	**6,265**	**4,915**

Um novo significado é atribuído a cada componente principal em virtude do mesmo ser uma combinação linear dos atributos originais. A seguir são apresentados os significados das novas variáveis considerando a importância dos atributos originais de forma ponderada, conforme Anexo A.3. A Tabela 4.4 resume a descrição abaixo.

C1 – Componente Linear - esse componente está relacionado com as grandezas lineares da bacia hidrográfica que são: área de drenagem (A), perímetro (P), linha de fundo (L), comprimento do curso d'água (Lt), comprimento da rede de drenagem (Ld), largura média (Lm), extensão média do escoamento superficial (Le) e lado maior (Lr);

C2 – Componente Forma – esse componente está relacionado com a forma da bacia hidrográfica e é explicada pelas variáveis: índice de circularidade (Ke), índice de compacidade (Kc) e índice de rugosidade (IR);

C3 – Componente Escoamento – esse componente está relacionado com a capacidade de escoamento da bacia hidrográfica e é explicado pelas variáveis: percentual do solo tipo I (SOLO1) e aptidão média de escoamento (L600);

C4 – Componente Relevo – esse componente está relacionado com o relevo da bacia hidrográfica e é explicado pela variável: desnível específico (DS);

C5 – Componente Drenagem – esse componente está relacionado com a eficiência da drenagem da bacia hidrográfica e é explicado pela variável: densidade de drenagem (Dd).

Tabela 4.4 – **Componentes Principais**

Componentes Principais	Significado	Variáveis Explicativas
C1	Linear	A, Pr, L, Lt, Ld, Lm, Le e Lr
C2	Forma	Ke, Kc, IR
C3	Escoamento	SOLO1, L600
C4	Relevo	DS
C5	Drenagem	Dd

Com os componentes principais gerados temos a criação do terceiro cenário, com a seguinte composição:

3. Cenário 3
    a. Componentes principais (Tabela 4.4)
    b. Número de bacias igual a 41

Um quarto cenário foi formado com o conjunto completo de atributos das bacias hidrográficas, como segue:

4. Cenário 4
    a. 32 atributos (Tabelas 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 e 3.7)
    b. Número de bacias igual a 41

4.2 RESULTADOS DOS ALGORITMOS

Todos os algoritmos foram executados utilizando os cenários descritos anteriormente.

4.2.1 Algoritmos Hierárquicos

Os resultados dos algoritmos hierárquicos serão representados pelos dendogramas que são estruturas em formato de árvore, nas quais os elementos são dispostos no eixo horizontal, e a distância (ou a similaridade) com que os os grupos são gerados, no eixo vertical. Desse modo, o dendograma é uma estrutura com toda a hierarquia dos agrupamentos gerados sobre o conjunto de elementos inicial.

4.2.1.1 *Single-Linkage*

Os dendogramas gerados por esse algoritmo demonstraram que não houve a formação de grupos bem definidos e separados, conforme ilustram as Figuras 4.1, 4.2, 4.3 e 4.4. Nessa apresentação não há diferenças significativas no tamanho dos arcos que representam os grupos, o que dificulta a escolha da altura para o corte do dendograma. O coeficiente de correlação cofenético, para cada cenário, é apresentado na Tabela 4.5. Apesar do algoritmo *Single-Linkage* apresentar um ótimo valor para esse coeficiente, não podemos afirmar que o dendograma gerado pelo mesmo representa uma boa estrutura de grupos formados, confirmando assim as afirmações apresentadas na literatura, [Metz 2006] e [RS 2006] de que a análise do coeficiente de correlação cofenético não é suficiente para identificar os melhores dendogramas nos modelos hierárquicos, sendo ainda necessária a inspeção visual dos mesmos.

Tabela 4.5 – **Coeficiente Cofenético**

Cenário	Algoritmos	Coeficiente Cofenético
I	*Single-Linkage*	0,71
	Complete-Linkage	0,57
	Ward	0,54
II	*Single-Linkage*	0,78
	Complete-Linkage	0,55
	Ward	0,54
II	*Single-Linkage*	0,70
	Complete-Linkage	0,51
	Ward	0,54
IV	*Single-Linkage*	0,71
	Complete-Linkage	0,52
	Ward	0,57

Cenário I

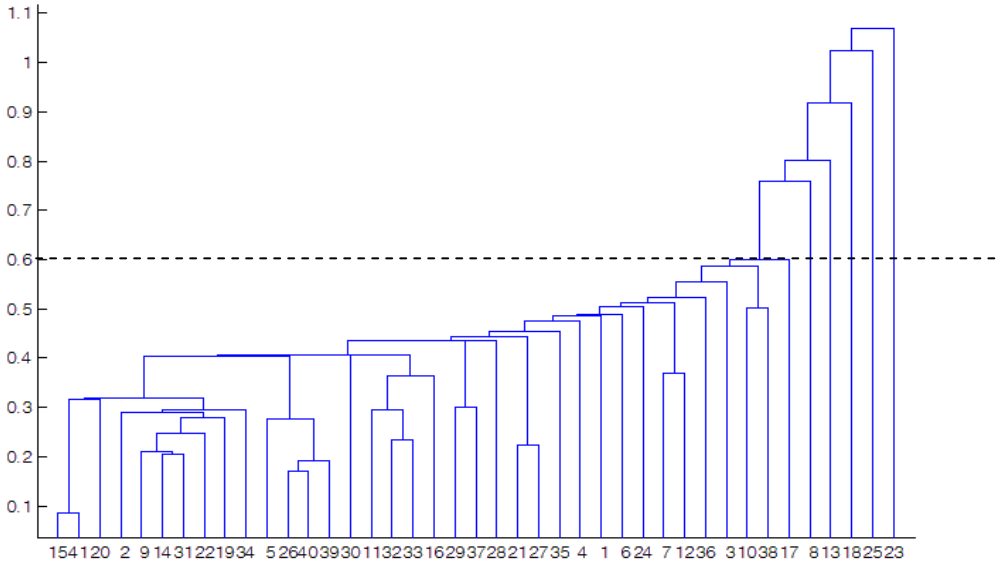


Figura 4.1 – **Dendograma gerado pelo *Single-Linkage* no Cenário I**

Cenário II

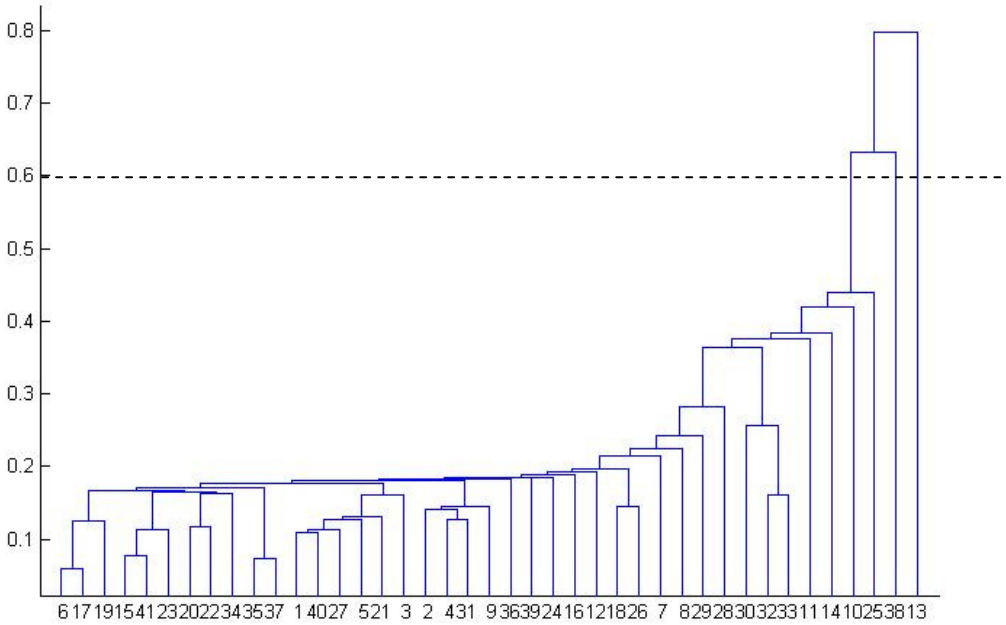


Figura 4.2 – **Dendograma gerado pelo *Single-Linkage* no Cenário II**

Cenário III

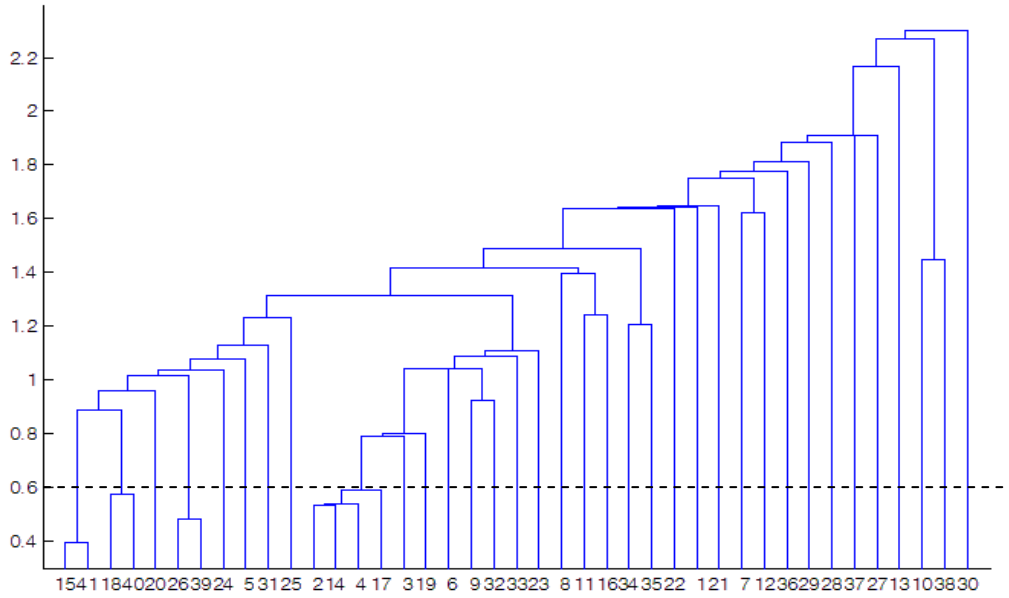


Figura 4.3 – **Dendograma gerado pelo *Single-Linkage* no Cenário III**

Cenário IV

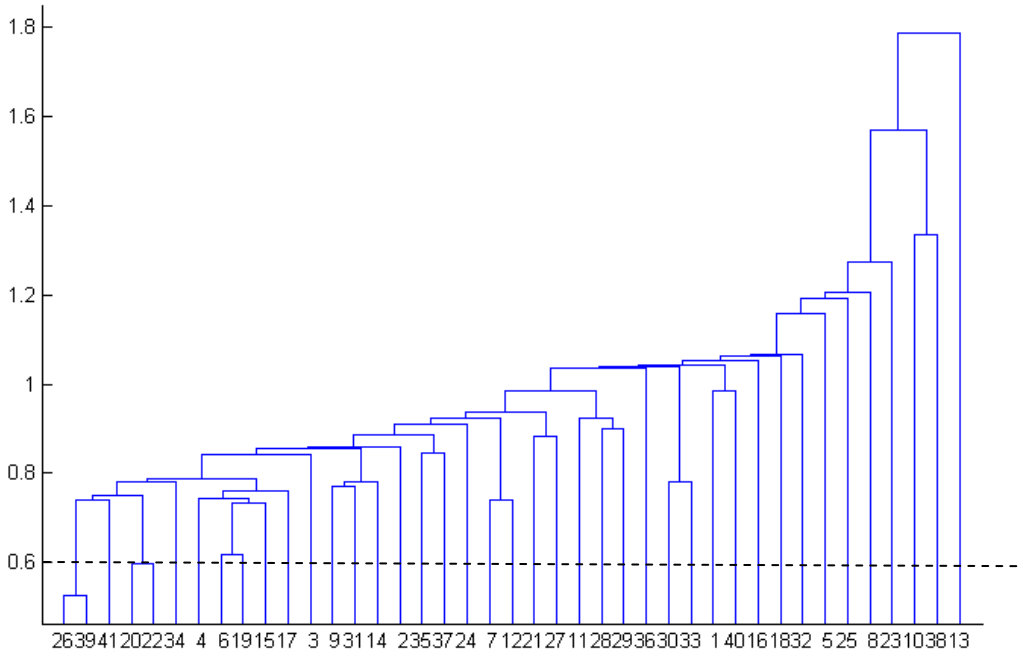


Figura 4.4 – **Dendograma gerado pelo *Single-Linkage* no Cenário IV**

4.2.1.2 *Complete-Linkage*

Conforme Figuras 4.5, 4.6, 4.7 e 4.8, os dendogramas gerados por esse algoritmo apresentaram grupos mais bem definidos quando comparado com o resultado produzido pelo algoritmo *Single-Linkage*, portanto realça ainda uma grande individualidade das bacias. Diferenças significativas entre as alturas dos arcos aparecem desde as primeiras junções significando um grau de dispersão elevado entre os elementos dentro do grupo. O coeficiente de correlação cofenético, para cada cenário, é apresentado na Tabela 4.5. Uma visualização conjunta dos dendogramas gerados pelos algoritmos hierárquicos encontra-se no Anexo B.

Cenário I

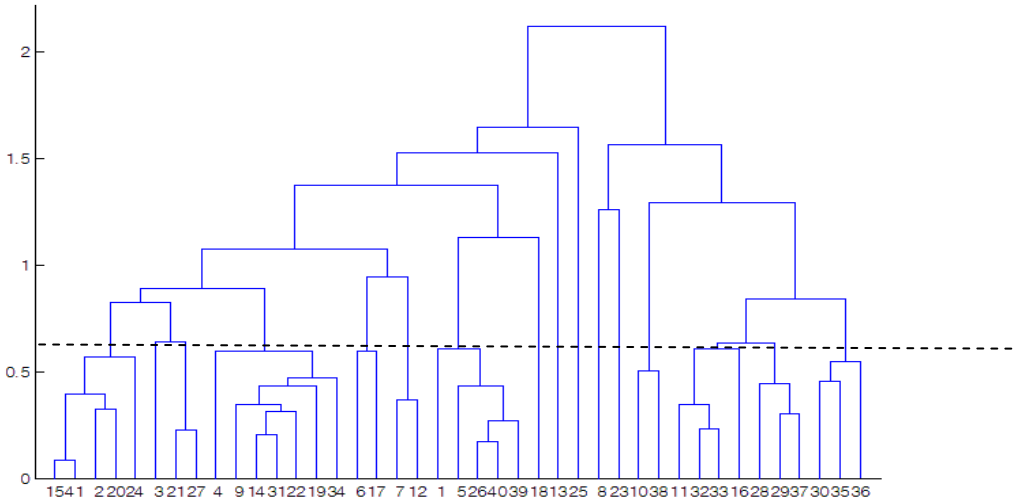


Figura 4.5 – **Dendograma gerado pelo *Complete-Linkage* no Cenário I**

Cenário II

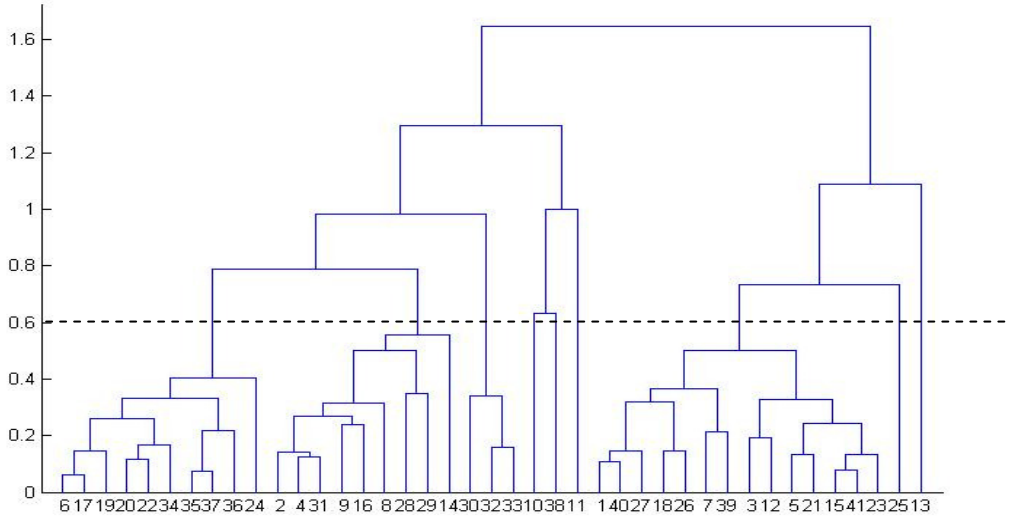


Figura 4.6 – **Dendograma gerado pelo *Complete-Linkage* no Cenário II**

Cenário III

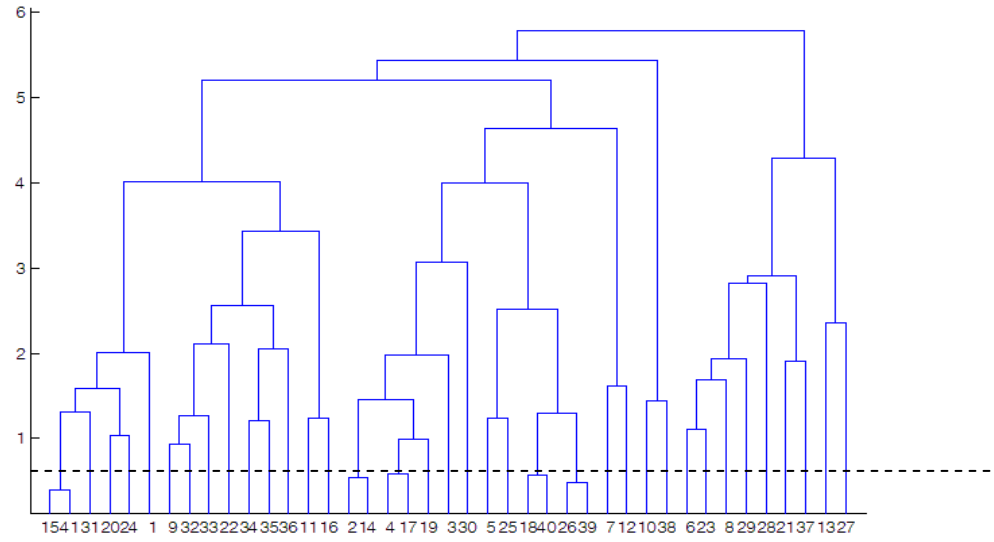


Figura 4.7 – **Dendograma gerado pelo *Complete-Linkage* no Cenário III**

Cenário IV

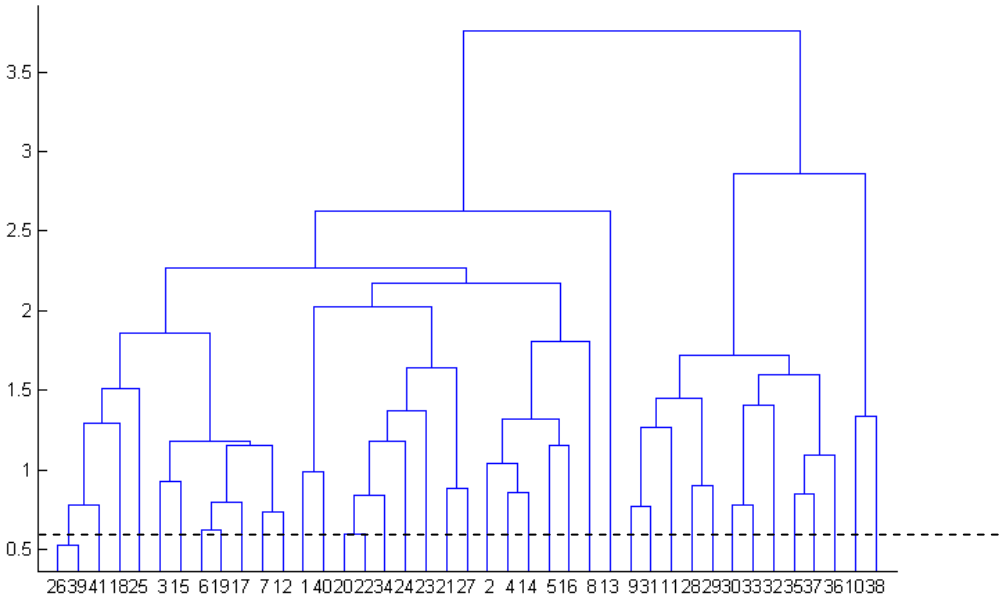


Figura 4.8 – **Dendograma gerado pelo *Complete-Linkage* no Cenário IV**

4.2.1.3 *Ward*

Apesar de apresentarem o coeficiente de correlação cofenético baixo, conforme Tabela 4.5, os dendogramas gerados pelo algoritmo *Ward*, como podemos observar nas Figuras 4.9, 4.10, 4.11 e 4.12, apresentaram uma melhor estrutura de grupos quando comparado com os dois algoritmos anteriores.

Cenário I

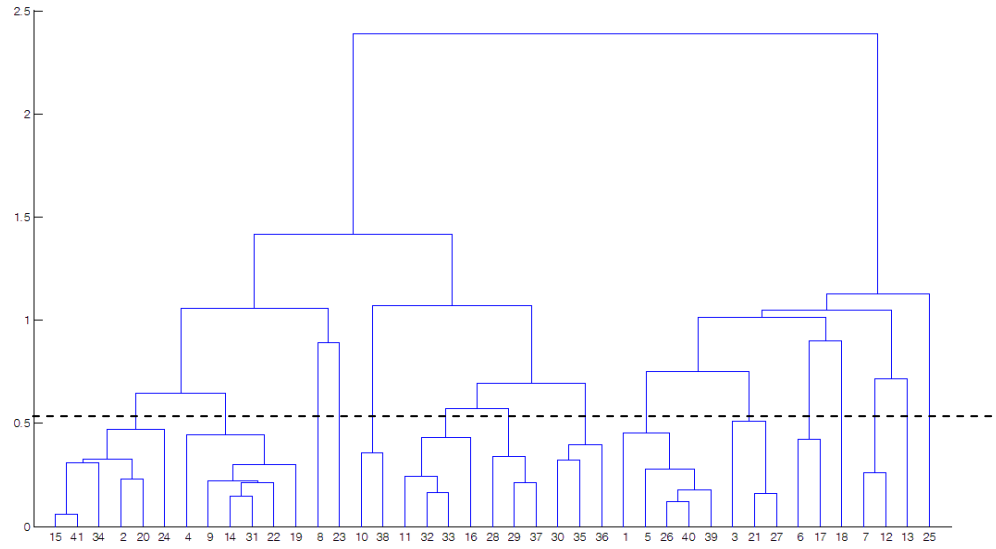


Figura 4.9 – **Dendograma gerado pelo *Ward* no Cenário I**

Cenário II

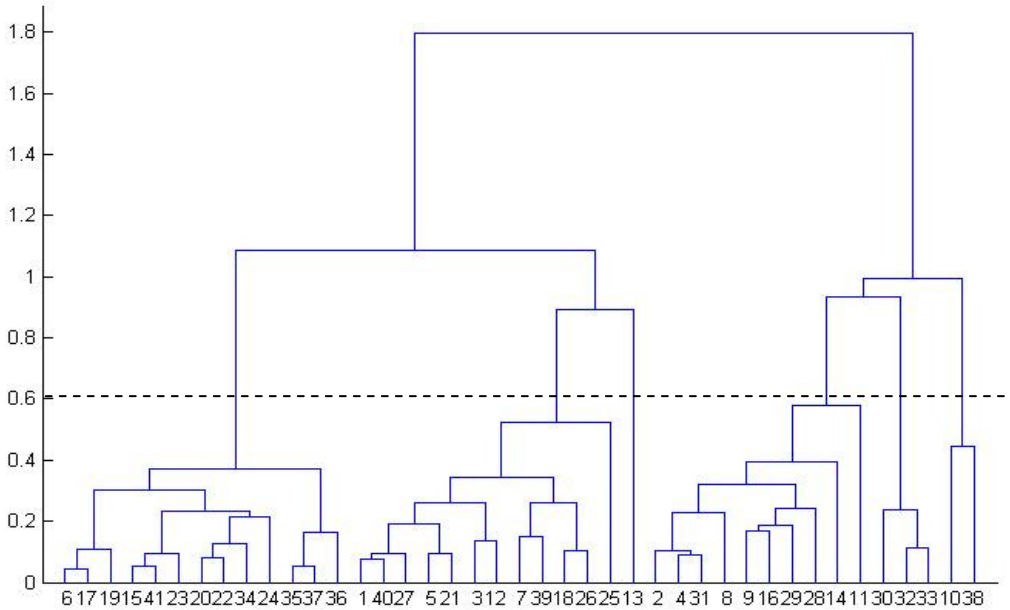


Figura 4.10 – **Dendograma gerado pelo *Ward* no Cenário II**

Cenário III

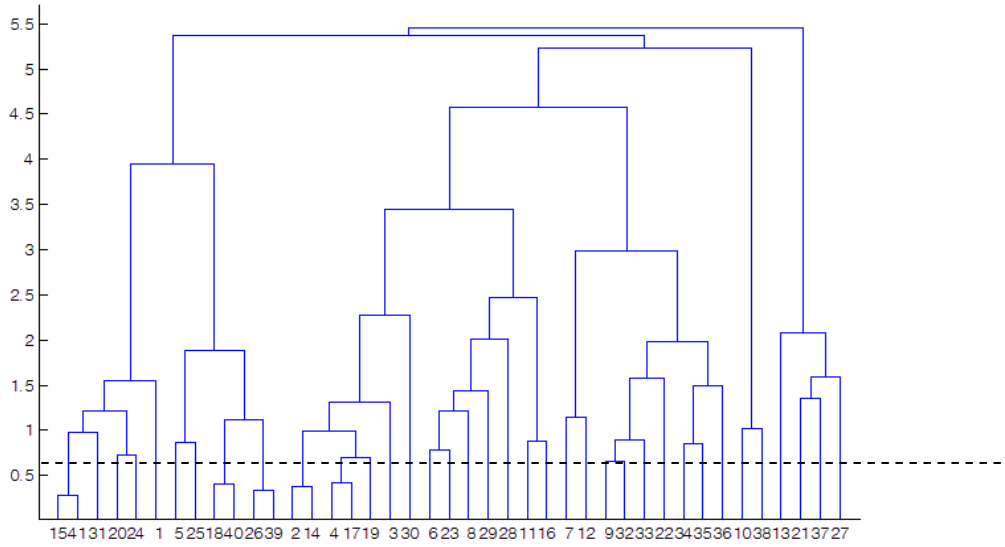


Figura 4.11 – **Dendograma gerado pelo *Ward* no Cenário III**

Cenário IV

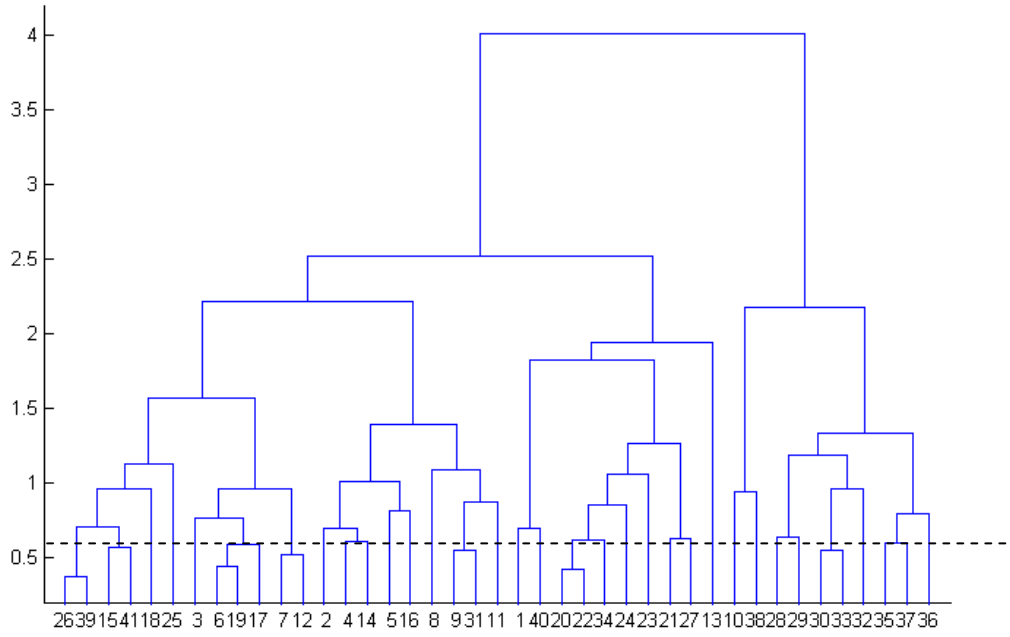


Figura 4.12 – **Dendograma gerado pelo *Ward* no Cenário IV**

Analisando os dendogramas gerados por cada um dos algoritmos hierárquicos conclui-se que o algoritmo *Ward* apresentou o melhor resultado. O dendograma gerado pelo algoritmo *Single-Linkage* mostrou uma degeneração na formação dos grupos, dando a impressão de um grupo único. Nos dendogramas gerados pelo algoritmo *Complete-Linkage* verifica-se que existe uma diferença significativa nos arcos desde as primeiras iterações. Ao fazer um corte na altura 0,6 observa-se a formação de vários grupos com apenas uma bacia. Fazendo um corte em uma altura maior

encontram-se agrupamentos com um número maior de bacias porém com um grau de dispersão elevado. Portanto, concluímos que os algoritmos *Single-Linkage* e *Complete-Linkage* não são indicados para a tarefa de clusterização a que se propõe esse trabalho.

Os dendogramas gerados pelo algoritmo *Ward* apresentaram uma formação de grupos bem compactos onde, os arcos iniciais que unem os grupos são bem pequenos, indicando assim uma distância, entre os elementos de cada grupo, próxima de zero. À medida que o número de iterações aumenta, os grupos vão se unindo em grupos maiores com distâncias maiores, e consequentemente arcos mais altos. Com isso percebe-se a formação de grupos bem definidos, o que facilita a escolha da altura onde o dendograma deve ser cortado, conforme pode ser observado na Figura 4.10.

Como já mencionado no Capítulo III, a validação dos agrupamentos permite comparar diversos algoritmos de clusterização, comparar duas partições, determinar o valor mais apropriado de algum parâmetro do algoritmo, entre outros. Os índices de validação foram aqui calculados sobre os resultados do algoritmo *Ward* nos quatro cenários propostos para execução dos algoritmos de Clusterização. A Tabela 4.6 apresenta os valores dos índices de *Silhouette*, *Davies-Bouldin* e *Dunn* em cada cenário.

Tabela 4.6 - **Índices de Validação dos Agrupamentos gerados pelo *Ward***

Cenários	Nº de Grupos					
	Nº de Grupos = 2			Nº de Grupos = 3		
	Silhouette	*D. Bouldin*	*Dunn*	*Silhouette*	*D. Bouldin*	*Dunn*
I	0,264	1,516	1,285	0,165	1,681	0,949
II	0,340	1,387	1,439	0,219	1,345	0,997
III	0,144	1,815	1,089	0,179	1,756	0,935
IV	0,237	1,395	1,431	0,171	1,649	1,004
Cenários	Nº de Grupos					
	Nº de Grupos = 4			Nº de Grupos = 6		
	Silhouette	*D. Bouldin*	*Dunn*	*Silhouette*	*D. Bouldin*	*Dunn*
I	0,196	1,468	1,026	0,204	1,497	0,711
II	0,253	1,502	0,976	0,365	1,085	1,218
III	0,227	1,703	0,928	0,043	1,669	0,816
IV	0,141	1,661	0,840	0,184	1,446	1,084

Os índices de *Silhouette* e *Dunn* sinalizam um bom resultado quando o seu valor é maximizado. O índice de *Davies-Bouldin* se comporta de forma contrária aos demais, sinalizando, portanto um bom resultado quando o seu valor é minimizado.

As Figuras 4.13, 4.14 e 4.15 apresentam a variação dos índices de validação em cada cenário e em cada configuração do número de grupos.

Com a análise do comportamento dos índices de validação conclui-se que os agrupamentos gerados no cenário II apresentaram os melhores resultados e quanto ao número de grupos, a partição ótima se deu com o número de grupos igual a 6.

O índice de *Silhouette* obteve o melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 6 grupos, conforme Figura 4.13. Pode-se observar também que este índice apresenta um comportamento semelhante para os números de grupos 2 e 6, onde os valores aumentam nos cenários II e IV e decrescem nos cenários I e III. Para os números de grupos 3 e 4, o índice de *Silhouette* apresenta comportamento idêntico, aumentando o seu valor no cenário II e tendo uma queda nos cenários I, III e IV .

O índice *Davies-Bouldin*, que identifica grupos compactos e separados, também obteve o melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 6 grupos, conforme Figura 4.14. O comportamento deste índice é semelhante nos números de grupos 2,3 e 6, onde apresenta os melhores valores nos cenários II e IV. No número de grupos 4 observa-se um comportamento diferenciado, apresentando seu melhor valor no cenário I, passando a aumentar nos cenários II e III e diminuir no cenário IV.

O índice *Dunn* obteve o seu melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 2 grupos, mas com uma variação pequena em relação ao valor obtido na divisão das bacias em 6 grupos, conforme Figura 4.15. O comportamento do índice *Dunn* é semelhante ao do índice *Davies-Bouldin* para os números de grupos 2,3 e 6. Para o número de grupos 4, este índice apresenta o melhor valor no cenário I, passando a diminuir nos cenários II, III e IV.

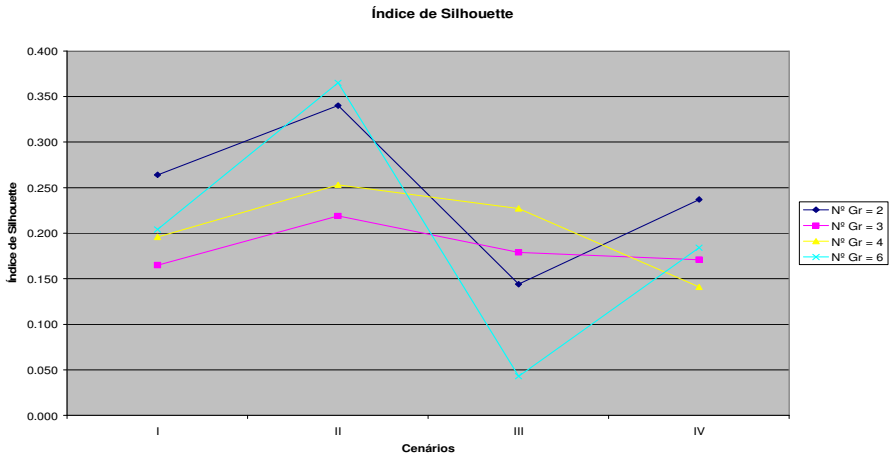


Figura 4.13 – **Comportamento do índice de *Silhouette* - *Ward*.**

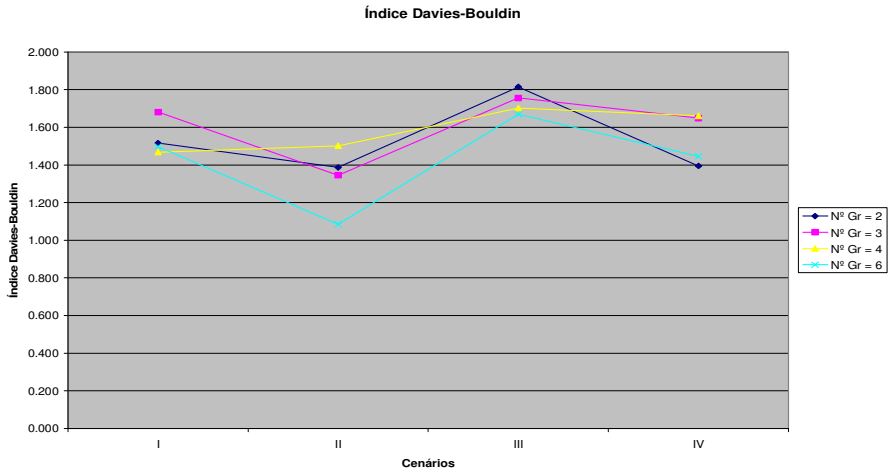


Figura 4.14 – **Comportamento do índice *Davies-Bouldin* - *Ward*.**

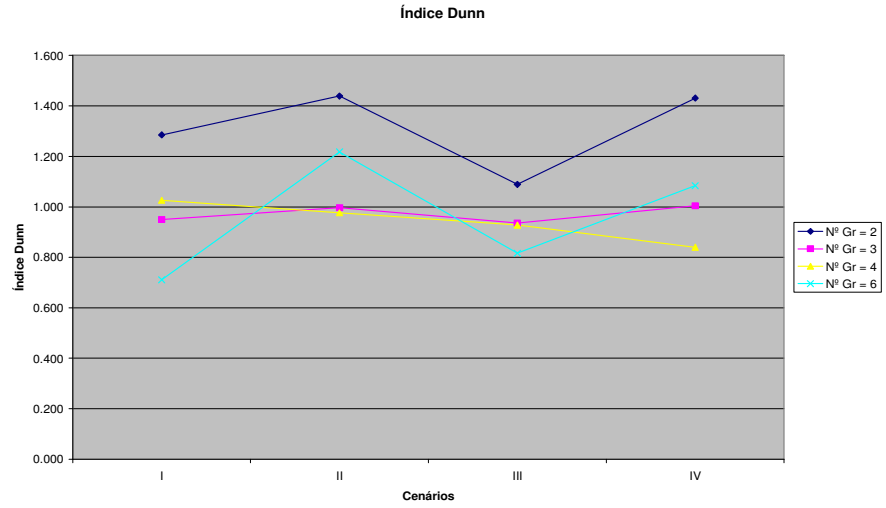


Figura 4.15 – **Comportamento do índice *Dunn* - *Ward*.**

4.2.2 Algoritmo Particional

4.2.2.1 *K-Means*

Na execução do algoritmo *K-Means*, foram utilizados como parâmetros: o número de grupos (k = 2,3,4 e 6), a medida de distância Euclidiana, a seleção dos k centróides iniciais de forma randômica e o número máximo de iterações igual a 100.

Foram realizadas para cada cenário 10 simulações para cada valor de k. Ao final de cada simulação é apresentado o valor da soma das distâncias *inter-cluster*, ou seja, a soma das distâncias de cada ponto do *cluster* ao seu centróide. Com os agrupamentos gerados em cada iteração foram calculados os índices de validação e obtida a média e o desvio padrão dos valores de cada índice. O resultado pode ser observado na Tabela 4.7.

Tabela 4.7. **Índices de Validação dos Agrupamentos gerados pelo *K-Means***

Cenários	Nº de Grupos											
	Nº de Grupos = 2						Nº de Grupos = 3					
	Silhouette		*D. Bouldin*		*Dunn*		*Silhouette*		*D. Bouldin*		*Dunn*	
	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd
I	0,261	0	1,516	0	1,285	0	0,224	0,039	1,604	0,187	1,002	0,061
II	0,309	0,007	1,380	0,031	1,396	0,061	0,295	0,043	1,453	0,113	1,046	0,090
III	0,176	0,046	1,818	0,118	1,015	0,042	0,177	0,016	1,822	0,117	0,943	0,053
IV	0,176	0,064	1,557	0,211	1,179	0,330	0,170	0,018	1,578	0,085	1,002	0,017
Cenários	Nº de Grupos											
	Nº de Grupos = 4						Nº de Grupos = 6					
	Silhouette		*D. Bouldin*		*Dunn*		*Silhouette*		*D. Bouldin*		*Dunn*	
	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd
I	0,202	0,022	1,620	0,099	0,998	0,093	0,205	0,025	1,502	0,102	0,875	0,137
II	0,315	0,038	1,380	0,094	0,980	0,059	0,281	0,066	1,233	0,117	0,713	0,249
III	0,200	0,021	1,673	0,080	0,946	0,068	0,243	0,019	1,436	0,101	0,985	0,163
IV	0,147	0,017	1,610	0,047	0,880	0,058	0,143	0,023	1,550	0,054	0,844	0,115

As Figuras 4.16, 4.17 e 4.18 apresentam a variação dos índices de validação em cada cenário e em cada configuração do número de grupos.

O índice de *Silhouette* obteve o melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 4 grupos, conforme Figura 4.16. Este índice apresentou um comportamento semelhante em todas configurações de números de grupos (2, 3, 4 e 6), com os menores valores nos cenários I, III e IV e melhor valor no cenário II.

O índice *Davies-Bouldin* obteve o melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 6 grupos, conforme Figura 4.17. O comportamento deste índice é semelhante para todos os números de grupos nos cenários I, II e III, apresentando uma diferença apenas para o número de grupos 6, no cenário IV, que ao contrário dos demais passa a ter menor valor.

O índice *Dunn* obteve o seu melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 2 grupos, conforme Figura 4.18. Os números de grupos 2 e 3 apresentaram um comportamento semelhante, com os melhores valores no cenário II e o pior valor no cenário III. No número de grupos igual 4, o índice teve um comportamento descendente a partir do cenário I até o cenário IV. No número de grupos 6, o índice *Dunn* apresentou um comportamento inverso ao apresentado pelos números de grupos 2 e 3, com o melhor valor no cenário III e o pior valor no cenário II.

Os resultados alcançados pela validação sobre os agrupamentos gerados pelo algoritmo *K-Means* não permitiram encontrar o número ótimo de grupos, mas confirmaram mais uma vez que a qualidade dos agrupamentos gerados no cenário II é superior aos gerados nos demais cenários.

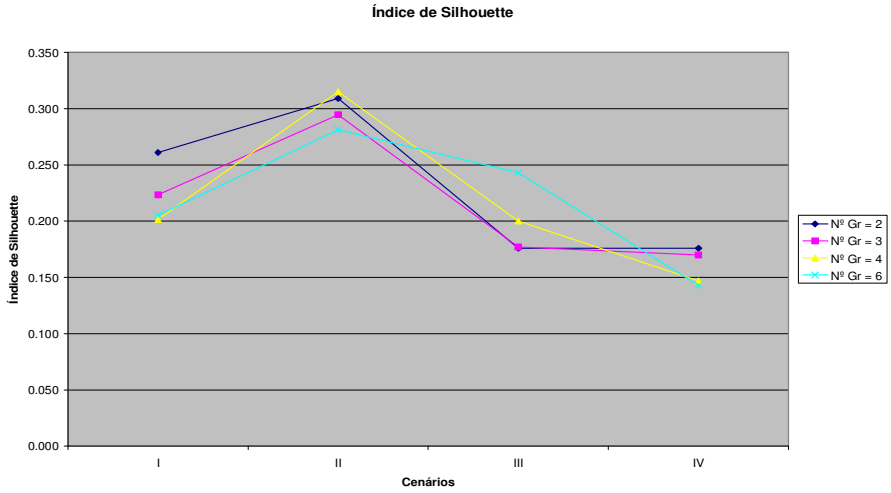


Figura 4.16– **Comportamento do índice de *Silhouette – K-Means*.**

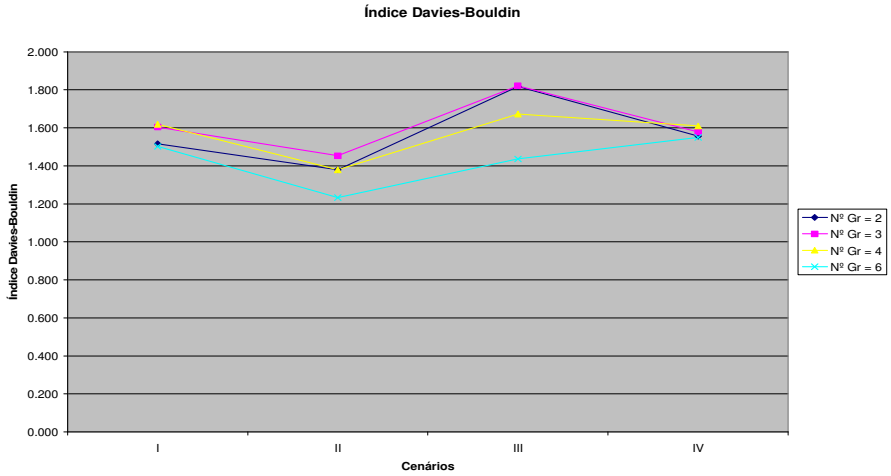


Figura 4.17 - **Comportamento do índice *Davies-Bouldin – K-Means*.**

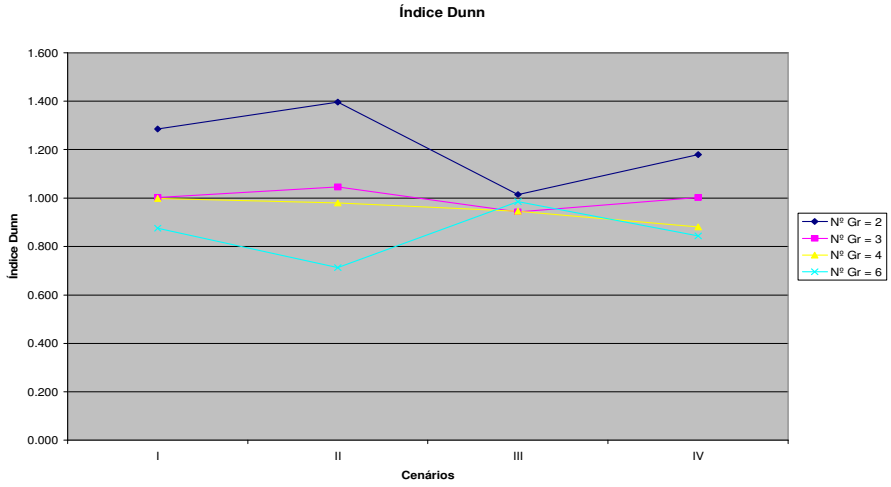


Figura 4.18- **Comportamento do índice *Dunn – K-Means*.**

4.2.3 **Rede Neural de *Kohonen***

4.2.3.1 **Configuração da Rede de *Kohonen***

Os parâmetros que regulam a rede neural de *Kohonen* são muitos, portanto é recomendado efetuar diversos testes com várias configurações para escolher, baseado em algum critério de qualidade, redes razoavelmente bem ajustadas.

Duas etapas foram realizadas para definir a melhor configuração da rede a ser utilizada no processo de agrupamento. A primeira etapa foi realizada para decidir qual a dimensão da rede a ser utilizada e a segunda, para definir a melhor configuração com relação aos parâmetros de topologia, função de vizinhança e número de iterações das fases de ordenação e convergência do processo adaptativo dos pesos sinápticos da rede.

Na primeira etapa foram testadas várias dimensões da rede definidas como 2x2, 3x3, 3x4, 4x5, 4x6, 4x7, 5x5, 5x6, 6x6, 6x7, 8x3. Em cada treinamento foram obtidos os valores do erro de quantização e o erro topográfico. Em virtude de na maior parte das configurações o erro topográfico apresentar valor igual a zero, conforme o Anexo C.33, a escolha da melhor configuração foi baseada no valor do erro de quantização. A Tabela 4.8 apresenta, para todos os cenários, os valores do erro de quantização para cada configuração com o destaque para a configuração 6x7 que apresentou o melhor desempenho sendo, portanto eleita a configuração a ser utilizada na aplicação da rede neural de *Kohonen*.

Tabela 4.8 – **Erro de Quantização – Dimensão da Rede**

Cenários	[2X2]	[3X3]	[3X4]	[4X5]	[4X6]	[4X7]	[5X5]	[5X6]	[6X6]	[6X7]	[8X3]
I	0,542	0,494	0,487	0,443	0,424	0,405	0,417	0,397	0,387	**0,359**	0,413
II	0,315	0,289	0,262	0,227	0,221	0,217	0,211	0,202	0,196	**0,184**	0,210
III	1,879	1,686	1,612	1,427	1,369	1,324	1,334	1,274	1,161	**1,149**	1,344
IV	1,023	0,947	0,918	0,850	0,828	0,801	0,820	0,767	0,770	**0,723**	0,825

Com o tamanho da rede definido passamos para a segunda etapa. Foram realizadas várias simulações para escolher a topologia (hexagonal ou retangular), a função de vizinhança (*guassian*, *cutgauss, epanechnikov, bubble*) e o número de iterações para a fase de ordenação e convergência dos pesos sinápticos. O valor do erro de quantização foi utilizado como referência para a escolha da melhor configuração. Para cumprir esta etapa vários passos foram seguidos e a seguir temos a descrição de cada um deles e seus respectivos resultados.

1º Passo: Simulação para escolher a topologia

- Tamanho da rede [6x7]
- Função de vizinhança gaussiana (default)

- Fases de treinamento 1x2

Tabela 4.9 – **Erro de Quantização - Topologia da rede**

Cenários	Hexagonal	Retangular
I	0,357	**0,353**
II	0,187	**0,186**
III	1,152	**1,062**
IV	0,731	**0,724**

2º Passo: Simulação para escolher a função de vizinhança

- Tamanho da rede [6x7]
- Topologia hexagonal e retangular
- Fases de treinamento 1x2

Tabela 4.10 – **Função de vizinhança**

	Topologia hexagonal				Topologia retangular			
Cenários	*gaussian*	*cutgauss*	*ep*	*bubble*	*gaussian*	*cutgauss*	*ep*	*bubble*
I	0,358	0,353	0,373	0,369	**0,348**	0,357	0,362	0,352
II	0,178	0,190	0,180	0,183	**0,178**	0,184	0,180	0,175
III	1,099	1,095	1,127	1,194	**1,072**	1,097	1,097	1,107
IV	0,734	0,715	0,746	0,743	**0,723**	0,731	0,730	0,730

No primeiro passo variamos a topologia da rede e utilizamos os valores default dos demais parâmetros. De acordo com a Tabela 4.9 podemos observar que o erro de quantização foi minimizado na topologia retangular. No segundo passo variamos a função de vizinhança e obtivemos os valores do erro de quantização utilizando as topologias hexagonal e retangular. Como mostra a Tabela 4.10, o erro de quantização apresentou melhor resultado na configuração com topologia retangular e função de vizinhança guassiana.

Com os parâmetros de topologia e função de vizinhança selecionados passamos para o terceiro passo que foi a escolha do número de épocas das fases de ordenação e convergência do processo adaptativo dos pesos sinápticos da rede.

3º Passo: Simulação para escolher o número de épocas da fase de treinamento da rede

- Tamanho da rede [6x7]
- Topologia retangular
- Função de vizinhança gaussiana

De acordo com os resultados apresentados na Tabela 4.11 e na Figura 4.19, a configuração escolhida foi a 30X120. Ao longo dos testes o erro de quantização teve o seu valor decrementado

até configuração 30X120 em todos os cenários, voltando a aumentar a partir da configuração 40x160 nos cenários I, II e III.

Após a definição dos parâmetros acima citados chegamos à seguinte configuração da rede: tamanho da rede [6x7], topologia retangular, função de vizinhança gaussiana e nº de épocas 30x120.

Tabela 4.11 – **Erro de Quantização – Nº de Épocas**

Cenários	Configurações					
	10x40	20x80	**30x120**	40x160	60x240	80x320
I	0,357	0,344	**0,319**	0,330	0,327	0,314
II	0,185	0,173	**0,165**	0,171	0,153	0,152
III	1,122	0,994	**0,949**	0,976	0,942	0,943
IV	0,727	0,698	**0,681**	0,663	0,664	0,663

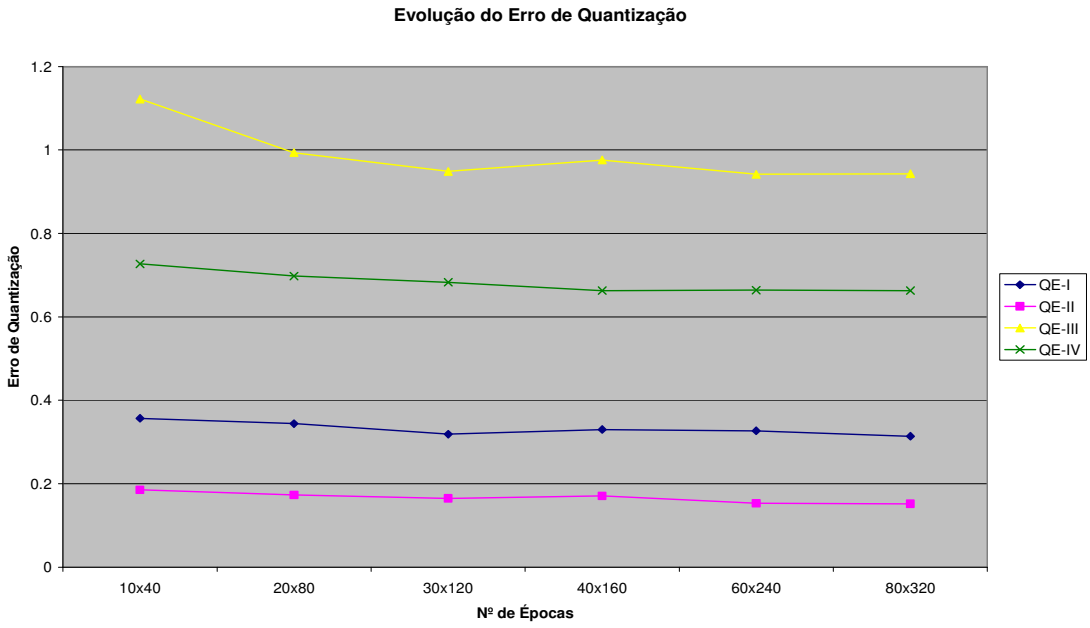


Figura 4.19 – **Comportamento do Erro de Quantização nas diversas configurações**

4.2.3.2 **Aplicação da Rede de *Kohonen***

Com a definição da configuração da rede passamos para a etapa de aplicação do algoritmo treinamento.

A Tabela 4.12 apresenta o erro de quantização encontrado na aplicação da rede de *Kohonen* em cada cenário.

Tabela 4.12 – **Erro de Quantização**

Cenários	Erro de Quantização
I	0,318
II	0,167
III	0,928
IV	0,666

As Figuras 4.20 a 4.23 apresentam os resultados da aplicação da rede de *Kohonen* em cada um dos quatro cenários.

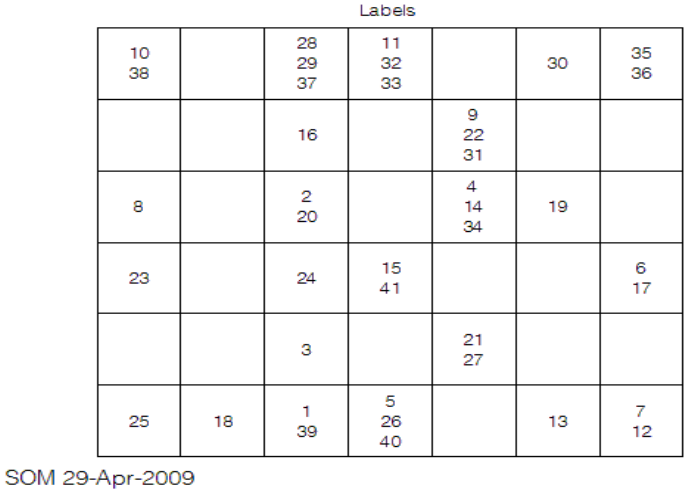


Figura 4.20 – **Rede treinada – Cenário I**

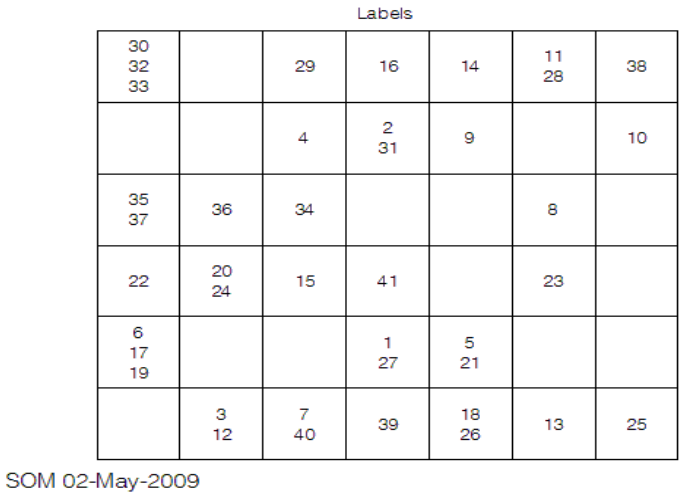


Figura 4.21 – **Rede treinada – Cenário II**

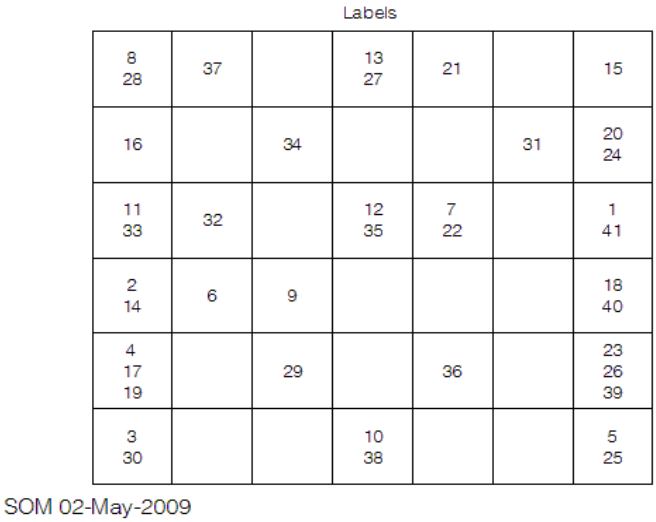


Figura 4.22 – **Rede treinada – Cenário III**

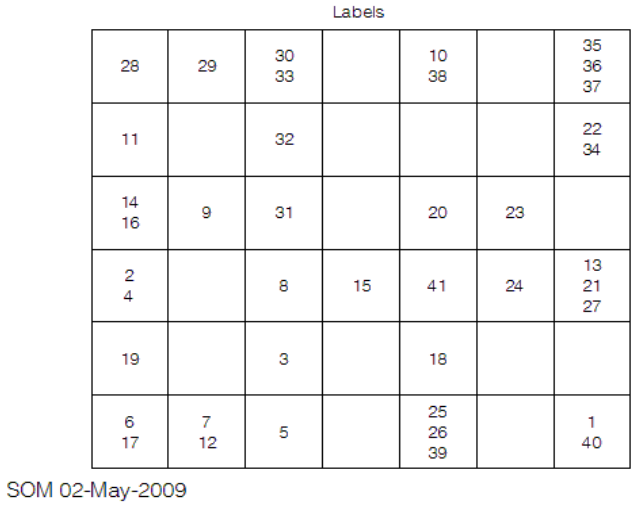


Figura 4.23 – **Rede treinada – Cenário IV**

Com o resultado obtido do treinamento da rede neural de *Kohonen* em cada um dos cenários, foram executadas, para cada valor de *k* (2,3,4 e 6), 10 simulações do algoritmo *K-Means*. Com os agrupamentos obtidos foram calculados os índices de validação e obtida a média e o desvio padrão. A Tabela 4.13 apresenta os valores obtidos de cada índice para cada valor de *k*.

O índice de *Silhouette* obteve o melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 2 grupos, conforme Figura 4.24. O comportamento deste índice é semelhante nos cenários I, II e IV. No cenário III os valores deste índice, para os números de grupos 2, 3 e 4, diminuem em relação aos valores obtidos no cenário II enquanto que para o número de grupos 6, o valor do índice *Silhouette* aumenta em relação ao obtido no cenário II.

O índice *Davies-Bouldin* obteve o melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 2 grupos, conforme Figura 4.25. Pode-se observar que o comportamento deste índice é o mesmo para todos os números de grupos até o cenário III. No cenário IV, para o número de grupos 6, o índice *Davies-Bouldin* aumentou em relação ao obtido no cenário III, nos demais números de grupos ocorreu uma diminuição em relação ao valor alcançado no cenário III.

O índice *Dunn* obteve o seu melhor resultado com os agrupamentos gerados no cenário II com a divisão das bacias hidrográficas em 2 grupos, conforme Figura 4.26. O comportamento deste índice foi semelhante nos números de grupos 4 e 6 em todos os cenários e nos grupos 2 e 3 até o cenário III.

Tabela 4.13 – **Índices de Validação dos Agrupamentos gerados pela Rede de *Kohonen***

Cenários	Nº de Grupos											
	Nº de Grupos = 2						Nº de Grupos = 3					
	Silhouette		*D. Bouldin*		*Dunn*		*Silhouette*		*D. Bouldin*		*Dunn*	
	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd
I	0,260	0	1,481	0	1,262	0	0,204	0,081	1,722	0,058	0,983	0,047
II	0,306	0,007	1,358	0	1,439	0	0,256	0,008	1,571	0	1,164	0
III	0,231	0,301	1,853	0,048	1,042	0,045	0,172	0,013	1,805	0,082	0,972	0,038
IV	0,221	0	1,407	0	1,412	0	0,147	0,039	1,614	0,041	0,951	0,092
Cenários	Nº de Grupos											
	Nº de Grupos = 4						Nº de Grupos = 6					
	Silhouette		*D. Bouldin*		*Dunn*		*Silhouette*		*D. Bouldin*		*Dunn*	
	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd	Média	DsvPd
I	0,171	0,022	1,661	0,030	1,036	0,052	0,161	0,022	1,630	0,117	0,751	0,055
II	0,223	0,034	1,526	0,107	0,881	0,020	0,193	0,075	1,373	0,041	0,566	0,137
III	0,216	0,017	1,661	0,049	1,032	0,050	0,220	0,025	1,529	0,088	0,912	0,113
IV	0,120	0,007	1,600	0,020	0,885	0,039	0,111	0,014	1,625	0,069	0,761	0,050

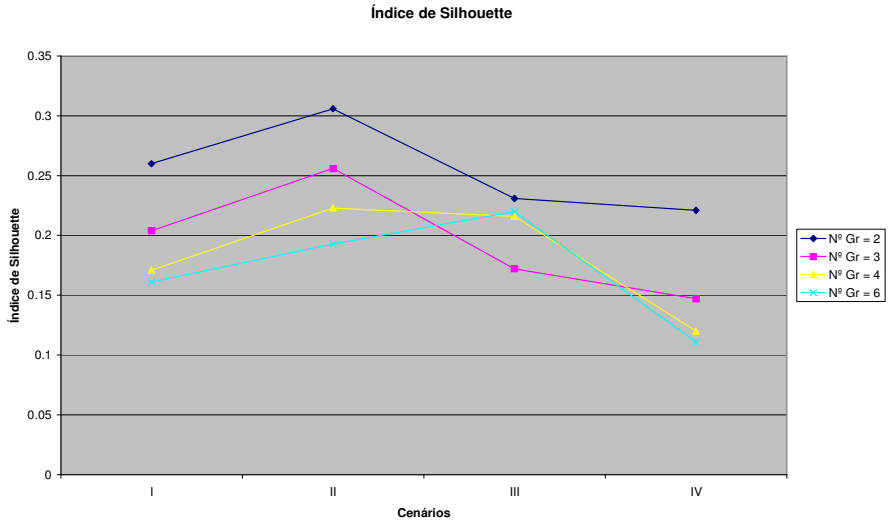


Figura 4.24 – **Comportamento do Índice de *Silhouette* – Rede de *Kohonen*.**

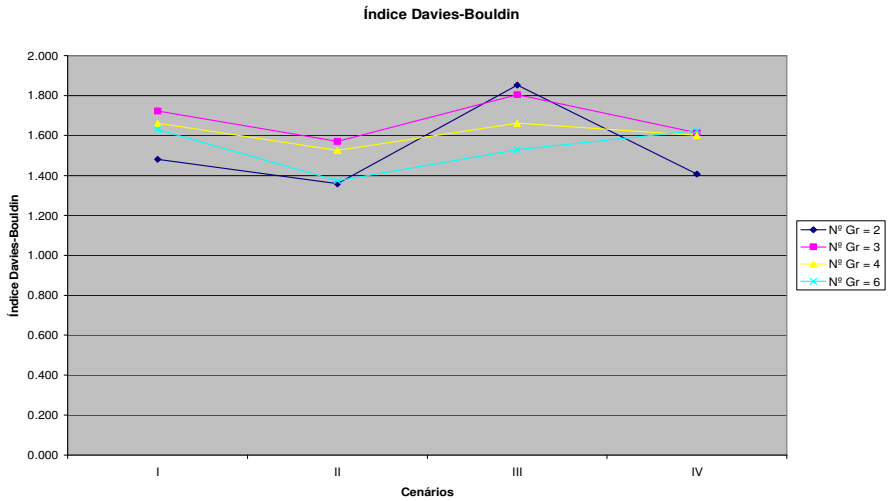


Figura 4.25 – **Comportamento do Índice *Davies-Bouldin* – Rede de *Kohonen*.**

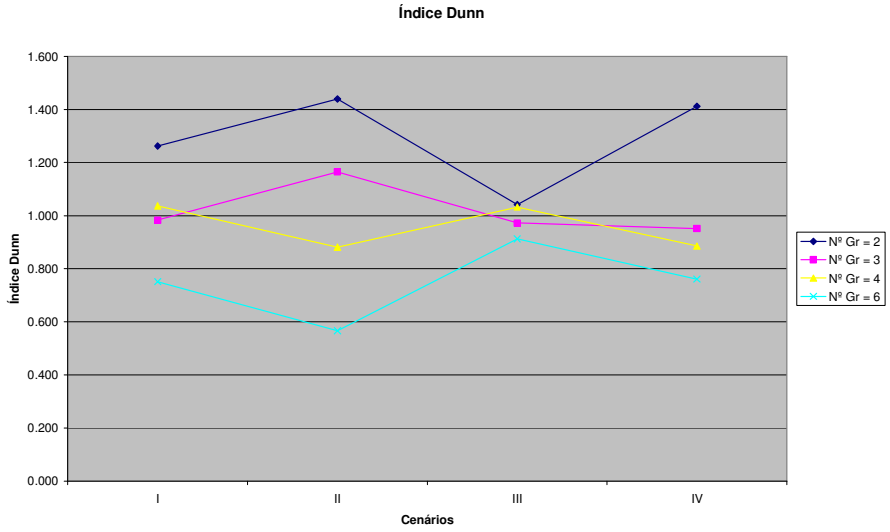


Figura 4.26 – Comportamento do Índice *Dunn* – Rede de *Kohonen*.

A pouca quantidade de dados obtidos para este trabalho, dados de apenas 41 bacias hidrográficas, não permitiram a formação de grupos de dados de treinamento e de avaliação dos resultados na aplicação da rede neural de *Kohonen*, limitando assim sua aplicação e possivelmente a geração de melhores resultados. Segundo Haykin [Haykin 2004], todos os neurônios de uma rede devem, portanto, ser expostos a um número suficiente de diferentes padrões de entradas para assegurar que o processo de auto-organização tenha uma chance de amadurecer apropriadamente.

4.3 ANÁLISE COMPARATIVA DOS ALGORITMOS

Nesta seção foi realizada uma análise comparativa dos resultados apresentados pelos três tipo de algoritmos (hierárquico, particional e baseado em modelos).

Em um primeiro momento, foi analizado o comportamento dos agrupamentos obtidos nos cenários I, II e III em relação aos agrupamentos obtidos com o conjunto completo de atributos (cenário IV). Em um segundo momento, foram analisados os resultados com relação aos cenários escolhidos para aplicação dos algoritmos e, em um terceiro momento, foram analisados os resultados com relação ao algoritmo e ao número grupos utilizados na divisão das 41 bacias hidrográficas.

4.3.1 Cenário IV x Cenários I, II e III

Para analisar o comportamento dos agrupamentos gerados nos três cenários (I, II e III) em relação aos agrupamentos obtidos com o conjunto completo de atributos foi aplicado o índice externo *rand* corrigido.

O índice externo avalia o agrupamento gerado baseado em uma estrutura pré-especificada, indicando se houve uma mudança na estrutura dos grupos com relação à estrutura previamente estabelecida.

Foi tomada como estrutura pré-especificada o agrupamento obtido com dados do cenário IV, composto pelo conjunto completo de atributos. A partir daí foram escolhidas algumas configurações de número de grupos e algoritmos nos demais cenários. Foi calculado o índice *rand* corrigido e os valores estão dispostos na Tabela 4.14. Pode-se verificar que os valores obtidos foram bem pequenos, demonstrando assim uma não concordância entre as partições obtidas no cenários IV e as partições obtidas nos demais cenários (I, II e III).

Tabela 4.14 – **Valores do índice externo *rand* corrigido**

Ward	Nº de Grupos	Correct Rand
Cenário I	6	0,176
Cenário II	6	0,267
Cenário III	4	0,036
K-Means	Nº de Grupos	Correct Rand
Cenário I	2	0,004
Cenário II	4	0,229
Cenário III	3	0,225
Kohonen	Nº de Grupos	Correct Rand
Cenário I	2	0,299
Cenário II	4	0,096
Cenário III	3	0,197

De acordo com a tabela acima, conclui-se que a redução do número de atributos mudou a estruturação dos grupos e que pelo os resultados apresentados na validação estatística (índices internos), a utilização de técnicas de seleção de atributos forneceu resultados melhores quando comparados com aos obtidos com o conjunto original de atributos.

4.3.2 **Cenários**

As Figuras 4.27, 4.28 e 4.29 apresentam o comportamento dos índices de validação nos agrupamentos gerados por cada algoritmo e em cada um dos quatro cenários.

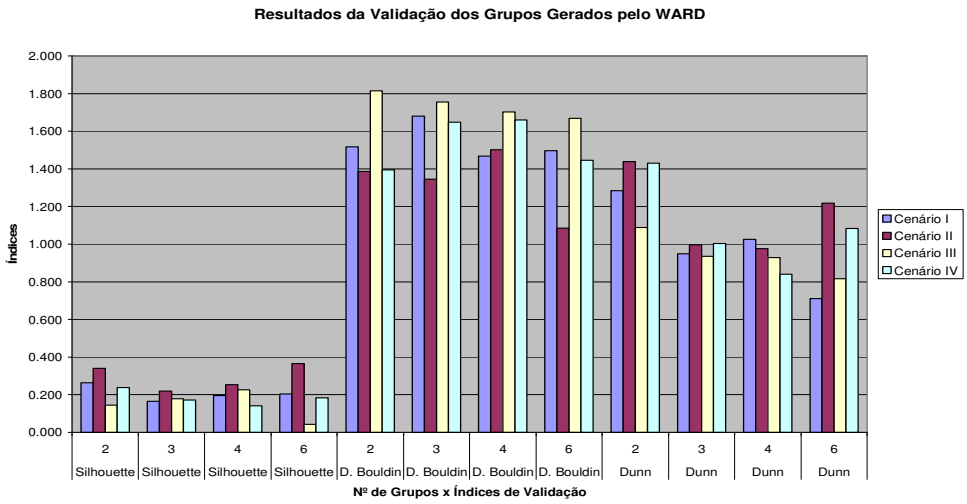


Figura 4.27 – **Resultados da validação dos grupos gerados pelo *Ward*.**

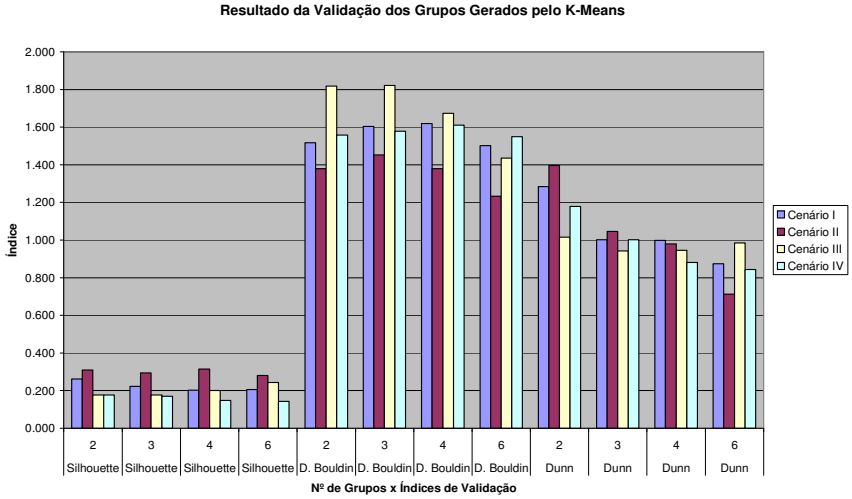


Figura 4.28 – **Resultados da validação dos grupos gerados pelo *K-Means*.**

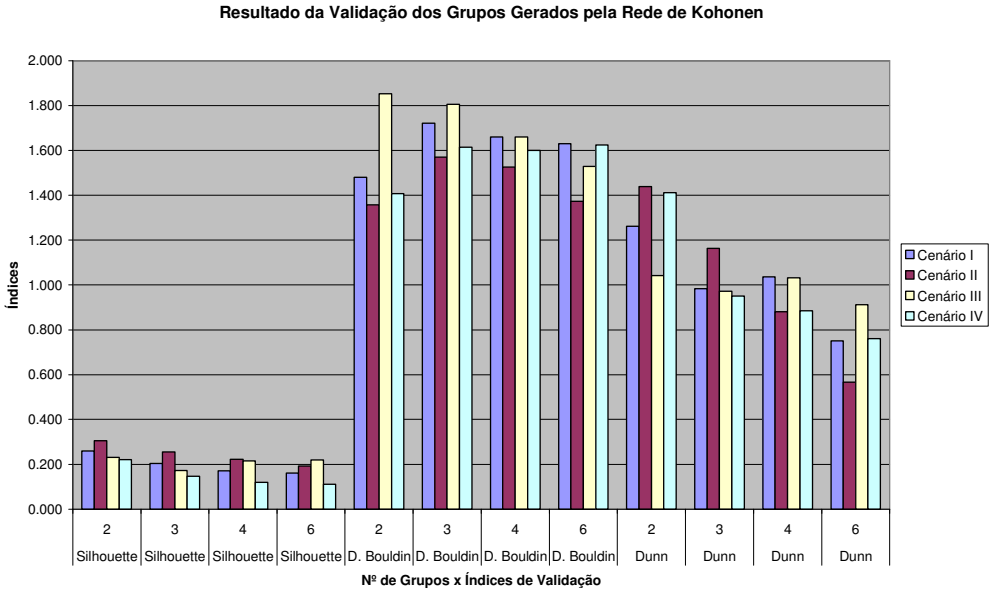


Figura 4.29 – **Resultados da validação dos grupos gerados pela Rede de *Kohonen*.**

Como se pode observar nas figuras acima, os melhores resultados dos índices de validação (*Silhouette*, *Davies-Bouldin* e *Dunn*) foram obtidos no contexto do cenário II para todos os algoritmos invariavelmente. As características que compõem o cenário II (A , Ke, Ra, IR e Imax), são representadas por grandezas que identificam a magnitude e a declividade da bacia hidrográfica e sua respectiva rede de drenagem, formando o conjunto de características capazes de permitir uma boa divisão de grupos homogêneos de bacias hidrográficas. Desta maneira, os futuros estudos de regionalização hidrológica devem adotar tais características na composição das variáveis explicativas em modelos matemáticos ou estatísticos. É importante ressaltar também que as medidas lineares e de declividade, que compõem esse cenário, apresentam a vantagem de serem facilmente determinadas a partir de mapas publicados para a maioria das regiões brasileiras com aceitável grau de confiabilidade.

4.3.3 **Algoritmos e Número de Grupos**

Os resultados obtidos pela validação dos agrupamentos gerados, pelos algoritmos nas quatro configurações do número de grupos utilizados, podem ser vistos nas Figura 4.30, 4.31 e 4.32. A Tabela 4.15 apresenta a classificação dos índices, indicando o algoritmo e o número de grupos utilizado. Como já foi mencionado anteriormente, os índices de *Silhouette* e *Dunn* sinalizam um bom resultado quando o seu valor é maximizado e o índice de *Davies-Bouldin* sinaliza um bom resultado quando o seu valor é minimizado. Os resultados analisados foram obtidos no contexto do cenário II que, como explicado na seção anterior, se destacou na qualidade dos agrupamentos gerados.

Tabela 4.15 – **Resultado da validação dos agrupamentos obtidos pelos algoritmos *Ward*, *K-Means* e Rede de *Kohonen***

Posição	Índice de Validação								
	Silhouette			*Davies-Bouldin*			*Dunn*		
1º	**6**	***Ward***	**0,365**	**6**	***Ward***	**1,085**	**2**	***Ward***	**1,439**
2º	2	*Ward*	0,340	6	*K-means*	1,233	2	*Kohonen*	1,439
3º	4	*K-means*	0,315	3	*Ward*	1,345	2	*K-means*	1,396
4º	2	*K-means*	0,309	2	*Kohonen*	1,358	6	*Ward*	1,218
5º	2	*Kohonen*	0,306	6	*Kohonen*	1,373	3	*Kohonen*	1,164
6º	3	*K-means*	0,295	2	*K-means*	1,380	3	*K-means*	1,046
7º	6	*K-means*	0,281	4	*K-means*	1,380	3	*Ward*	0,997
8º	3	*Kohonen*	0,256	2	*Ward*	1,387	4	*K-means*	0,980
9º	4	*Ward*	0,253	3	*K-means*	1,453	4	*Ward*	0,976
10º	4	*Kohonen*	0,223	4	*Ward*	1,502	4	*Kohonen*	0,881
11º	3	*Ward*	0,219	4	*Kohonen*	1,526	6	*K-means*	0,713
12º	6	*Kohonen*	0,193	3	*Kohonen*	1,571	6	*Kohonen*	0,566

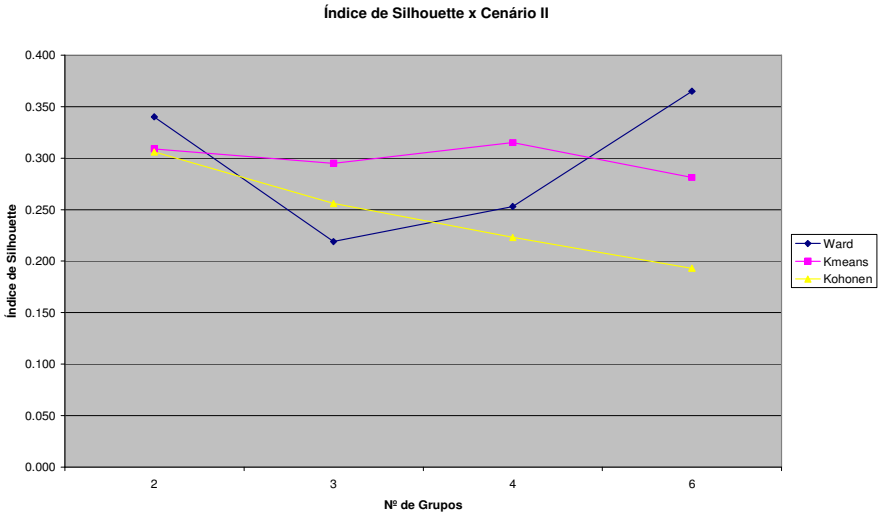


Figura 4.30 – **Comportamento do Índice de *Silhouette* nos resultados obtidos pelos algoritmos de clusterização.**

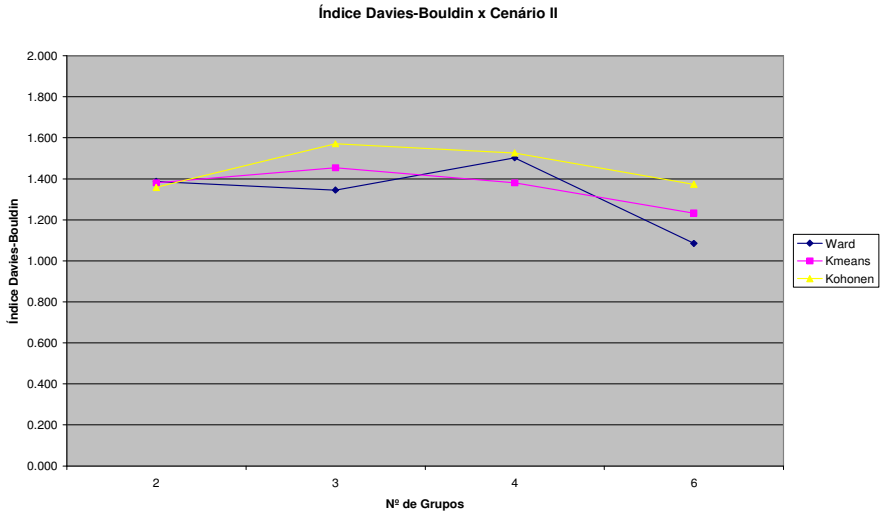


Figura 4.31 – **Comportamento do Índice de *Davies-Bouldin* nos resultados obtidos pelos algoritmos de clusterização.**

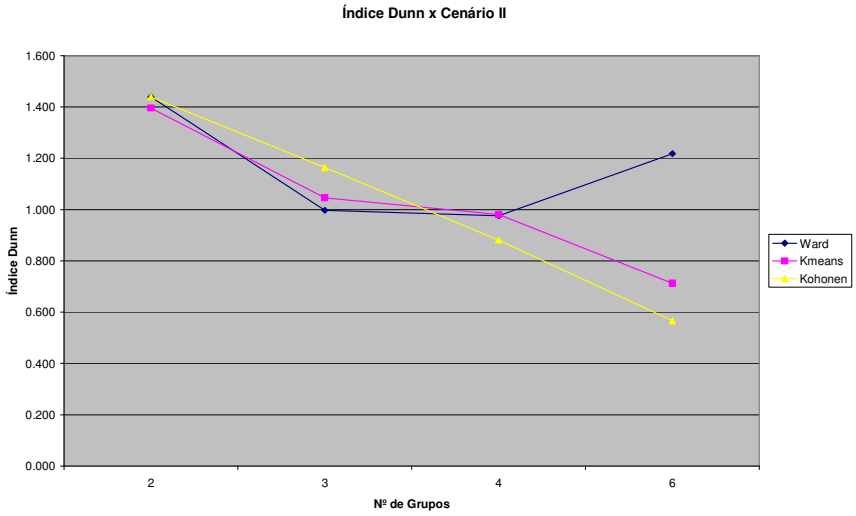


Figura 4.32 – **Comportamento do Índice *Dunn* nos resultados obtidos pelos algoritmos de clusterização.**

Observando os números apresentados pode-se concluir que o algoritmo *Ward* se destacou na tarefa de clusterização das 41 bacias hidrográficas. O mesmo obteve os melhores resultados nos três índices de validação utilizados. Com relação ao número de grupos, tivemos o número de grupos igual a 6 obtendo os melhores resultados em dois dos índices de validação (*Silhouette* e *Davies-Bouldin*) .

A Tabela 4.16 apresenta a composição dos grupos obtidos com os dados do cenário II e a divisão das bacias em 6 grupos. Com os dados apresentados na Tabela 4.16 percebe-se qual(is) característica(s) influenciou(aram) na formação dos grupos.

Tabela 4.16 - **Divisão das bacias hidrográficas em 6 grupos**

Bacias Hidrográficas		**Características das bacias idrográficas**					**Grupos**	**Característica determinante**
		A	Ke	Ra	IR	Imax		
13	Gamela	12,8	0,645	0,400	193,36	85,106	1	Imax
1	Albino	9,5	0,639	0,400	52,31	0,408	2	Ke
3	Bartolomeu I	59,5	0,584	0,189	342,98	5,799	2	
5	Chupadouro	17,8	0,619	0,406	127,73	9,836	2	
7	Emas	35,0	0,800	0,165	354,08	0,215	2	
12	Frutuoso II	19,1	0,670	0,321	483,69	0,329	2	
18	Tamanduá I	23,4	0,747	0,423	69,99	18,975	2	
21	Campos	181,2	0,568	0,320	224,61	15,306	2	
25	Livramento	37,0	0,898	0,996	18,77	3,788	2	
26	Namorados	14,2	0,703	0,203	87,52	17,857	2	
27	Soledade	313,1	0,606	0,171	46,57	2,871	2	
39	Umburana/Sumé	10,4	0,779	0,243	64,99	0,269	2	
40	Jatobá/Sumé	26,8	0,659	0,242	125,05	0,136	2	
30	Piancó	4.710,0	0,404	0,147	1.409,30	0,026	3	A,Ra, IR, Imax
32	Patos	1.850,0	0,291	0,172	1.072,18	0,034	3	
33	Serra Negra Norte	3.330,0	0,260	0,164	1.248,92	0,030	3	
2	Arrojado	30,1	0,367	0,653	522,69	6,791	4	Ke, IR
4	Cach. dos Alves	110,5	0,420	0,500	552,56	0,197	4	
8	Engo Arcoverde	126,9	0,338	0,932	344,62	0,136	4	
9	Engo Ávidos	1.009,5	0,406	0,923	688,77	0,047	4	
11	Farinha	747,9	0,246	1,799	746,41	4,351	4	
14	Jatobá I	94,0	0,415	0,862	944,40	28,074	4	
16	Riac dos Cavalos	161,4	0,271	0,724	644,27	0,075	4	
28	Antenor Navarro	1.720,0	0,227	1,164	715,74	0,054	4	
29	Aparecida	3.720,0	0,315	0,650	784,18	0,045	4	
31	Emas	530,0	0,419	0,688	636,00	0,069	4	
10	Epitácio Pessoa	10.659,0	0,283	0,488	499,08	2,794	5	A, Ke
38	Guarita	17.220,0	0,282	1,320	705,70	0,055	5	
6	Cochos	56,5	0,456	0,077	452,20	0,150	6	Ke, Ra
15	Queimadas	124,3	0,495	0,336	269,57	0,142	6	
17	Serra Vermelha	55,7	0,495	0,070	438,69	0,076	6	
19	Vazantes	137,0	0,457	0,197	595,93	0,044	6	

Continua na próxima página...

Tabela 4.16 - **Divisão das bacias hidrográficas em 6 grupos (Continuação)**

Bacias Hidrográficas		Características das bacias idrográficas					Grupos	Característica determinante
		A	Ke	Ra	IR	Imax		
20	Camalau	1.054,0	0,387	0,305	300,37	5,300	6	Ke, Ra
22	Cordeiro	1.665,8	0,407	0,217	428,27	7,219	6	
23	Santo Antonio	340,6	0,493	0,497	300,35	5,099	6	
24	São Domingos	65,5	0,397	0,157	102,66	10,060	6	
34	Faz. Alagamar	2.270,0	0,454	0,412	401,19	0,053	6	
35	Pedro Velho	3.590,0	0,370	0,105	512,33	0,113	6	
36	Caraúbas	5.120,0	0,441	0,286	419,21	0,018	6	
37	Poço de Pedras	3,140,0	0,339	0,105	439,79	0,018	6	
41	Gangorra/Sumé	137,4	0,512	0,375	172,64	0,056	6	

A Figura 4.33 apresenta, de forma visual, a composição dos grupos obtidos com os dados do cenário II e a divisão das bacias em 6 grupos.

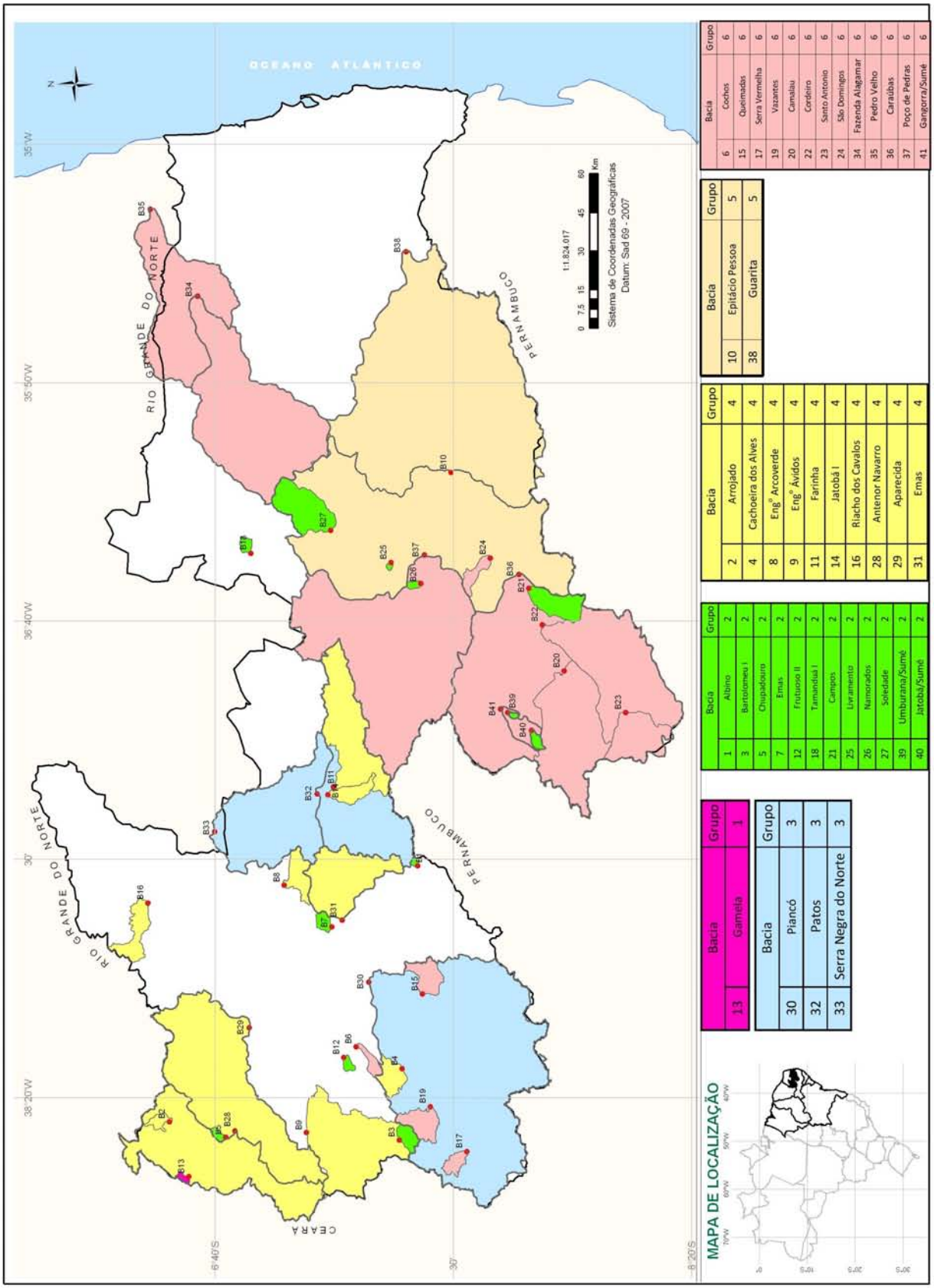


Figura 4.33 – **Divisão das 41 bacias hidrográficas em 6 grupos**

4.4 **CONCLUSÃO**

Neste capítulo foram apresentados os resultados obtidos com a aplicação dos algoritmos hierárquicos, Particional e Rede Neural de *Kohonen*. Com os agrupamentos obtidos pelos algoritmos *Ward*, *K-Means* e Rede de *Kohonen* foram calculados os índices de validação. Após uma análise em separado dos resultados de cada algoritmo, foi feita uma análise comparativa dos algoritmos acima citados enfocando os cenários utilizados e o número de grupos. No próximo capítulo será apresentada a conclusão da pesquisa bem como as suas contribuições para o estudo de regionalização hidrológica.

CAPÍTULO V

CONCLUSÃO

Neste trabalho apresentamos uma metodologia com base na utilização da análise de agrupamento (Clusterização) para identificação de áreas hidrologicamente homogêneas no Estado da Paraíba. No desenvolvimento da pesquisa, além do que foi proposto neste trabalho, algumas contribuições foram acrescentadas e que juntamente com as conclusões serão dispostas a seguir:

5.1 Conclusões

1. Bacias hidrográficas

A pouca quantidade de dados obtidos para este trabalho, dados de apenas 41 bacias hidrográficas, se deu pela dificuldade de obter as informações necessárias haja vista requerer enorme esforço de cálculo, interpretação de mapas e a baixa qualidade dos dados das estações fluviométricas em operação no Estado.

Uma possibilidade para obtenção de mais dados referentes às características físicas das bacias hidrográficas será a automação do processo de obtenção dos mesmos por meio de algoritmos que utilizem técnicas de geoprocessamento e linguagem de programação avançada, além da recuperação e implantação de novas estações fluviométricas.

Até o momento não se tem conhecimento de estudos e até mesmo de informações oficiais com relação à definição de agrupamentos de bacias hidrográficas similares no Estado da Paraíba. Essa ausência de informação não permitiu comparar e validar os resultados obtidos neste trabalho.

2. Seleção de Atributos

Pela análise dos resultados obtidos nos quatro cenários definidos neste trabalho conclui-se que a etapa de pré-processamento, especificamente a seleção de atributos, é de fundamental importância na etapa de mineração de dados. Isto pôde ser constatado no cenário IV, composto pelos 32 atributos das bacias hidrográficas, cujos resultados se mostraram inferiores aos obtidos nos cenários I e II, compostos por um menor número de características resultantes do processo de seleção de atributos.

Como citado anteriormente, a etapa de pré-processamento possui fundamental relevância no processo de descoberta de conhecimento porém é negligenciada por muitos devido ao excesso de trabalho manual, mas é nesta etapa que se garante a qualidade dos dados assegurando uma maior fidelidade aos resultados obtidos pelos algoritmos de mineração de dados.

Os algoritmos aplicados obtiveram os melhores resultados com os dados do cenário II. As características que compõem o cenário II (A , Ke, Ra, IR e Imax), se mostraram capazes de permitir uma boa divisão de grupos homogêneos, de forma que futuros estudos de regionalização hidrológica devem adotar tais características na composição das variáveis explicativas em modelos matemáticos ou estatísticos.

Com relação aos componentes principais, largamente utilizado em aplicações de clusterização [Demirel et. al 2007] [Júnior et. al 2006] [Llanillo et. al 2006], o seu uso não se mostrou eficiente do ponto de vista de prover uma boa classificação, para os dados das bacias hidrográficas levantadas neste trabalho.

3. Metodologia

A sequência metodológica proposta neste trabalho pode ser usada em outras regiões sem perdas da confiabilidade dos resultados, uma vez que mudariam-se apenas a base de dados das bacias hidrográficas e a consequente definição das regiões hidrologicamente homogêneas.

Vale destacar a introdução de índices de validação estatística nos agrupamentos gerados normalmente ausentes em trabalhos na área de engenharia no Brasil, reduzindo o empirismo que tem caracterizado as análises e aplicações feitas em estudos afins.

O aprendizado não supervisionado é complexo no sentido de não se ter um conhecimento a priori dos dados e com isso a validação estatística se torna uma ferramenta importantíssima na avaliação dos resultados obtidos.

4. Algoritmos de Clusterização

Não existe um algoritmo específico que seja apropriado a todos os tipos de dados e adoção de um ou outro depende de uma análise aprofundada como a aqui apresentada. Entretanto podemos destacar, para os dados utilizados, o algoritmo *Ward* como aquele que apresentou melhores resultados no contexto das bacias hidrográficas estudadas, confirmando supremacia no conjunto dos algoritmos hierárquicos já discorridos na literatura especializada.

Os agrupamentos gerados pelo algoritmo *Ward* apresentaram melhor desempenho na validação estatística em relação aos algoritmos *K-Means* e Rede Neural de *Kohonen*. Tomando o melhor valor obtido pelos índices, nos algoritmos *Ward*, *K-Means* e Rede Neural de *Kohonen* respectivamente, tem-se os seguintes valores para o índice de *Silhouette*, 0.365, 0.315 e 0.306, o índice *Davies-Bouldin*, 1.085, 1.233 e 1.345 e o índice *Dunn*, 1.439, 1.396 e 1.439.

Os algoritmos *Single-Linkage* e *Complete-Linkage* demonstraram uma não adequação a tarefa de Clusterização a que se propõe este trabalho.

Os resultados da validação estatística apresentados pela Rede Neural de *Kohonen* foram inferiores aos apresentados pelos algoritmos *Ward* e *K-Means*. A pouca quantidade de dados utilizada na aplicação deste algoritmo, não permitiu uma exposição suficiente de dados de entrada para assegurar um melhor processo de auto-organização como recomenda a literatura, impossibilitando assim a geração de melhores resultados.

5. Trabalhos relacionados

Todos os trabalhos citados anteriormente aplicaram a técnica de clusterização para identificação de regiões hidrologicamente homogêneas.

Com relação aos dados utilizados, os trabalhos de Demirel et al. [Demirel 2007] e Júnior et al. [Júnior 2006] utilizaram apenas dados pluviométricos. O trabalho de Rao e Srinivas [RS 2006] utilizou 9 (nove) atributos fisicos e climatológicos e Porto et al. [Porto 2004] utilizou apenas 4 (quatro) atributos físicos das bacias hidrográficas. Em nenhum dos trabalhos citados foi aplicada alguma técnica de seleção de atributos. Nesta dissertação foram utilizados 32 (trinta e dois) atributos físicos e climatológicos e aplicadas técnicas de seleção de atributos. Com a seleção de atributos foram criados mais três cenários para a aplicação dos algoritmos, permitindo assim uma análise comparativa entre os resultados obtidos em cada cenário.

Com relação aos algoritmos de clusterização, dentre os trabalhos citados, a maioria aplicou apenas uma abordagem de algoritmo de clusterização ([Demirel 2007]; [Porto 2004]; [Júnior 2006]). Nesta dissertação, foram aplicados algoritmos hierárquicos, particional e rede neural de Kohonen com o objetivo de identificar qual deles adequa-se aos estudos de regionalização hidrológica no Estado da Paraíba.

Assim como o trabalho de Rao e Srinivas [RS 2006], esta dissertação aplicou a validação estatística para os agrupamentos obtidos permitindo assim comparar os diversos algoritmos de clusterização utilizados.

5.2 Contribuições

1. Os resultados obtidos neste trabalho se constituem uma referência metodológica em estudos de regionalização hidrológica, assim como indicativo para aplicações práticas em engenharia de recursos hidrícos no âmbito do Estado da Paraíba.
2. Aplicação de métodos de seleção de atributos com o objetivo de assegurar uma maior fidelidade aos resultados obtidos pelos algoritmos de mineração de dados.
3. A metodologia aqui proposta pode ser facilmente aplicada em outras regiões sem perdas da confiabilidade dos resultados, uma vez que para isto, só será necessário mudar apenas

a base de dados das bacias hidrográficas e a consequente definição das regiões hidrologicamente homogêneas.

4. Introdução de índices de validação estatística nos agrupamentos gerados normalmente ausentes em trabalhos na área de engenharia no Brasil, reduzindo o empirismo que tem caracterizado as análises e aplicações feitas em estudos afins.

5.3 **Trabalhos Futuros**

1. Ampliação da base de dados por meio de automatização dos trabalhos de obtenção das caracterísiticas físicas e climáticas das bacias hidrográficas.
2. Incorporar dados de outras sub-regiões do Nordeste Semi-árido visando aumentar a representatividade dos dados e avaliar a flexibilidade dos algoritmos.
3. Estudar a possibilidade do uso de caracterísiticas das bacias hidrográficas obtidas de imagens de satélite, cujas informações podem ser obtidas em tempo real.
4. A partir da metodologia desenvolvida neste trabalho projetar um sistema iteligente capaz de executar os algoritmos de clusterização, a validação estatística e sugerir opções que auxiliem o analista do domíno no processo de tomada de decisão.

BIBLIOGRAFIA

[ANA 2007] Rede Hidrometeorológica administrada pela ANA. Apresentação de slides realizada em junho de 2007, 89 p.

[ANA 2009] Agência Nacional de Águas. Sistema de Informações Hidrológicas. Disponível em **http://hidroweb.ana.gov.br/** . Acessado em 10/09/2008.

[BL 2004] M. J. A. Berry, G. Linoff. Data Mining Techniques For marketing, Sales and Customer Relationship Managemet, Second Edition. Wiley Publishing, Inc., 2004.

[Bogorny 2003] V. Bogorny. Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos – Trabalho Individual. UFRGS, Porto Alegre - RS, 2003.

[Bolshakova 2009] N. Bolshakova. Machaon clustering and validation environment. Disponível em: **https://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html**. Acessado em: 13/04/2009.

[Borges 2006] H. B. Borges. Redução de dimensionalidade em base de dados de expressão gênica. Dissertação de Mestrado, PUC, Curitiba – PR, 2006.

[Brasil 2006] Plano Nacional de Recursos Hídricos: Síntese Executiva. Brasília, 2006, p. 142. Disponível em: **http://pnrh.cnrh-srh.gov.br/ .** Acessado em 01/07/2009.

[Cataldi 2007] M. Cataldi, C. C. L. Achão, B. G. F. Machado, S. B. Silva, L.G. F. Guilhon. Aplicação das técnicas de Mineração de Dados como complemento às previsões estocásticas univariadas de vazão natural: estudo de caso para a bacia do rio Iguaçu. Revista Brasileira de Recursos Hídricos, v. 12, p. 83-92, 2007.

[Cortês 2002] S. C. Cortês, R. M. Porcaro, S. Lifschitz. Mineração de Dados – Funcionalidades, Técnicas e Abordagens. Acessado em 04/02/2008. **ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf**.

[DB 1979] D. L. Davies, D. W. Bouldin. A cluster separation measure. IEEE Transaction on Pattern Analysis and Machine Intelligence 1, 224-227.

[Demirel 2007] M. C. Demirel, A. J. Mariano, E. Kahya. Performing *k-means* analysis to drought principal components of Turkish Rivers. Hidrology Days 2007, **http://hydrologydays.colostate.edu/Proceedings_2007.htm.** Acessado em 20/01/2009.

[Diniz 2006] L. S. Diniz. Legislação de Saneamento e Recursos Hidrícos. Paraíba, 2006.

[Diniz 2008] L. S. Diniz. Regionalização de parâmetros de modelo chuva-vazão usando redes neurais. Tese de Doutorado, IPH/UFRGS – RS, 2008.

[Espenchitt 2008] D. G. Espenchitt. Segmentação de dados em um número desconhecido de grupos utilizando algoritmo de colônia de formigas. Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro - RJ, 2008.

[Faceli 2005] K. Faceli, A. C. P. L. F. Carvalho, M. C. P. Souto. Validação de Algoritmos de Agrupamento. Relatórios Técnicos do ICMC, ISSN – 0103-2569. São Carlos – SP.

[Faria 2006] M. P. C. Faria. Análise de Crédito à Pequena Empresa – Um Modelo de Escoragem Baseado nas Metodologias Estatísticas : Análise Fatorial e Lógica Fuzzy. Dissertação de Mestrado. Faculdade de Economia e Finanças IBMEC. Rio de Janeiro - RJ, 2006.

[Fayyad 1996] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery in database. AI Magazine, p.37-54, 1996.

[Francisco 2004] C. A. C. Francisco.Rede de Kohonen: Uma ferramenta no estudo das relações tróficas entre espécies de peixes. Dissertação de Mestrado, UFPR, Curitiba – PR, 2004.

[GP 2005] R. Goldschmidt, E. Passos. Data Mining – Um Guia Prático. Editora Campus, 2º Tiragem, 2005.

[Hall 1999] M. A. Hall. Correlation-based Feature Selection For Machine Learning. Doctoral dissertation, The University of Waikato, Department of Computer Science. Hamilton, NewZealand, 1999.

[Haykin 2004] S. Haykin. Rede Neurais – Princípios e Práticas. 2ª Edição – Porto Alegre – Bookman. Reimpressão 2004.

[HK 2000] Jiawei Han; Micheline Kamber - Data Mining - Concepts and Techniques. Morgan Kaufmann Publishers, 2000.

[Júnior 2006] J. C. F. M. Júnior, G. C. Sediyama, P. A. Ferreira, B. G. Leal. Determinação de regiões homogêneas quanto à distribuição de frequência de chuvas no leste do Estado de Minas Gerais. Revista Brasileira de Engenharia Agrícola e Ambiental, v.10, n.2, p.408–416, 2006.

[Llanillo 2006] R. F. Llanillo, M. E. Del Grossi, F. O. Santos, P. D. Munhos, M. F. Guimarães. Regionalização da agricultura do Estado do Paraná, Brasil. Cienc. Rural vol.36 no.1 Santa Maria Jan./Feb. 2006.

[Larose 2005] D. T. Larose. Discovering Knowledge in Data – An Introduction to Data Mining. Wiley-Interscience, 2005.

[Menezes 2007] R. H. N. Menezes, R. T. Dantas, F. A. S. Sousa. Regiões pluviométricas homogêneas no Estado do Maranhão, Brasil. Revista Brasileira de Agrometeorologia, Piracicaba, v.15, n.2, p. 152-160, 2007.

[Metz 2006] J. Metz. Interpretação de Clustering gerados por algoritmos de clustering hierárquico. Dissertação de Mestrado, USP, São Carlos - SP, 2006.

[Meyer 2002] A. S. Meyer. Comparação de coeficientes de similaridade usados em análise de agrupamento com dados de marcadores moleculares dominantes. Dissertação de Mestrado, Escola Superior de Agricultura Luis de Queiroz, USP, Piracicaba – SP, 2002.

[Mitra 2002] P. Mitra, C. A. Murthy, S. K. Pal. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, nº 3, March 2002.

[MZ 2002] P. Moscato, F. J. Von Zuben. Um visão geral de clusterização de dados. **ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia368_02/topico5_02.pdf**. Acessado em 07/03/2008.

[Peralta 2003] A. S. Peralta. Análise de regionalização de vazão máxima para pequenas bacias hidrográficas. Dissertação de Mestrado, UNICAMP – SP, 2003.

[Porto 2004] M. M. Porto, E. M. Andrade, R. N. T. Costa, L. C. A. L. Filho, M. Meireles. Identificação de bacias hidrográficas com características físicas similares no Estado do Ceará. Revista Ciência Agronômica, vol.35, Nº 1, p.17-25.

[Rabus 2003] B. Rabus, M. Eineder, A. Roth, R. Bamler. The shuttle radar topography mission - A new class of digital elevation models acquired by spaceborne radar. Journal of Photogrammetry & Remote Sensing, v. 57, p. 241-262.

[RS 2006] A. R. Rao, V.V. Srinivas. Regionalization of watersheds by hybrid-cluster analysis. Journal of Hydrology 318, p. 37-56, 2006.

[Romão 2002] W. Romão. Descoberta de Conhecimento Relevante em Banco de Dados sobre Ciência e Tecnologia. Dissertação de Mestrado, UFSC – PR, 2002.

[Silva 2004] M. P. S. Silva – Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka – INPE. Disponível em **http://bibliotecadigital.sbc.org.br/download.php?paper=35**. Acessado em 05/09/2007.

[Souto 2009] M. Souto. Validação de Agrupamentos. Disponível em **http://www.dimap.ufrn.br/~marcilio/AM/validacao-cluster.ppt**. Acessado em 11/04/2009.

[Thomé 2002] A. C. G. Thomé. Redes Neurais – Uma ferramenta para KDD e Data Mining. Disponível em **http://equipe.nce.ufrj.br/thome/grad/nn/mat_didatico/apostila_kdd_mbi.pdf** . Acessado em 04/02/2008.

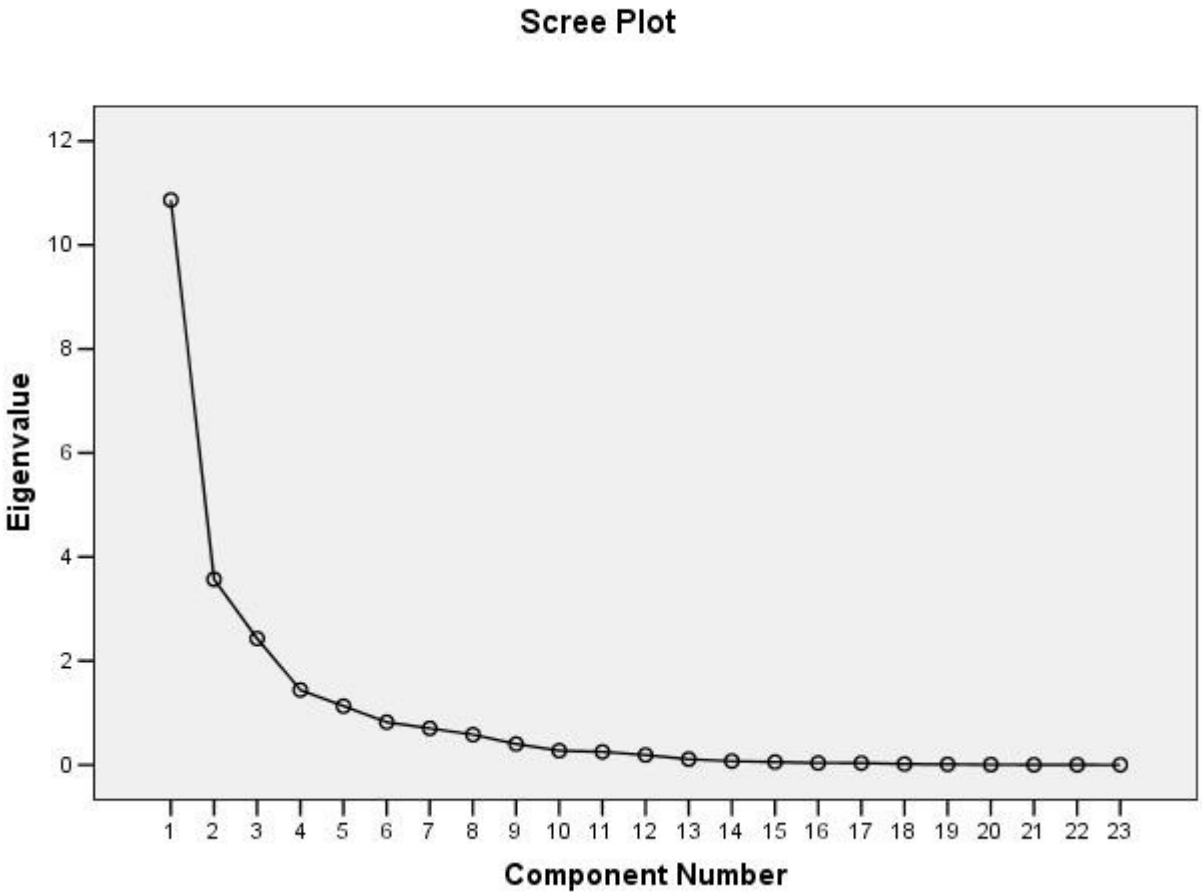
[Vesanto 2000] J. Vesanto. Using SOM in Data Mining. Licentiate´s Thesis, Helsink University of Technology. Finland, 2000. Disponível em **http://www.cis.hut.fi/~juuso/**.

[Ward 1963] J. H. Ward. Hierarchical Grouping to Optimize an Objetive Function. American Statistical Association Journal, March 1963.

[Zuchini 2003] M. H. Zuchini. Aplicações de Mapas Auto-Organizáveis em Mineração de Dados e Recuperação de Informação.Dissertação de Mestrado, Unicamp, Campinas – SP, 2003.

ANEXO A: ANÁLISE DE COMPONENTES PRINCIPAIS

Anexo A.1 – **Scree Plot**



Anexo A.2 – **Autovalores associados à matriz de correlação correspondentes as respectivas porcentagens de explicação da variação total.**

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.866	47.242	47.242	10.866	47.242	47.242	9.324	40.538	40.538
2	3.570	15.521	62.763	3.570	15.521	62.763	2.530	11.002	51.540
3	2.430	10.566	73.328	2.430	10.566	73.328	2.122	9.227	60.768
4	1.441	6.265	79.593	1.441	6.265	79.593	1.431	6.221	66.988
5	1.131	4.915	84.508	1.131	4.915	84.508	1.221	5.307	72.295
6	0.822	3.572	88.081	0.822	3.572	88.081	1.179	5.124	77.420
7	0.702	3.053	91.134	0.702	3.053	91.134	1.074	4.671	82.090
8	0.581	2.527	93.661	0.581	2.527	93.661	1.059	4.606	86.696
9	0.402	1.747	95.408	0.402	1.747	95.408	0.961	4.176	90.872
10	0.273	1.185	96.593	0.273	1.185	96.593	0.927	4.031	94.904
11	0.252	1.095	97.689	0.252	1.095	97.689	0.405	1.759	96.662
12	0.190	0.824	98.513	0.190	0.824	98.513	0.214	0.932	97.594
13	0.110	0.479	98.992	0.110	0.479	98.992	0.179	0.778	98.371
14	0.072	0.315	99.306	0.072	0.315	99.306	0.168	0.729	99.100
15	0.055	0.241	99.547	0.055	0.241	99.547	0.073	0.319	99.419
16	0.037	0.160	99.708	0.037	0.160	99.708	0.039	0.168	99.587
17	0.034	0.147	99.855	0.034	0.147	99.855	0.033	0.141	99.729
18	0.017	0.074	99.929	0.017	0.074	99.929	0.029	0.124	99.853
19	0.010	0.043	99.972	0.010	0.043	99.972	0.024	0.103	99.957
20	0.005	0.021	99.993	0.005	0.021	99.993	0.008	0.033	99.990
21	0.002	0.007	99.999	0.002	0.007	99.999	0.002	0.009	99.999
22	0.000	0.001	100.000	0.000	0.001	100.000	0.000	0.001	100.000
23	0.000	0.000	100.000	0.000	0.000	100.000	0.000	0.000	100.000

Extraction Method: Principal Component Analysis.

Anexo A.3 – **Matriz de Coeficientes dos Componentes Principais**

Rotated Component Matrix(a)

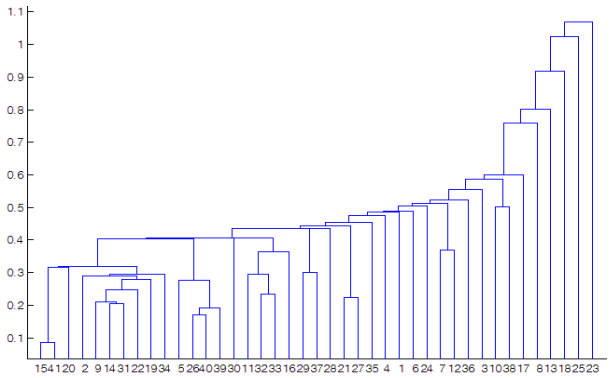
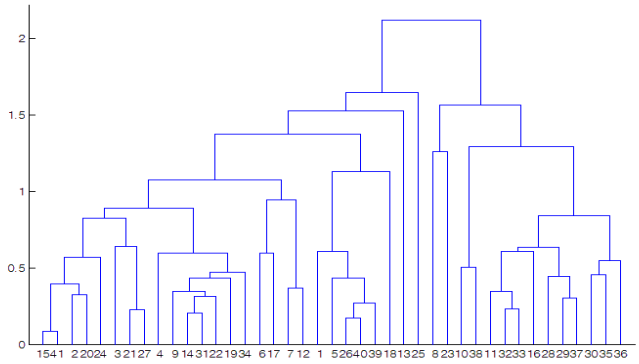
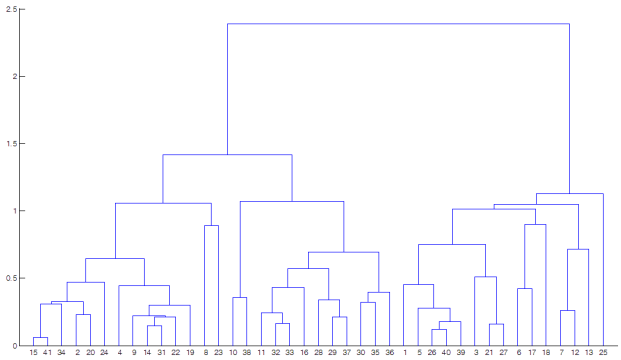
	Component																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A	0.976	0.056	0.005	-0.010	-0.042	0.015	0.001	0.023	0.001	-0.040	0.064	-0.055	-0.072	-0.023	0.011	-0.163	-0.005	-0.018	0.019	0.002	0.000	0.000	0.000
Ld	0.966	0.089	-0.006	0.040	0.082	0.002	0.013	0.029	-0.044	0.011	0.025	-0.027	-0.053	0.142	-0.002	-0.024	-0.008	-0.151	0.015	0.002	0.000	0.000	0.000
Pr	0.959	0.240	0.044	0.034	-0.033	-0.054	0.053	0.043	-0.073	-0.018	-0.014	-0.002	0.007	0.011	0.000	0.001	0.040	0.023	0.043	0.021	-0.021	-0.003	-0.001
Lr	0.956	0.252	0.039	0.028	-0.029	-0.049	0.057	0.038	-0.067	-0.019	0.002	-0.007	-0.001	0.013	0.002	-0.004	0.051	0.021	0.062	0.022	-0.023	-0.003	0.001
Le	0.945	0.171	0.049	-0.015	-0.019	-0.056	0.017	0.081	-0.021	-0.015	-0.007	0.206	0.091	-0.007	0.006	0.078	-0.018	0.016	0.014	-0.077	0.001	0.000	0.000
Ir	0.944	0.114	0.074	0.080	-0.065	-0.104	0.023	0.077	-0.120	-0.012	-0.146	0.039	0.076	0.000	-0.014	0.046	-0.042	0.035	-0.120	0.007	-0.002	0.000	0.000
Lm	0.935	0.173	0.017	0.002	-0.046	-0.043	0.050	0.105	-0.010	-0.029	-0.034	0.253	0.087	-0.013	0.001	0.032	-0.006	0.038	-0.031	0.026	0.021	0.004	0.000
Lt	0.917	0.213	0.092	0.101	-0.044	-0.100	0.020	0.021	-0.141	0.016	-0.096	-0.223	-0.024	-0.002	-0.007	0.032	-0.011	0.028	-0.010	-0.008	0.024	-0.006	0.000
L	0.913	0.223	0.084	0.108	-0.053	-0.100	0.021	0.021	-0.144	0.011	-0.099	-0.224	-0.017	-0.006	-0.008	0.034	-0.007	0.026	-0.013	0.000	0.008	0.012	0.000
Or	0.664	0.466	0.016	0.102	0.240	-0.222	0.087	0.135	0.005	0.074	-0.222	0.043	0.372	0.055	0.011	0.007	-0.001	0.004	-0.003	-0.001	0.000	0.000	0.000
Ke	-0.400	-0.886	0.019	-0.069	-0.092	0.070	-0.062	-0.086	-0.018	-0.068	0.074	-0.004	-0.014	0.028	0.006	0.003	0.112	0.008	0.027	0.003	0.000	0.000	0.000
Kc	0.431	0.878	-0.025	0.045	0.100	-0.073	0.100	0.027	0.001	0.002	0.011	-0.008	-0.010	0.029	0.001	0.005	0.116	0.009	0.027	0.003	0.000	0.000	0.000
IR	0.381	0.483	-0.144	0.436	0.421	-0.038	0.194	0.017	-0.050	0.196	-0.093	-0.002	0.056	0.374	-0.003	0.003	0.003	-0.004	0.000	0.000	0.000	0.000	0.000
SOLO 1	0.119	0.029	0.952	-0.005	-0.143	0.112	0.002	0.013	-0.085	-0.056	0.019	-0.015	-0.035	0.003	-0.183	0.004	-0.001	0.000	-0.003	0.000	0.000	0.000	0.000
L600 (mm)	0.086	-0.079	0.931	0.109	-0.052	0.208	-0.063	0.008	-0.089	-0.086	0.021	0.012	0.040	-0.019	0.198	-0.003	0.001	0.001	0.002	0.000	0.000	0.000	0.000
DS	0.163	0.104	0.109	0.931	0.061	0.151	0.068	-0.044	-0.044	0.215	0.049	-0.006	0.008	0.003	0.003	0.001	-0.001	0.000	-0.001	0.000	0.000	0.000	0.000
Dd	-0.169	0.184	-0.203	0.072	0.882	-0.081	0.000	-0.008	0.264	0.192	-0.003	0.000	0.012	0.003	0.001	0.000	0.000	-0.002	0.001	0.000	0.000	0.000	0.000
Imax	-0.183	-0.108	0.370	0.154	-0.080	0.884	-0.042	-0.010	-0.025	-0.024	0.047	0.001	-0.012	-0.001	0.001	-0.001	0.000	-0.001	0.001	0.000	0.000	0.000	0.000
Cmed	0.066	0.120	-0.046	0.069	0.012	-0.038	0.977	0.062	-0.056	0.100	-0.014	0.001	0.007	0.012	-0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ip	0.150	0.084	0.015	-0.047	-0.007	-0.019	0.063	0.972	-0.112	-0.073	-0.024	0.003	0.010	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AE/A	-0.274	0.007	-0.199	-0.046	0.274	-0.014	-0.083	-0.163	0.858	0.191	0.026	0.006	0.000	-0.006	0.000	-0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000
P (mm)	-0.047	0.081	-0.172	0.332	0.238	-0.012	0.155	-0.109	0.205	0.846	0.029	-0.002	0.007	0.017	-0.001	0.001	-0.001	0.000	0.000	0.000	0.000	0.000	0.000
IG	-0.357	-0.364	0.194	0.405	-0.036	0.455	-0.063	-0.117	0.092	0.115	0.535	0.013	-0.056	-0.030	0.001	-0.003	0.001	-0.001	0.002	0.000	0.000	0.000	0.000

Extraction Method: Principal Component Analysis.

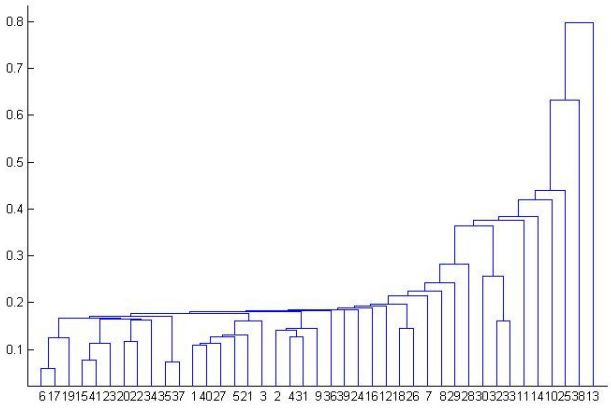
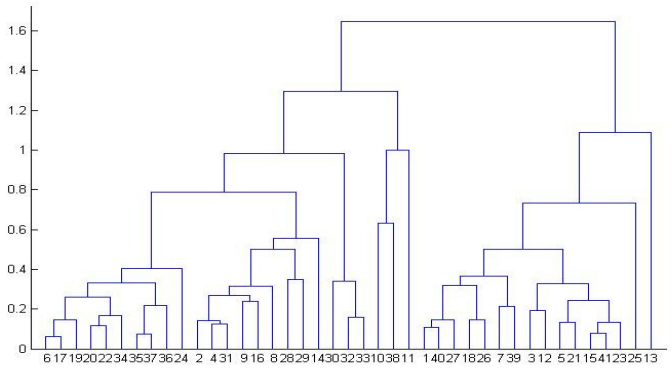
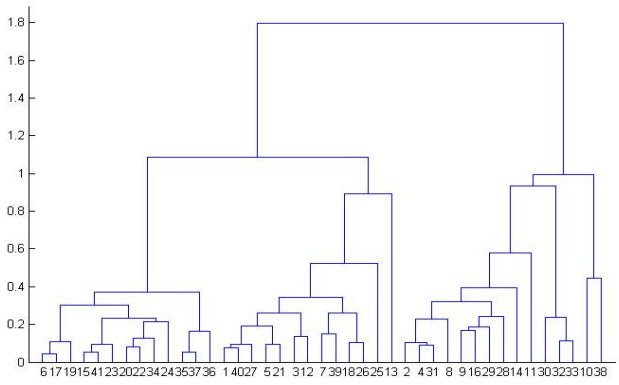
a. Rotation converged in 7 iterations.

ANEXO B: GRÁFICOS OBTIDOS NA APLICAÇÃO DOS ALGORITMOS DE CLUSTERIZAÇÃO

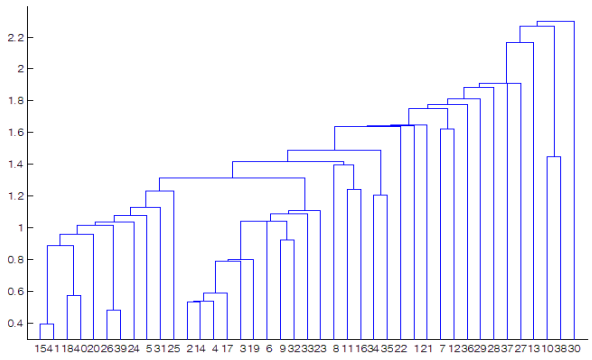
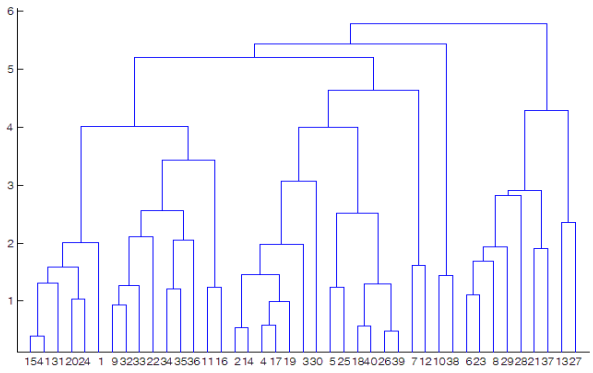
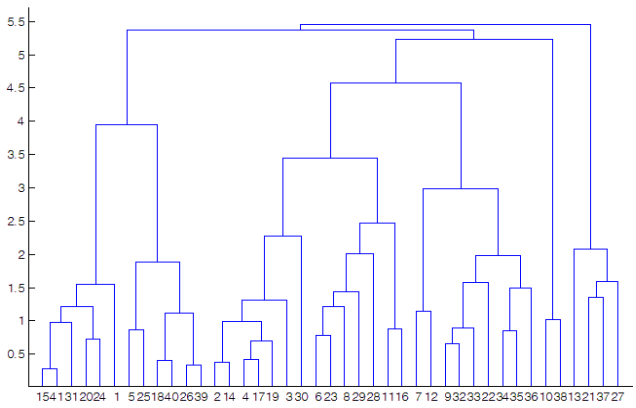
Anexo B.1 – **Dendogramas obtidos pelos algoritmos hierárquicos no Cenário I**

Algoritmo	Dendograma	Cenário
Single-Linkage		I
Complete-Linkage		
Ward		

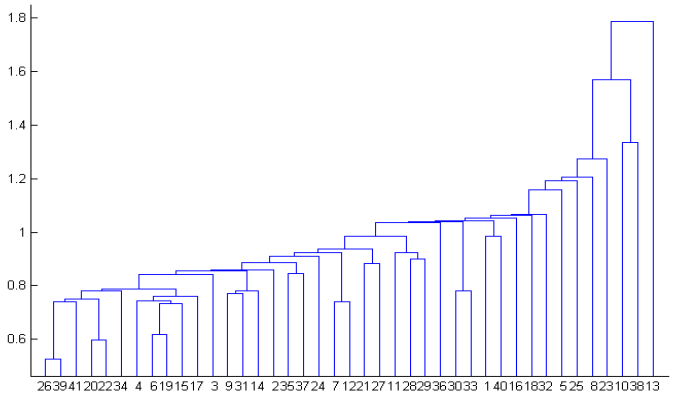
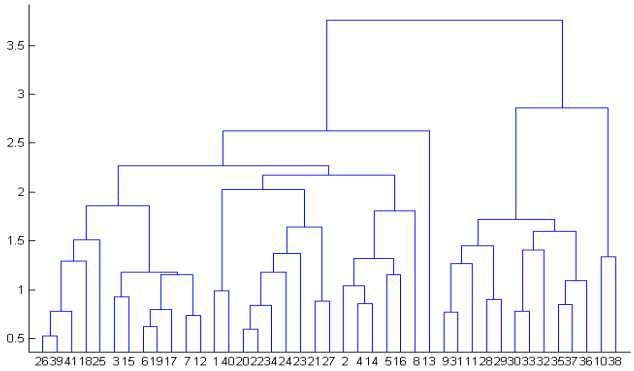
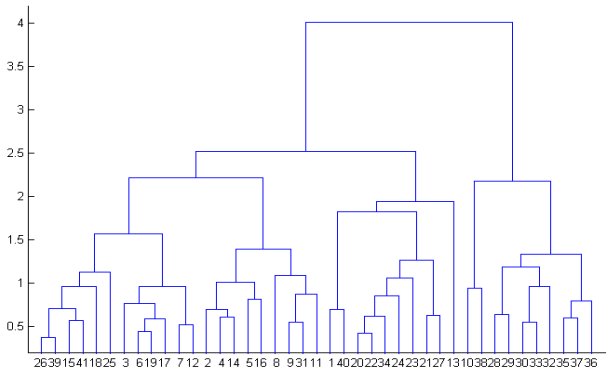
Anexo B.2 – **Dendogramas obtidos pelos algoritmos hierárquicos no Cenário II**

Algoritmo	Dendograma	Cenário
Single-Linkage		II
Complete-Linkage		
Ward		

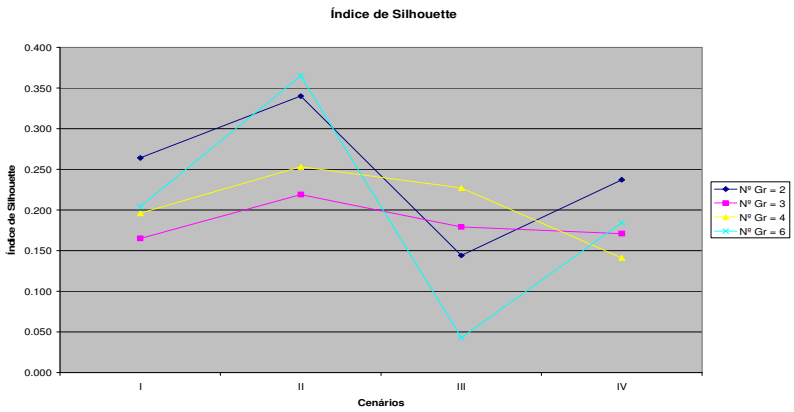
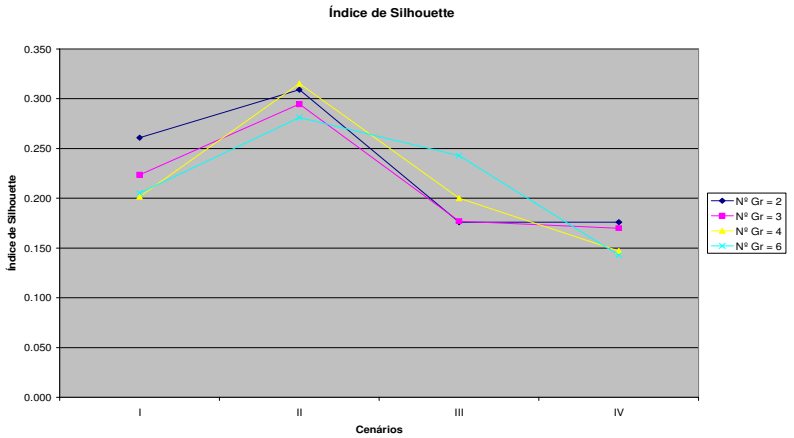
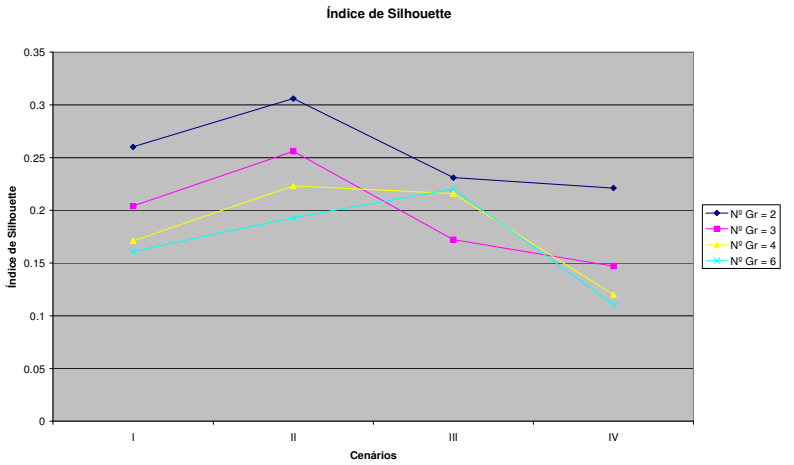
Anexo B.3 – **Dendogramas obtidos pelos algoritmos hierárquicos no Cenário III**

Algoritmo	Dendograma	Cenário
Single-Linkage		III
Complete-Linkage		
Ward		

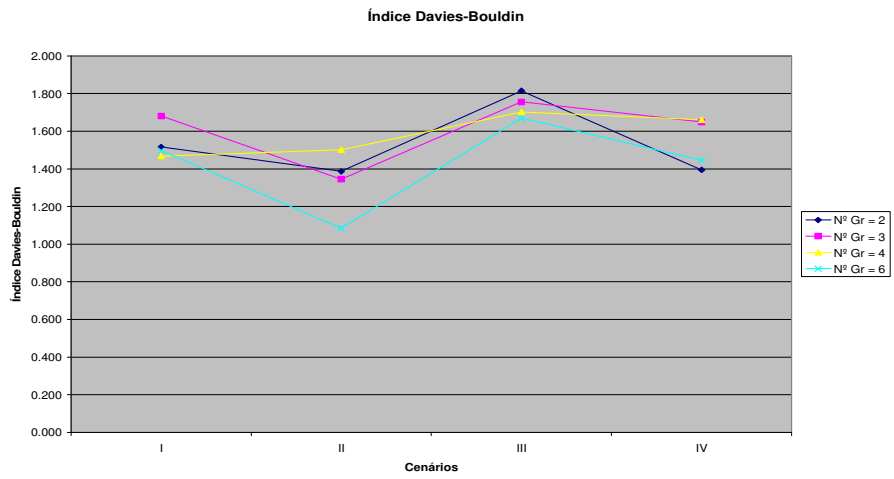
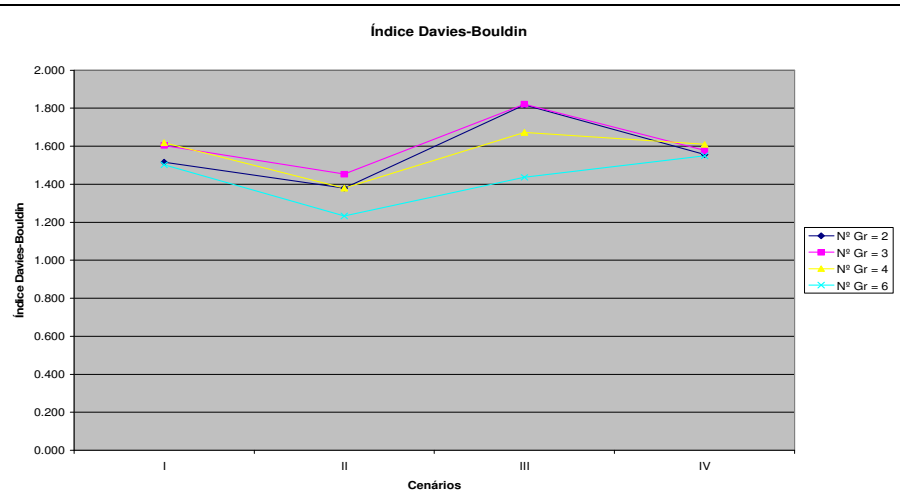
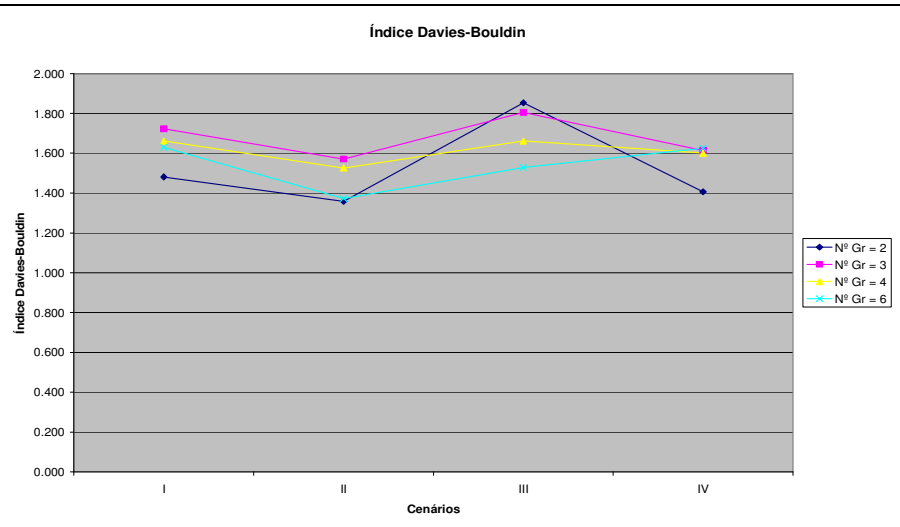
Anexo B.4 – **Dendogramas obtidos pelos algoritmos hierárquicos no Cenário IV**

Algoritmo	Dendograma	Cenário
Single-Linkage		IV
Complete-Linkage		
Ward		

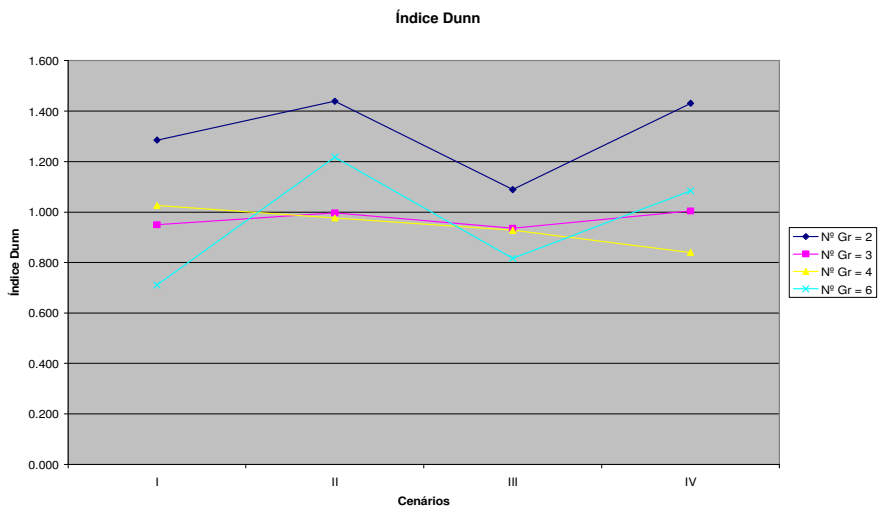
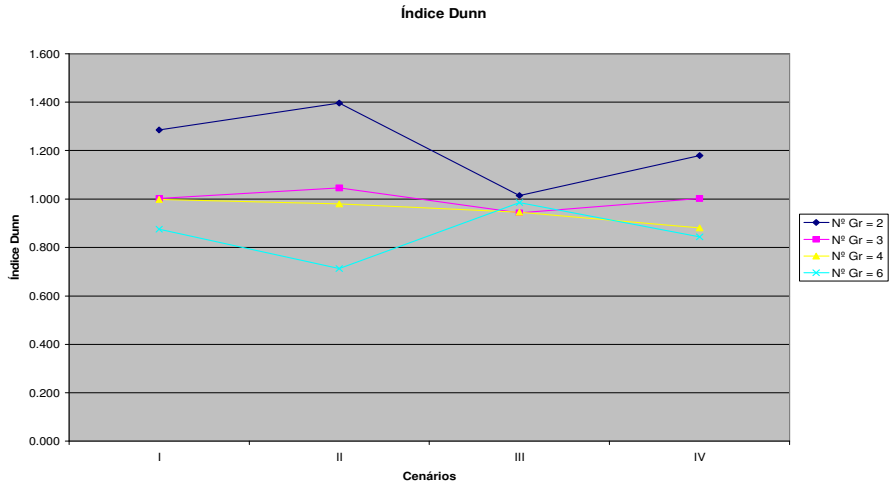
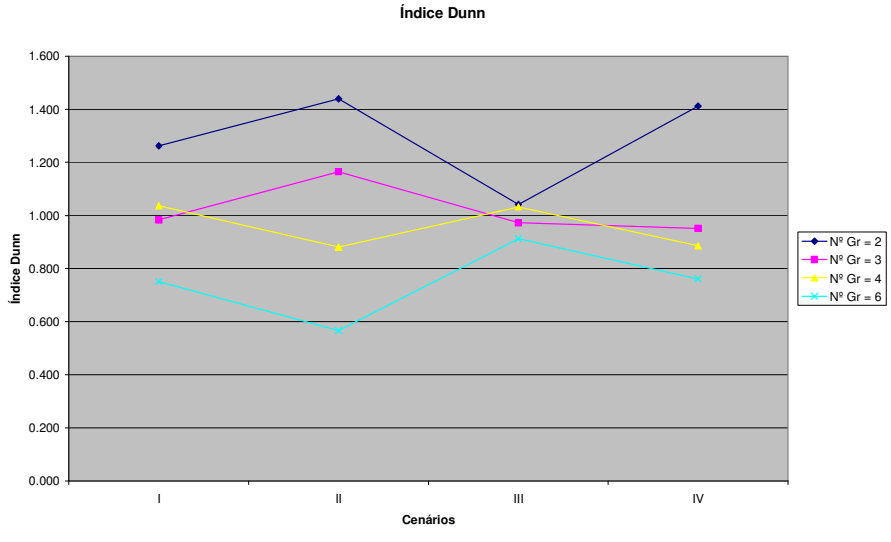
Anexo B.5 – **Comportamento do Índice de *Silhouette***

Algoritmo	Índice de *Silhouette*
Ward	
K-Means	
Rede de *Kohonen*	

Anexo B.6 – **Comportamento do Índice *Davies-Bouldin***

Algoritmo	Índice *Davies-Bouldin*
Ward	
K-Means	
Rede de *Kohonen*	

Anexo B.7 – **Comportamento do Índice *Dunn***

Algoritmo	Índice *Dunn*
Ward	
K-Means	
Rede de *Kohonen*	

ANEXO C: SIMULAÇÕES DO ALGORITMO *K-MEANS*

Anexo C.1 – Matriz de Simulações do *K-Means* (Nº de Grupos = 2) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	2	2	2	2	2	2	2	2	2
2	Arrojado	1	1	1	1	1	1	1	1	1	1
3	Bartolomeu I	2	2	2	2	2	2	2	2	2	2
4	Cachoeira Alves	1	1	1	1	1	1	1	1	1	1
5	Chupadouro	2	2	2	2	2	2	2	2	2	2
6	Cochos	2	2	2	2	2	2	2	2	2	2
7	Emas	2	2	2	2	2	2	2	2	2	2
8	Engo Arcoverde	1	1	1	1	1	1	1	1	1	1
9	Engo Ávidos	1	1	1	1	1	1	1	1	1	1
10	Epitácio Pessoa	1	1	1	1	1	1	1	1	1	1
11	Farinha	1	1	1	1	1	1	1	1	1	1
12	Frutuoso II	2	2	2	2	2	2	2	2	2	2
13	Gamela	2	2	2	2	2	2	2	2	2	2
14	Jatobá I	1	1	1	1	1	1	1	1	1	1
15	Queimadas	2	2	2	2	2	2	2	2	2	2
16	Riac. dos Cavalos	1	1	1	1	1	1	1	1	1	1
17	Serra Vermelha	2	2	2	2	2	2	2	2	2	2
18	Tamanduá I	2	2	2	2	2	2	2	2	2	2
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	1	1	1	1	1	1	1	1	1	1
21	Campos	2	2	2	2	2	2	2	2	2	2
22	Cordeiro	1	1	1	1	1	1	1	1	1	1
23	Santo Antonio	1	1	1	1	1	1	1	1	1	1
24	São Domingos	2	2	2	2	2	2	2	2	2	2
25	Livramento	2	2	2	2	2	2	2	2	2	2
26	Namorados	2	2	2	2	2	2	2	2	2	2
27	Soledade	2	2	2	2	2	2	2	2	2	2
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	1	1	1	1	1	1	1	1	1	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	1	1	1	1	1	1	1	1	1	1
35	Pedro Velho	1	1	1	1	1	1	1	1	1	1
36	Caraúbas	1	1	1	1	1	1	1	1	1	1
37	Poço de Pedras	1	1	1	1	1	1	1	1	1	1
38	Guarita	1	1	1	1	1	1	1	1	1	1
39	Umburana/Sumé	2	2	2	2	2	2	2	2	2	2
40	Jatobá/Sumé	2	2	2	2	2	2	2	2	2	2
41	Gangorra/Sumé	2	2	2	2	2	2	2	2	2	2

Anexo C.2 – Matriz de Simulações do *K-Means* (Nº de Grupos = 3) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	3	2	3	2	1	3	3	2	3
2	Arrojado	1	1	1	2	1	2	1	1	3	2
3	Bartolomeu I	2	3	2	1	2	1	3	3	2	3
4	Cachoeira Alves	1	1	1	2	1	2	1	1	3	2
5	Chupadouro	2	3	2	3	2	1	3	3	2	3
6	Cochos	2	3	3	1	1	1	3	1	2	3
7	Emas	2	3	2	1	2	3	3	3	2	3
8	Engo Arcoverde	3	1	1	2	1	2	1	1	3	2
9	Engo Ávidos	1	1	1	2	1	2	1	1	3	2
10	Epitácio Pessoa	1	1	1	2	3	2	2	2	3	1
11	Farinha	1	1	1	2	3	2	1	1	3	1
12	Frutuoso II	2	3	2	1	2	3	3	3	2	3
13	Gamela	2	3	2	1	2	3	3	3	1	3
14	Jatobá I	1	1	1	2	1	2	1	1	3	2
15	Queimadas	2	3	2	1	1	1	3	3	2	2
16	Riac. dos Cavalos	1	1	1	2	3	2	1	1	3	1
17	Serra Vermelha	2	3	3	1	1	1	3	1	2	3
18	Tamanduá I	2	3	3	3	2	1	3	3	2	3
19	Vazantes	1	3	1	1	1	2	1	1	3	2
20	Camalau	1	1	1	2	1	2	1	1	3	2
21	Campos	2	3	2	1	2	3	3	3	2	3
22	Cordeiro	1	1	1	2	1	2	1	1	3	2
23	Santo Antonio	3	2	2	1	1	1	1	3	2	2
24	São Domingos	2	3	2	1	1	1	3	3	2	2
25	Livramento	2	2	2	3	2	1	3	3	2	3
26	Namorados	2	3	2	3	2	1	3	3	2	3
27	Soledade	2	3	2	1	2	3	3	3	2	3
28	Antenor Navarro	1	1	1	2	3	2	1	1	3	1
29	Aparecida	1	1	1	2	3	2	1	1	3	1
30	Piancó	1	1	1	2	3	2	1	1	3	1
31	Emas	1	1	1	2	1	2	1	1	3	2
32	Patos	1	1	1	2	3	2	1	1	3	1
33	Serra Negra Norte	1	1	1	2	3	2	1	1	3	1
34	Faz. Alagamar	1	1	1	1	1	2	1	1	3	2
35	Pedro Velho	1	1	1	2	3	2	1	1	3	1
36	Caraúbas	1	1	1	2	3	2	1	1	3	1
37	Poço de Pedras	1	1	1	2	3	2	1	1	3	1
38	Guarita	1	1	1	2	3	2	2	2	3	1
39	Umburana/Sumé	2	3	2	3	2	1	3	3	2	3
40	Jatobá/Sumé	2	3	2	3	2	1	3	3	2	3
41	Gangorra/Sumé	2	3	2	1	1	1	3	3	2	2

Anexo C.3 – Matriz de Simulações do *K-Means* (Nº de Grupos = 4) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	4	1	1	2	3	2	4	1	3	4
2	Arrojado	2	3	3	1	2	4	2	3	1	2
3	Bartolomeu I	3	1	1	2	1	3	4	1	3	4
4	Cachoeira Alves	2	3	3	1	2	4	2	3	1	2
5	Chupadouro	4	1	1	2	3	2	4	1	3	4
6	Cochos	3	3	3	2	1	3	2	2	4	4
7	Emas	3	1	1	2	1	3	4	2	4	4
8	Engo Arcoverde	2	3	3	1	4	4	1	3	1	2
9	Engo Ávidos	2	3	3	1	2	4	2	3	1	2
10	Epitácio Pessoa	1	4	4	3	2	1	3	3	2	3
11	Farinha	2	4	2	3	2	1	2	3	2	2
12	Frutuoso II	3	1	1	2	1	3	4	2	4	4
13	Gamela	3	2	1	2	1	3	4	2	4	4
14	Jatobá I	2	3	3	1	2	4	2	3	1	2
15	Queimadas	3	3	3	1	1	4	4	1	1	4
16	Riac. dos Cavalos	2	4	2	3	2	1	2	3	2	2
17	Serra Vermelha	3	3	2	2	1	3	2	2	4	4
18	Tamanduá I	4	1	1	2	3	2	4	1	3	4
19	Vazantes	3	3	3	1	1	4	2	2	1	2
20	Camalau	2	3	3	1	2	4	2	3	1	2
21	Campos	3	1	1	2	1	3	4	1	3	4
22	Cordeiro	2	3	3	1	2	4	2	3	1	2
23	Santo Antonio	3	3	3	1	4	4	1	4	1	1
24	São Domingos	3	3	3	1	1	4	4	1	1	4
25	Livramento	4	1	1	4	3	2	4	4	3	1
26	Namorados	4	1	1	2	3	2	4	1	3	4
27	Soledade	3	1	1	2	1	3	4	1	3	4
28	Antenor Navarro	2	4	2	3	2	1	2	3	2	2
29	Aparecida	2	4	2	3	2	1	2	3	2	2
30	Piancó	2	4	2	3	2	1	2	3	2	2
31	Emas	2	3	3	1	2	4	2	3	1	2
32	Patos	2	4	2	3	2	1	2	3	2	2
33	Serra Negra Norte	2	4	2	3	2	1	2	3	2	2
34	Faz. Alagamar	3	3	3	1	2	4	2	3	1	2
35	Pedro Velho	2	4	2	3	2	1	2	3	2	2
36	Caraúbas	2	4	2	3	2	1	2	3	2	2
37	Poço de Pedras	2	4	2	3	2	1	2	3	2	2
38	Guarita	1	4	4	3	2	1	3	3	2	3
39	Umburana/Sumé	4	1	1	2	3	2	4	1	3	4
40	Jatobá/Sumé	4	1	1	2	3	2	4	1	3	4
41	Gangorra/Sumé	3	3	3	1	1	4	4	1	1	4

Anexo C.4 – Matriz de Simulações do *K-Means* (Nº de Grupos = 6) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	5	4	6	5	2	4	4	5	6	2
2	Arrojado	1	6	1	6	3	6	1	6	4	1
3	Bartolomeu I	5	4	5	4	2	4	5	6	6	2
4	Cachoeira Alves	1	6	1	6	3	6	1	6	4	1
5	Chupadouro	5	4	6	5	2	4	4	5	6	2
6	Cochos	4	5	5	3	4	1	5	2	6	1
7	Emas	5	4	2	1	2	1	5	1	2	2
8	Engo Arcoverde	1	2	1	4	1	2	1	6	4	3
9	Engo Ávidos	1	6	1	6	3	6	1	6	4	1
10	Epitácio Pessoa	3	3	3	2	5	3	3	4	3	6
11	Farinha	2	1	3	6	5	5	2	3	5	6
12	Frutuoso II	5	4	2	1	2	1	5	1	2	2
13	Gamela	5	4	2	1	2	1	5	1	2	2
14	Jatobá I	1	6	1	6	3	6	1	6	4	1
15	Queimadas	1	6	5	4	3	6	1	6	4	1
16	Riac. dos Cavalos	2	1	3	3	4	5	2	3	5	6
17	Serra Vermelha	4	5	5	3	4	1	5	2	6	1
18	Tamanduá I	4	5	6	5	2	4	4	2	6	4
19	Vazantes	1	6	5	6	3	6	1	6	4	1
20	Camalau	1	6	1	4	3	6	1	6	4	1
21	Campos	5	4	5	1	2	4	5	6	6	2
22	Cordeiro	1	6	1	6	3	6	1	6	4	1
23	Santo Antonio	1	2	1	4	1	2	1	6	4	3
24	São Domingos	1	6	5	4	3	6	1	6	4	1
25	Livramento	6	4	4	5	6	4	6	5	1	5
26	Namorados	5	4	6	5	2	4	4	5	6	2
27	Soledade	5	4	5	1	2	4	5	6	6	2
28	Antenor Navarro	2	1	3	2	5	5	2	3	5	6
29	Aparecida	2	1	3	2	5	5	2	3	5	6
30	Piancó	2	1	3	6	5	5	2	3	5	6
31	Emas	1	6	1	6	3	6	1	6	4	1
32	Patos	2	1	3	6	5	5	2	3	5	6
33	Serra Negra Norte	2	1	3	6	5	5	2	3	5	6
34	Faz. Alagamar	1	6	1	6	3	6	1	6	4	1
35	Pedro Velho	2	1	3	3	4	5	2	3	5	6
36	Caraúbas	2	1	3	3	4	5	2	3	5	6
37	Poço de Pedras	2	1	3	2	5	5	2	3	5	6
38	Guarita	3	3	3	2	5	3	3	4	3	6
39	Umburana/Sumé	5	4	6	5	2	4	4	5	6	2
40	Jatobá/Sumé	5	4	6	5	2	4	4	5	6	2
41	Gangorra/Sumé	1	6	5	4	3	6	1	6	4	1

Anexo C.5 – Matriz de Simulações do *K-Means* (Nº de Grupos = 2) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	1	1	1	2	2	2	2	2	1
2	Arrojado	1	2	2	2	1	1	1	1	1	2
3	Bartolomeu I	2	1	1	1	2	2	2	2	2	1
4	Cachoeira Alves	1	2	2	2	1	1	1	1	1	2
5	Chupadouro	2	1	1	1	2	2	2	2	2	1
6	Cochos	2	1	1	1	2	2	2	2	2	1
7	Emas	2	1	1	1	2	2	2	2	2	1
8	Engo Arcoverde	1	2	2	2	1	1	1	1	1	2
9	Engo Ávidos	1	2	2	2	1	1	1	1	1	2
10	Epitácio Pessoa	1	2	2	2	1	1	1	1	1	2
11	Farinha	1	2	2	2	1	1	1	1	1	2
12	Frutuoso II	2	1	1	1	2	2	2	2	2	1
13	Gamela	2	1	1	1	2	2	2	2	2	1
14	Jatobá I	1	2	2	2	1	1	1	1	1	2
15	Queimadas	2	1	1	1	2	2	2	2	2	1
16	Riac. dos Cavalos	1	2	2	2	1	1	1	1	1	2
17	Serra Vermelha	2	1	1	1	2	2	2	2	2	1
18	Tamanduá I	2	1	1	1	2	2	2	2	2	1
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	1	1	1	1	1	1	1	1	1	1
21	Campos	2	1	1	1	2	2	2	2	2	1
22	Cordeiro	1	1	1	1	1	1	1	1	1	1
23	Santo Antonio	2	1	1	1	2	2	2	2	2	1
24	São Domingos	2	1	1	1	2	2	2	2	2	1
25	Livramento	2	1	1	1	2	2	2	2	2	1
26	Namorados	2	1	1	1	2	2	2	2	2	1
27	Soledade	2	1	1	1	2	2	2	2	2	1
28	Antenor Navarro	1	2	2	2	1	1	1	1	1	2
29	Aparecida	1	2	2	2	1	1	1	1	1	2
30	Piancó	1	2	2	2	1	1	1	1	1	2
31	Emas	1	2	2	2	1	1	1	1	1	2
32	Patos	1	2	2	2	1	1	1	1	1	2
33	Serra Negra Norte	1	2	2	2	1	1	1	1	1	2
34	Faz. Alagamar	1	1	1	1	1	1	1	1	1	1
35	Pedro Velho	1	2	2	2	1	1	1	1	1	2
36	Caraúbas	1	2	2	2	1	1	1	1	1	2
37	Poço de Pedras	1	2	2	2	1	1	1	1	1	2
38	Guarita	1	2	2	2	1	1	1	1	1	2
39	Umburana/Sumé	2	1	1	1	2	2	2	2	2	1
40	Jatobá/Sumé	2	1	1	1	2	2	2	2	2	1
41	Gangorra/Sumé	2	1	1	1	2	2	2	2	2	1

Anexo C.6 – Matriz de Simulações do *K-Means* (Nº de Grupos = 3) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	3	1	3	3	1	2	2	1	3	2
2	Arrojado	2	3	1	2	3	3	3	3	1	1
3	Bartolomeu I	3	1	3	3	1	2	2	1	3	2
4	Cachoeira Alves	2	3	1	2	3	1	2	3	1	1
5	Chupadouro	3	1	3	3	1	2	2	1	3	2
6	Cochos	2	1	3	2	3	1	2	1	1	2
7	Emas	3	1	3	3	1	2	2	1	3	2
8	Engo Arcoverde	2	3	1	1	3	3	3	3	2	1
9	Engo Ávidos	1	3	1	1	3	3	3	3	2	1
10	Epitácio Pessoa	1	3	2	1	2	1	3	3	2	3
11	Farinha	1	3	1	1	2	3	3	3	2	1
12	Frutuoso II	3	1	3	3	1	2	2	1	3	2
13	Gamela	3	1	3	3	1	2	1	1	3	2
14	Jatobá I	1	3	1	1	3	3	3	3	2	1
15	Queimadas	2	1	3	2	1	2	2	1	3	2
16	Riac. dos Cavalos	1	3	1	1	3	3	3	3	2	1
17	Serra Vermelha	2	1	3	2	3	1	2	1	1	2
18	Tamanduá I	3	1	3	3	1	2	2	1	3	2
19	Vazantes	2	3	3	2	3	1	2	1	1	1
20	Camalau	2	3	3	2	3	1	2	1	1	1
21	Campos	3	1	3	3	1	2	2	1	3	2
22	Cordeiro	2	3	3	2	3	1	2	1	1	1
23	Santo Antonio	2	1	3	2	1	2	2	1	3	2
24	São Domingos	2	1	3	2	1	2	2	1	3	2
25	Livramento	3	1	3	3	1	2	2	1	3	2
26	Namorados	3	1	3	3	1	2	2	1	3	2
27	Soledade	3	1	3	3	1	2	2	1	3	2
28	Antenor Navarro	1	3	1	1	2	3	3	3	2	1
29	Aparecida	1	3	1	1	3	3	3	3	2	3
30	Piancó	1	2	1	2	3	1	3	2	1	3
31	Emas	2	3	1	2	3	3	3	3	1	1
32	Patos	1	2	1	2	3	1	3	2	1	3
33	Serra Negra Norte	1	2	1	2	3	1	3	2	1	3
34	Faz. Alagamar	2	3	3	2	3	1	2	1	1	1
35	Pedro Velho	2	3	3	2	3	1	2	3	1	3
36	Caraúbas	2	3	3	2	3	1	2	3	1	3
37	Poço de Pedras	2	3	3	2	3	1	2	3	1	3
38	Guarita	1	3	2	1	2	3	3	3	2	3
39	Umburana/Sumé	3	1	3	3	1	2	2	1	3	2
40	Jatobá/Sumé	3	1	3	3	1	2	2	1	3	2
41	Gangorra/Sumé	2	1	3	3	1	2	2	1	3	2

Anexo C.7 – Matriz de Simulações do *K-Means* (Nº de Grupos = 4) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	1	1	3	1	3	1	2	4	3	1
2	Arrojado	4	2	2	3	1	4	4	2	4	4
3	Bartolomeu I	3	1	3	1	3	1	2	4	3	1
4	Cachoeira Alves	3	2	2	3	1	4	1	2	4	4
5	Chupadouro	1	1	3	1	3	1	2	4	3	1
6	Cochos	3	1	3	3	1	4	1	2	4	1
7	Emas	1	1	3	1	3	1	2	4	3	1
8	Engo Arcoverde	4	2	2	4	1	4	4	1	4	4
9	Engo Ávidos	4	2	2	4	1	4	4	1	1	4
10	Epitácio Pessoa	2	3	1	4	2	3	3	1	2	2
11	Farinha	4	2	2	4	2	3	4	1	1	4
12	Frutuoso II	1	1	3	1	3	1	2	4	3	1
13	Gamela	1	4	3	2	4	1	2	4	3	3
14	Jatobá I	4	2	2	4	1	4	4	1	1	4
15	Queimadas	3	1	3	3	3	4	2	2	4	1
16	Riac. dos Cavalos	4	2	2	4	1	4	4	1	1	4
17	Serra Vermelha	3	1	3	3	1	4	1	2	4	1
18	Tamanduá I	1	1	3	1	3	1	2	4	3	1
19	Vazantes	3	1	3	3	1	4	1	2	4	1
20	Camalau	3	1	3	3	1	4	1	2	4	1
21	Campos	1	1	3	1	3	1	2	4	3	1
22	Cordeiro	3	1	3	3	1	4	1	2	4	1
23	Santo Antonio	3	1	3	3	3	4	2	2	4	1
24	São Domingos	3	1	3	3	3	4	2	2	4	1
25	Livramento	1	1	3	1	3	1	2	4	3	1
26	Namorados	1	1	3	1	3	1	2	4	3	1
27	Soledade	1	1	3	1	3	1	2	4	3	1
28	Antenor Navarro	4	2	2	4	2	3	4	1	1	4
29	Aparecida	2	2	2	4	1	4	4	1	1	4
30	Piancó	2	2	4	3	1	2	1	3	1	2
31	Emas	4	2	2	3	1	4	4	2	1	4
32	Patos	2	2	4	3	1	2	1	3	1	2
33	Serra Negra Norte	2	2	4	3	1	2	1	3	1	2
34	Faz. Alagamar	3	1	3	3	1	4	1	2	4	1
35	Pedro Velho	3	1	3	3	1	4	1	2	4	1
36	Caraúbas	3	1	3	3	1	4	1	2	4	1
37	Poço de Pedras	3	1	3	3	1	4	1	2	4	1
38	Guarita	2	3	1	4	2	3	3	1	2	2
39	Umburana/Sumé	1	1	3	1	3	1	2	4	3	1
40	Jatobá/Sumé	1	1	3	1	3	1	2	4	3	1
41	Gangorra/Sumé	3	1	3	1	3	4	2	2	4	1

Anexo C.8 – Matriz de Simulações do *K-Means* (Nº de Grupos = 6) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	5	4	2	1	1	1	3	2	3	5
2	Arrojado	1	6	5	6	4	2	6	1	5	1
3	Bartolomeu I	5	5	2	1	1	4	3	2	3	2
4	Cachoeira Alves	4	6	5	6	3	2	6	3	5	1
5	Chupadouro	5	4	2	1	1	1	3	2	3	5
6	Cochos	4	5	5	6	3	4	6	3	5	2
7	Emas	5	4	2	1	5	1	5	2	3	5
8	Engo Arcoverde	1	6	5	5	4	5	6	1	5	1
9	Engo Ávidos	1	6	1	5	4	5	4	1	4	6
10	Epitácio Pessoa	3	3	3	5	6	2	1	6	1	3
11	Farinha	1	2	1	5	4	5	4	1	6	6
12	Frutuoso II	5	4	2	1	1	1	3	2	3	5
13	Gamela	6	4	6	4	2	6	2	4	2	4
14	Jatobá I	1	6	1	5	4	5	4	1	4	6
15	Queimadas	4	5	5	6	3	4	6	3	5	2
16	Riac. dos Cavalos	1	6	1	5	4	5	4	1	4	6
17	Serra Vermelha	4	5	5	6	3	4	6	3	5	2
18	Tamanduá I	5	4	2	1	5	1	5	2	3	5
19	Vazantes	4	5	5	6	3	4	6	3	5	1
20	Camalau	4	5	5	6	3	4	6	3	5	1
21	Campos	5	4	2	1	1	4	3	2	3	2
22	Cordeiro	4	5	5	6	3	2	6	3	5	1
23	Santo Antonio	4	5	5	6	3	4	6	3	5	2
24	São Domingos	4	5	5	6	3	4	6	3	5	2
25	Livramento	5	4	4	2	5	1	5	5	3	5
26	Namorados	5	4	2	1	1	1	3	2	3	5
27	Soledade	5	4	2	1	1	1	3	2	3	5
28	Antenor Navarro	1	2	1	5	4	5	4	1	6	6
29	Aparecida	1	6	1	5	6	5	4	6	4	6
30	Piancó	2	1	1	3	6	3	4	6	4	6
31	Emas	1	6	1	6	4	2	4	1	5	1
32	Patos	2	1	1	3	6	3	4	6	4	6
33	Serra Negra Norte	2	1	1	3	6	3	4	6	4	6
34	Faz. Alagamar	4	5	5	6	3	2	6	3	5	1
35	Pedro Velho	4	5	5	6	3	2	6	3	1	1
36	Caraúbas	4	5	5	6	3	2	6	3	1	1
37	Poço de Pedras	4	5	5	6	3	2	6	3	1	1
38	Guarita	3	3	3	5	6	5	1	6	6	3
39	Umburana/Sumé	5	4	2	1	5	1	5	2	3	5
40	Jatobá/Sumé	5	4	2	1	1	1	3	2	3	5
41	Gangorra/Sumé	4	5	5	6	1	4	3	3	5	2

Anexo C.9 – Matriz de Simulações do *K-Means* (Nº de Grupos = 2) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	1	2	2	1	1	1	1	2	2	1
2	Arrojado	1	2	1	2	1	1	2	1	1	2
3	Bartolomeu I	1	2	1	2	1	1	2	1	1	2
4	Cachoeira Alves	1	2	1	2	1	1	2	1	1	2
5	Chupadouro	1	2	2	2	1	1	2	2	2	1
6	Cochos	1	2	1	2	1	1	2	1	2	2
7	Emas	1	2	1	2	2	2	1	2	1	1
8	Engo Arcoverde	1	2	2	1	1	1	2	1	2	2
9	Engo Ávidos	1	2	1	2	2	1	1	1	1	2
10	Epitácio Pessoa	2	2	2	1	2	1	2	1	1	1
11	Farinha	1	2	1	1	2	1	1	1	1	2
12	Frutuoso II	1	2	1	2	2	2	1	1	1	1
13	Gamela	1	1	1	1	2	2	1	1	2	1
14	Jatobá I	1	2	1	2	1	1	2	1	1	2
15	Queimadas	1	2	2	1	1	1	1	2	2	1
16	Riac. dos Cavalos	1	2	1	1	1	1	1	1	1	2
17	Serra Vermelha	1	2	1	2	1	1	2	1	1	2
18	Tamanduá I	1	2	2	2	1	1	1	2	2	1
19	Vazantes	1	2	1	2	2	1	2	1	1	2
20	Camalau	1	2	2	1	1	1	1	2	2	1
21	Campos	1	1	2	1	1	2	1	1	2	1
22	Cordeiro	1	2	1	1	2	2	1	1	1	1
23	Santo Antonio	1	2	2	2	1	1	2	1	2	1
24	São Domingos	1	2	2	1	1	1	1	2	2	1
25	Livramento	1	2	2	2	1	1	2	2	2	1
26	Namorados	1	2	2	2	1	1	2	2	2	1
27	Soledade	1	1	2	1	1	2	2	1	2	1
28	Antenor Navarro	1	1	1	1	1	1	2	1	2	2
29	Aparecida	1	2	2	1	1	1	2	1	1	2
30	Piancó	2	2	1	2	2	1	2	1	1	2
31	Emas	1	2	2	1	2	2	1	2	2	1
32	Patos	1	2	1	1	2	1	1	1	1	2
33	Serra Negra Norte	1	2	1	1	2	1	1	1	1	2
34	Faz. Alagamar	1	2	2	1	2	2	1	1	2	1
35	Pedro Velho	1	2	1	1	2	2	1	1	1	1
36	Caraúbas	2	2	2	1	2	1	1	2	1	1
37	Poço de Pedras	1	1	2	1	1	1	2	1	2	1
38	Guarita	2	2	2	1	2	1	2	1	1	1
39	Umburana/Sumé	1	2	2	2	1	1	2	2	2	1
40	Jatobá/Sumé	1	2	2	2	1	1	1	2	2	1
41	Gangorra/Sumé	1	2	2	1	1	1	1	2	2	1

Anexo C.10 – Matriz de Simulações do *K-Means* (Nº de Grupos = 3) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	3	2	3	3	1	2	1	1	1
2	Arrojado	2	1	3	1	1	2	1	2	2	1
3	Bartolomeu I	1	1	3	1	1	3	1	2	2	1
4	Cach. Alves	2	1	3	1	1	2	1	2	2	1
5	Chupadouro	1	1	3	1	1	3	2	1	2	1
6	Cochos	2	1	3	1	1	2	1	2	2	1
7	Emas	3	3	2	2	2	2	1	3	1	3
8	Engº Arcoverde	2	1	2	3	3	3	2	2	2	1
9	Engº Ávidos	2	1	2	2	2	2	1	2	2	3
10	Epit. Pessoa	1	2	2	2	2	1	3	2	3	2
11	Farinha	2	1	2	3	1	2	1	2	2	3
12	Frutuoso II	3	1	3	2	2	2	1	3	1	3
13	Gamela	3	2	1	2	3	3	2	3	1	2
14	Jatobá I	2	1	3	1	1	2	1	2	2	1
15	Queimadas	2	3	2	3	3	1	2	1	1	1
16	Riac. Cavalos	2	1	2	3	1	2	1	2	2	1
17	Serra Vermelha	1	1	3	1	1	2	1	2	2	1
18	Tamanduá I	1	3	2	3	3	1	2	1	1	1
19	Vazantes	2	1	3	1	1	2	1	2	2	3
20	Camalau	2	3	2	3	3	1	2	1	1	1
21	Campos	3	2	1	1	3	3	2	3	1	2
22	Cordeiro	2	3	2	2	2	2	1	3	1	3
23	Santo Antonio	1	1	3	1	1	3	2	1	2	1
24	São Domingos	2	3	2	3	3	1	2	1	1	1
25	Livramento	1	1	3	1	1	3	2	1	2	1
26	Namorados	1	1	3	1	1	3	2	1	2	1
27	Soledade	3	2	1	1	3	3	2	3	1	2
28	Ant. Navarro	2	2	1	1	3	3	2	2	2	2
29	Aparecida	1	2	3	1	1	3	2	2	2	2
30	Piancó	1	1	3	2	2	2	1	2	2	3
31	Emas	2	3	2	3	3	1	2	1	1	3
32	Patos	2	1	2	2	1	2	1	2	2	3
33	Serra N. Norte	2	1	2	2	2	2	1	2	2	3
34	Faz. Alagamar	3	2	2	2	3	3	2	3	1	2
35	Pedro Velho	3	2	2	2	2	2	1	3	1	3
36	Caraúbas	1	3	2	2	2	1	3	1	1	3
37	Poço Pedras	3	2	1	1	3	3	2	2	1	2
38	Guarita	1	2	2	2	2	1	3	2	3	2
39	Umbur/Sumé	1	1	3	1	1	3	2	1	2	1
40	Jatobá/Sumé	1	3	2	3	3	1	2	1	1	1
41	Gangor/Sumé	2	3	2	3	3	1	2	1	1	1

Anexo C.11 – Matriz de Simulações do *K-Means* (Nº de Grupos = 4) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	4	4	4	1	3	4	2	1	2	4
2	Arrojado	3	2	2	3	2	3	3	4	4	2
3	Bartolomeu I	2	2	2	4	2	3	4	4	4	3
4	Cachoeira Alves	3	2	2	3	2	3	3	4	4	2
5	Chupadouro	2	2	2	4	2	3	4	4	2	3
6	Cochos	3	2	2	3	2	3	3	4	4	2
7	Emas	4	3	4	1	1	2	1	1	1	2
8	Engo Arcoverde	3	4	3	3	4	4	3	1	4	4
9	Engo Ávidos	3	3	2	3	1	3	3	1	4	2
10	Epitácio Pessoa	1	3	1	4	4	1	3	3	3	1
11	Farinha	3	4	3	3	4	4	3	1	4	4
12	Frutuoso II	3	3	2	3	1	2	1	1	1	2
13	Gamela	1	1	3	2	1	2	1	2	1	2
14	Jatobá I	3	2	2	3	2	3	3	4	4	2
15	Queimadas	4	4	4	1	3	4	2	1	2	4
16	Riac. Cavalos	3	4	3	3	4	4	3	1	4	4
17	Serra Vermelha	3	2	2	3	2	3	3	4	4	2
18	Tamanduá I	4	4	4	1	3	3	2	1	2	3
19	Vazantes	3	2	2	3	2	3	3	4	4	2
20	Camalau	4	4	4	1	3	4	2	1	2	4
21	Campos	1	1	3	2	4	2	4	2	1	2
22	Cordeiro	4	3	4	1	1	4	1	1	1	4
23	Santo Antonio	2	2	2	4	2	3	4	4	2	3
24	São Domingos	4	4	4	1	3	4	2	1	2	4
25	Livramento	2	2	2	4	2	3	4	4	2	3
26	Namorados	2	2	2	4	2	3	4	4	2	3
27	Soledade	1	1	3	2	4	2	4	2	1	2
28	Antenor Navarro	1	1	3	2	4	4	3	2	4	2
29	Aparecida	1	2	3	4	4	3	3	4	4	2
30	Piancó	3	3	2	4	2	3	3	4	4	2
31	Emas	4	4	4	1	3	4	2	1	2	4
32	Patos	3	3	3	3	4	4	3	1	4	2
33	Serra N. Norte	3	3	3	3	4	4	3	1	4	2
34	Faz. Alagamar	1	3	3	2	4	2	1	1	1	2
35	Pedro Velho	1	3	3	2	1	2	1	1	1	2
36	Caraúbas	4	3	4	1	3	1	2	1	3	1
37	Poço de Pedras	1	1	3	2	4	2	4	2	1	2
38	Guarita	1	3	1	4	4	1	3	3	3	1
39	Umburana/Sumé	2	2	2	4	2	3	4	4	2	3
40	Jatobá/Sumé	4	4	4	1	3	4	2	1	2	3
41	Gangorra/Sumé	4	4	4	1	3	4	2	1	2	4

Anexo C.12 – Matriz de Simulações do *K-Means* (Nº de Grupos = 6) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	1	4	2	3	5	6	5	2	1	5
2	Arrojado	5	3	3	4	2	2	6	5	3	2
3	Bartolomeu I	1	1	6	4	6	2	6	5	4	2
4	Cachoeira Alves	5	3	3	4	2	2	6	5	4	2
5	Chupadouro	1	1	6	2	6	6	5	4	2	1
6	Cochos	5	3	3	4	2	2	6	5	3	2
7	Emas	3	5	5	1	1	4	2	3	1	4
8	Engo Arcoverde	5	3	3	6	4	5	3	2	3	3
9	Engo Ávidos	2	5	3	4	2	2	6	5	4	2
10	Epitácio Pessoa	6	6	4	5	3	3	1	6	5	6
11	Farinha	5	3	3	6	4	5	6	5	3	3
12	Frutuoso II	3	5	5	1	1	4	2	3	4	4
13	Gamela	4	2	1	1	1	1	4	1	6	4
14	Jatobá I	5	3	3	4	2	2	6	5	4	2
15	Queimadas	1	4	2	3	5	6	5	2	1	5
16	Riac. dos Cavalos	5	3	3	6	4	5	3	2	3	3
17	Serra Vermelha	5	3	3	4	2	2	6	5	4	2
18	Tamanduá I	1	4	2	3	5	6	5	4	2	1
19	Vazantes	5	3	3	4	2	2	6	5	4	2
20	Camalau	5	4	2	3	5	5	5	2	1	5
21	Campos	4	2	1	2	3	1	4	1	6	4
22	Cordeiro	2	5	5	1	1	4	2	3	1	5
23	Santo Antonio	1	1	6	2	6	6	5	4	2	1
24	São Domingos	5	4	2	3	5	5	5	2	1	5
25	Livramento	1	1	6	2	6	6	5	4	2	1
26	Namorados	1	1	6	2	6	6	5	4	2	1
27	Soledade	4	2	1	2	3	1	4	1	6	4
28	Antenor Navarro	5	3	1	6	4	5	3	1	3	3
29	Aparecida	5	3	3	6	3	2	3	5	3	3
30	Piancó	2	5	3	4	2	2	6	5	4	2
31	Emas	2	4	2	3	5	4	2	2	1	5
32	Patos	5	3	3	6	2	2	6	5	3	3
33	Serra Negra Norte	2	3	3	6	2	2	6	5	3	3
34	Faz. Alagamar	2	5	5	1	3	4	2	3	1	4
35	Pedro Velho	2	5	5	1	1	4	2	3	1	4
36	Caraúbas	2	5	2	3	3	4	2	6	1	5
37	Poço de Pedras	4	2	1	6	3	1	3	1	6	3
38	Guarita	6	6	4	5	3	3	1	6	5	6
39	Umburana/Sumé	1	1	6	2	6	6	5	4	2	1
40	Jatobá/Sumé	1	4	2	3	5	6	5	4	1	5
41	Gangorra/Sumé	1	4	2	3	5	6	5	2	1	5

Anexo C.13 – Matriz de Simulações do *K-Means* (Nº de Grupos = 2) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	1	1	1	1	1	1	1	1	1	1
2	Arrojado	1	1	1	1	2	1	2	2	1	1
3	Bartolomeu I	1	1	1	1	2	1	2	2	1	1
4	Cachoeira Alves	1	1	1	1	2	1	2	2	1	1
5	Chupadouro	1	1	1	1	2	1	2	2	1	1
6	Cochos	1	1	1	1	2	1	2	2	1	1
7	Emas	1	1	1	1	2	1	2	2	1	1
8	Engo Arcoverde	1	1	1	1	2	1	2	2	1	1
9	Engo Ávidos	1	1	1	1	2	1	2	2	1	1
10	Epitácio Pessoa	2	2	2	2	1	2	1	1	2	2
11	Farinha	2	2	2	2	2	2	2	2	2	2
12	Frutuoso II	1	1	1	1	2	1	2	2	1	1
13	Gamela	1	1	1	1	1	1	1	1	1	1
14	Jatobá I	1	1	1	1	2	1	2	2	1	1
15	Queimadas	1	1	1	1	2	1	2	2	1	1
16	Riac. dos Cavalos	1	1	1	1	2	1	2	2	1	1
17	Serra Vermelha	1	1	1	1	2	1	2	2	1	1
18	Tamanduá I	1	1	1	1	1	1	1	1	1	1
19	Vazantes	1	1	1	1	2	1	2	2	1	1
20	Camalau	1	1	1	1	1	1	1	1	1	1
21	Campos	1	1	1	1	1	1	1	1	1	1
22	Cordeiro	2	2	2	2	1	2	1	1	2	2
23	Santo Antonio	1	1	1	1	1	1	1	1	1	1
24	São Domingos	1	1	1	1	1	1	1	1	1	1
25	Livramento	1	1	1	1	1	1	1	1	1	1
26	Namorados	1	1	1	1	1	1	1	1	1	1
27	Soledade	1	1	1	1	1	1	1	1	1	1
28	Antenor Navarro	2	2	2	2	2	2	2	2	2	2
29	Aparecida	2	2	2	2	2	2	2	2	2	2
30	Piancó	2	2	2	2	2	2	2	2	2	2
31	Emas	1	1	1	1	2	1	2	2	1	1
32	Patos	2	2	2	2	2	2	2	2	2	2
33	Serra Negra Norte	2	2	2	2	2	2	2	2	2	2
34	Faz. Alagamar	2	2	2	2	1	2	1	1	2	2
35	Pedro Velho	2	2	2	2	1	2	1	1	2	2
36	Caraúbas	2	2	2	2	1	2	1	1	2	2
37	Poço de Pedras	2	2	2	2	1	2	1	1	2	2
38	Guarita	2	2	2	2	1	2	1	1	2	2
39	Umburana/Sumé	1	1	1	1	1	1	1	1	1	1
40	Jatobá/Sumé	1	1	1	1	1	1	1	1	1	1
41	Gangorra/Sumé	1	1	1	1	1	1	1	1	1	1

Anexo C.14 – Matriz de Simulações do *K-Means* (Nº de Grupos = 3) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	2	1	1	1	3	1	2	1	1
2	Arrojado	2	3	1	1	1	1	1	2	2	1
3	Bartolomeu I	2	3	1	2	3	1	1	3	2	1
4	Cachoeira Alves	2	3	1	2	1	1	1	3	2	1
5	Chupadouro	2	3	1	1	1	1	1	2	2	1
6	Cochos	2	3	1	2	3	1	1	3	2	1
7	Emas	2	3	1	2	3	1	1	3	2	1
8	Engo Arcoverde	2	3	1	2	1	1	1	3	2	1
9	Engo Ávidos	2	3	3	3	1	1	1	1	2	1
10	Epitácio Pessoa	3	1	2	3	2	2	3	1	3	2
11	Farinha	3	3	3	3	1	1	2	1	2	2
12	Frutuoso II	2	3	1	2	3	1	1	3	2	1
13	Gamela	1	2	1	2	3	3	1	3	1	3
14	Jatobá I	2	3	1	2	1	1	1	3	2	1
15	Queimadas	2	3	1	2	3	1	1	3	2	1
16	Riac. dos Cavalos	2	3	1	1	1	1	1	2	2	1
17	Serra Vermelha	2	3	1	2	3	1	1	3	2	1
18	Tamanduá I	2	2	1	2	3	1	1	3	1	1
19	Vazantes	2	3	1	2	3	1	1	3	2	1
20	Camalau	2	2	3	2	3	1	2	3	1	1
21	Campos	1	2	1	2	3	3	2	3	1	3
22	Cordeiro	3	1	3	3	2	2	2	1	3	2
23	Santo Antonio	2	2	1	2	3	1	2	3	1	1
24	São Domingos	2	2	1	2	3	3	2	3	1	1
25	Livramento	2	2	1	2	3	1	1	3	1	1
26	Namorados	2	2	1	2	3	1	1	3	1	1
27	Soledade	1	2	1	2	3	3	2	3	1	3
28	Antenor Navarro	3	1	3	3	2	2	2	1	3	2
29	Aparecida	3	1	3	3	2	2	3	1	3	2
30	Piancó	3	1	3	3	2	2	3	1	3	2
31	Emas	2	3	1	2	1	1	1	3	2	1
32	Patos	3	1	3	3	2	2	2	1	3	2
33	Serra Negra Norte	3	1	3	3	2	2	3	1	3	2
34	Faz. Alagamar	3	1	3	3	2	2	2	1	3	2
35	Pedro Velho	3	1	3	3	2	2	2	1	3	2
36	Caraúbas	3	1	3	3	2	2	3	1	3	2
37	Poço de Pedras	3	1	3	3	2	2	2	1	3	2
38	Guarita	3	1	2	3	2	2	3	1	3	2
39	Umburana/Sumé	2	2	1	2	3	1	1	3	1	1
40	Jatobá/Sumé	2	2	1	1	1	3	1	2	1	1
41	Gangorra/Sumé	2	2	1	2	3	1	1	3	1	1

Anexo C.15 – Matriz de Simulações do *K-Means* (Nº de Grupos = 4) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	3	2	1	2	3	4	2	1	1	4
2	Arrojado	3	3	3	1	2	3	4	3	3	2
3	Bartolomeu I	1	3	3	1	4	2	4	3	3	2
4	Cachoeira Alves	1	3	3	1	2	3	4	3	3	2
5	Chupadouro	3	3	3	1	2	3	4	3	3	2
6	Cochos	1	3	3	1	4	2	4	3	3	2
7	Emas	1	3	1	1	4	2	1	3	4	2
8	Engo Arcoverde	1	3	3	1	2	3	4	2	3	2
9	Engo Ávidos	1	4	3	1	2	3	4	2	3	1
10	Epitácio Pessoa	4	1	2	3	1	1	3	4	2	3
11	Farinha	1	4	3	1	2	3	4	2	3	1
12	Frutuoso II	1	3	3	1	4	2	1	3	3	2
13	Gamela	2	2	1	4	3	4	1	1	1	4
14	Jatobá I	1	3	3	1	2	3	4	2	3	2
15	Queimadas	1	3	3	1	4	2	1	3	4	2
16	Riac. dos Cavalos	3	3	3	1	2	3	4	2	3	2
17	Serra Vermelha	1	3	3	1	4	2	4	3	3	2
18	Tamanduá I	2	2	1	2	4	4	1	3	4	4
19	Vazantes	1	3	3	1	4	2	4	3	3	2
20	Camalau	2	4	4	2	3	4	2	2	4	4
21	Campos	2	2	4	4	3	4	2	1	4	4
22	Cordeiro	2	4	4	3	3	1	2	2	4	1
23	Santo Antonio	2	2	4	2	3	4	2	2	4	4
24	São Domingos	2	2	4	2	3	4	2	1	4	4
25	Livramento	2	2	1	2	4	4	1	3	4	4
26	Namorados	2	2	1	2	4	4	1	3	4	4
27	Soledade	2	2	4	4	3	4	2	1	1	4
28	Antenor Navarro	4	4	2	3	2	3	3	2	2	1
29	Aparecida	4	4	2	3	1	1	3	4	2	1
30	Piancó	4	4	2	3	1	1	3	4	2	1
31	Emas	1	3	3	1	2	3	4	2	3	2
32	Patos	4	4	2	3	1	1	3	2	2	1
33	Serra Negra Norte	4	4	2	3	1	1	3	4	2	1
34	Faz. Alagamar	2	4	4	3	3	1	2	4	2	3
35	Pedro Velho	4	4	2	3	1	1	3	4	2	3
36	Caraúbas	4	4	2	3	1	1	3	4	2	3
37	Poço de Pedras	4	4	2	3	1	1	3	4	2	3
38	Guarita	4	1	2	3	1	1	3	4	2	3
39	Umburana/Sumé	2	2	1	2	4	4	1	3	4	4
40	Jatobá/Sumé	3	2	1	2	3	4	2	1	1	4
41	Gangorra/Sumé	2	2	1	2	3	4	2	3	4	4

Anexo C.16 – Matriz de Simulações do *K-Means* (Nº de Grupos = 6) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	1	2	3	5	1	3	1	4	6	2
2	Arrojado	4	5	2	5	4	1	4	5	6	6
3	Bartolomeu I	5	6	1	6	2	5	6	6	3	4
4	Cachoeira Alves	4	5	2	6	4	1	4	5	3	6
5	Chupadouro	4	2	2	5	4	3	4	3	6	6
6	Cochos	5	6	1	6	2	5	6	6	3	4
7	Emas	5	6	1	6	2	5	6	6	5	4
8	Engo Arcoverde	4	5	2	2	4	1	4	3	3	6
9	Engo Ávidos	6	5	2	2	6	1	3	5	3	6
10	Epitácio Pessoa	2	4	4	3	3	4	5	2	2	3
11	Farinha	6	5	2	2	6	1	3	5	3	6
12	Frutuoso II	5	6	1	6	2	5	6	6	3	4
13	Gamela	3	1	5	4	1	6	1	4	4	2
14	Jatobá I	4	5	2	2	4	1	4	5	3	6
15	Queimadas	5	6	1	6	2	5	6	6	5	4
16	Riac. dos Cavalos	4	5	2	2	4	1	4	5	6	6
17	Serra Vermelha	5	6	1	6	2	5	6	6	3	4
18	Tamanduá I	1	6	3	1	2	6	1	6	5	4
19	Vazantes	5	6	1	6	2	5	6	6	3	4
20	Camalau	6	3	6	1	5	6	2	1	1	5
21	Campos	3	1	6	1	5	6	2	4	4	5
22	Cordeiro	6	3	6	1	5	2	2	1	1	5
23	Santo Antonio	1	6	6	1	5	6	2	4	1	5
24	São Domingos	1	6	6	1	5	6	2	4	1	5
25	Livramento	1	6	3	1	2	6	1	6	5	4
26	Namorados	1	6	3	1	2	6	1	6	5	4
27	Soledade	3	1	6	1	5	6	2	4	4	5
28	Antenor Navarro	6	5	2	2	6	1	3	5	2	1
29	Aparecida	2	3	4	3	3	2	5	1	2	1
30	Piancó	2	3	4	3	3	2	5	1	2	1
31	Emas	6	5	2	2	6	1	3	5	3	6
32	Patos	6	3	4	2	6	2	3	1	2	1
33	Serra Negra Norte	6	3	4	3	6	2	3	1	2	1
34	Faz. Alagamar	6	3	6	1	5	2	2	1	1	5
35	Pedro Velho	2	3	4	3	3	2	5	1	2	1
36	Caraúbas	2	3	4	3	3	2	5	1	2	1
37	Poço de Pedras	2	3	4	3	3	2	5	1	2	1
38	Guarita	2	4	4	3	3	4	5	2	2	3
39	Umburana/Sumé	1	6	3	1	2	6	1	6	5	4
40	Jatobá/Sumé	1	2	3	5	1	3	1	4	6	2
41	Gangorra/Sumé	1	6	3	1	5	6	1	6	5	5

Anexo C.17 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 2) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
	Albino	2	2	2	2	2	2	2	2	2	2
2	Arrojado	1	1	1	1	1	1	1	1	1	1
3	Bartolomeu I	2	2	2	2	2	2	2	2	2	2
4	Cachoeira Alves	1	1	1	1	1	1	1	1	1	1
5	Chupadouro	2	2	2	2	2	2	2	2	2	2
6	Cochos	2	2	2	2	2	2	2	2	2	2
7	Emas	2	2	2	2	2	2	2	2	2	2
8	Engo Arcoverde	1	1	1	1	1	1	1	1	1	1
9	Engo Ávidos	1	1	1	1	1	1	1	1	1	1
10	Epitácio Pessoa	1	1	1	1	1	1	1	1	1	1
11	Farinha	1	1	1	1	1	1	1	1	1	1
12	Frutuoso II	2	2	2	2	2	2	2	2	2	2
13	Gamela	2	2	2	2	2	2	2	2	2	2
14	Jatobá I	1	1	1	1	1	1	1	1	1	1
15	Queimadas	2	2	2	2	2	2	2	2	2	2
16	Riac. dos Cavalos	1	1	1	1	1	1	1	1	1	1
17	Serra Vermelha	2	2	2	2	2	2	2	2	2	2
18	Tamanduá I	2	2	2	2	2	2	2	2	2	2
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	1	1	1	1	1	1	1	1	1	1
21	Campos	2	2	2	2	2	2	2	2	2	2
22	Cordeiro	1	1	1	1	1	1	1	1	1	1
23	Santo Antonio	2	2	2	2	2	2	2	2	2	2
24	São Domingos	2	2	2	2	2	2	2	2	2	2
25	Livramento	2	2	2	2	2	2	2	2	2	2
26	Namorados	2	2	2	2	2	2	2	2	2	2
27	Soledade	2	2	2	2	2	2	2	2	2	2
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	1	1	1	1	1	1	1	1	1	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	1	1	1	1	1	1	1	1	1	1
35	Pedro Velho	1	1	1	1	1	1	1	1	1	1
36	Caraúbas	1	1	1	1	1	1	1	1	1	1
37	Poço de Pedras	1	1	1	1	1	1	1	1	1	1
38	Guarita	1	1	1	1	1	1	1	1	1	1
39	Umburana/Sumé	2	2	2	2	2	2	2	2	2	2
40	Jatobá/Sumé	2	2	2	2	2	2	2	2	2	2
41	Gangorra/Sumé	2	2	2	2	2	2	2	2	2	2

Anexo C.18 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 3) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	3	3	3	3	3	3	3	3	3	3
2	Arrojado	2	2	2	2	2	2	2	2	2	2
3	Bartolomeu I	2	2	2	2	2	3	3	3	2	3
4	Cachoeira Alves	2	2	2	2	2	2	2	2	2	2
5	Chupadouro	2	2	2	2	2	3	3	3	2	3
6	Cochos	2	2	2	2	2	3	3	3	2	3
7	Emas	2	2	2	2	2	3	3	3	2	3
8	Engo Arcoverde	2	2	2	2	2	1	1	1	2	1
9	Engo Ávidos	2	2	2	2	2	1	1	1	2	1
10	Epitácio Pessoa	1	1	1	1	1	2	2	2	1	2
11	Farinha	2	2	2	2	2	1	1	1	2	1
12	Frutuoso II	2	2	2	2	2	3	3	3	2	3
13	Gamela	3	3	3	3	3	3	3	3	3	3
14	Jatobá I	2	2	2	2	2	1	1	1	2	1
15	Queimadas	2	2	2	2	3	3	3	3	2	3
16	Riac. dos Cavalos	2	2	2	2	2	1	1	1	2	1
17	Serra Vermelha	2	2	2	2	2	3	3	3	2	3
18	Tamanduá I	3	3	3	3	3	3	3	3	3	3
19	Vazantes	2	2	2	2	2	3	3	3	2	3
20	Camalau	3	3	3	3	3	3	3	3	3	3
21	Campos	3	3	3	3	3	3	3	3	3	3
22	Cordeiro	1	1	1	1	1	2	2	2	1	2
23	Santo Antonio	3	3	3	3	3	3	3	3	3	3
24	São Domingos	3	3	3	3	3	3	3	3	3	3
25	Livramento	3	3	3	3	3	3	3	3	3	3
26	Namorados	3	3	3	3	3	3	3	3	3	3
27	Soledade	3	3	3	3	3	3	3	3	3	3
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	1	1	1	1	2	2	2	1	2
31	Emas	2	2	2	2	2	1	1	1	2	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	2	2	2	1	2
34	Faz. Alagamar	1	1	1	1	1	2	2	2	1	2
35	Pedro Velho	1	1	1	1	1	2	2	2	1	2
36	Caraúbas	1	1	1	1	1	2	2	2	1	2
37	Poço de Pedras	1	1	1	1	1	2	2	2	1	2
38	Guarita	1	1	1	1	1	2	2	2	1	2
39	Umburana/Sumé	3	3	3	3	3	3	3	3	3	3
40	Jatobá/Sumé	3	3	3	3	3	3	3	3	3	3
41	Gangorra/Sumé	3	3	3	3	3	3	3	3	3	3

Anexo C.19 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 4) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	3	3	3	3	3	3	3	3	3	3
2	Arrojado	1	2	2	1	2	2	1	1	2	1
3	Bartolomeu I	3	3	3	3	4	3	3	3	3	3
4	Cachoeira Alves	1	2	2	1	2	2	1	1	2	1
5	Chupadouro	3	3	3	3	4	3	3	3	3	3
6	Cochos	4	4	4	4	4	4	4	4	4	4
7	Emas	4	4	4	4	4	4	4	4	4	4
8	Engo Arcoverde	1	2	2	2	2	2	1	2	2	2
9	Engo Ávidos	1	1	1	1	2	1	1	1	2	1
10	Epitácio Pessoa	2	1	1	1	1	1	2	1	1	1
11	Farinha	1	1	1	1	2	1	1	1	1	1
12	Frutuoso II	4	4	4	4	4	4	4	4	4	4
13	Gamela	4	4	4	4	4	4	4	4	4	4
14	Jatobá I	1	2	2	1	2	2	1	1	2	1
15	Queimadas	4	2	2	4	4	2	4	4	2	4
16	Riac. dos Cavalos	1	1	1	1	2	1	1	1	1	1
17	Serra Vermelha	4	4	4	4	4	4	4	4	4	4
18	Tamanduá I	3	3	3	3	3	3	3	3	3	3
19	Vazantes	1	2	2	1	2	2	1	1	2	1
20	Camalau	1	2	2	1	2	2	1	1	2	1
21	Campos	4	3	3	4	4	4	4	4	3	4
22	Cordeiro	1	1	1	1	2	1	1	1	2	1
23	Santo Antonio	3	2	2	2	3	2	3	2	2	2
24	São Domingos	4	2	2	4	4	2	4	4	2	4
25	Livramento	3	3	3	3	3	3	3	3	3	3
26	Namorados	3	3	3	3	4	3	3	3	3	3
27	Soledade	4	3	3	4	4	4	4	4	3	4
28	Antenor Navarro	1	1	1	1	2	1	1	1	1	1
29	Aparecida	1	1	1	1	2	1	1	1	1	1
30	Piancó	1	1	1	1	2	1	1	1	1	1
31	Emas	1	1	1	1	2	1	1	1	2	1
32	Patos	1	1	1	1	2	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	2	1	1	1	1	1
34	Faz. Alagamar	1	1	1	1	2	1	1	1	1	1
35	Pedro Velho	1	1	1	1	2	1	1	1	1	1
36	Caraúbas	1	1	1	1	2	1	1	1	1	1
37	Poço de Pedras	1	1	1	1	2	1	1	1	1	1
38	Guarita	2	1	1	1	1	1	2	1	1	1
39	Umburana/Sumé	3	3	3	3	3	3	3	3	3	3
40	Jatobá/Sumé	3	3	3	3	4	3	3	3	3	3
41	Gangorra/Sumé	4	2	2	4	4	2	4	4	2	4

Anexo C.20 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 6) – Cenário I

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	5	6	6	6	6	5	6	6	6	6
2	Arrojado	2	4	4	4	4	2	4	4	4	4
3	Bartolomeu I	6	6	6	6	6	6	6	6	6	6
4	Cachoeira Alves	2	4	4	4	4	2	4	4	4	4
5	Chupadouro	6	6	6	6	6	6	6	6	6	6
6	Cochos	4	5	5	4	5	4	5	5	5	4
7	Emas	6	5	5	5	5	6	5	5	5	5
8	Engo Arcoverde	3	3	2	3	3	3	2	3	3	3
9	Engo Ávidos	2	2	4	2	2	2	1	1	4	2
10	Epitácio Pessoa	1	1	1	1	1	1	1	2	1	1
11	Farinha	2	2	1	2	2	2	1	1	2	2
12	Frutuoso II	6	5	5	5	5	6	5	5	5	5
13	Gamela	6	5	5	5	5	6	5	5	5	5
14	Jatobá I	2	4	4	4	4	2	4	4	4	4
15	Queimadas	6	4	4	4	4	6	4	4	5	4
16	Riac. dos Cavalos	2	2	1	2	2	2	1	1	2	2
17	Serra Vermelha	4	5	5	4	5	4	5	5	5	4
18	Tamanduá I	5	6	3	6	6	5	3	6	6	6
19	Vazantes	2	4	4	4	4	2	4	4	4	4
20	Camalau	2	4	4	4	4	2	4	4	4	4
21	Campos	6	6	6	6	6	6	6	5	5	6
22	Cordeiro	2	2	4	2	2	2	1	1	4	2
23	Santo Antonio	3	3	2	3	3	3	2	3	3	3
24	São Domingos	6	4	4	4	4	6	4	4	5	4
25	Livramento	5	6	3	6	6	5	3	6	6	6
26	Namorados	6	6	6	6	6	6	6	6	6	6
27	Soledade	6	6	6	6	6	6	6	5	5	6
28	Antenor Navarro	2	2	1	2	2	2	1	1	2	2
29	Aparecida	2	2	1	2	2	2	1	1	2	2
30	Piancó	2	2	1	2	2	2	1	1	4	2
31	Emas	2	2	4	2	2	2	1	1	4	2
32	Patos	2	2	1	2	2	2	1	1	2	2
33	Serra Negra Norte	2	2	1	2	2	2	1	1	2	2
34	Faz. Alagamar	2	4	4	4	4	2	4	4	4	4
35	Pedro Velho	2	2	1	2	2	2	1	1	4	2
36	Caraúbas	2	2	1	2	2	2	1	1	4	2
37	Poço de Pedras	2	2	1	2	2	2	1	1	2	2
38	Guarita	1	1	1	1	1	1	1	2	1	1
39	Umburana/Sumé	5	6	6	6	6	5	6	6	6	6
40	Jatobá/Sumé	6	6	6	6	6	6	6	6	6	6
41	Gangorra/Sumé	6	4	4	4	4	6	4	4	5	4

Anexo C.21 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 2) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	2	2	2	1	1	2	2	2	1
2	Arrojado	1	1	1	1	2	2	1	1	1	2
3	Bartolomeu I	2	2	2	2	1	1	2	2	2	1
4	Cachoeira Alves	1	1	1	1	2	2	1	1	1	2
5	Chupadouro	2	2	2	2	1	1	2	2	2	1
6	Cochos	2	2	2	2	1	1	2	2	2	1
7	Emas	2	2	2	2	1	1	2	2	2	1
8	Engo Arcoverde	1	1	1	1	2	2	1	1	1	2
9	Engo Ávidos	1	1	1	1	2	2	1	1	1	2
10	Epitácio Pessoa	1	1	1	1	2	2	1	1	1	2
11	Farinha	1	1	1	1	2	2	1	1	1	2
12	Frutuoso II	2	2	2	2	1	1	2	2	2	1
13	Gamela	2	2	2	2	1	1	2	2	2	1
14	Jatobá I	1	1	1	1	2	2	1	1	1	2
15	Queimadas	2	2	2	2	1	1	2	2	2	1
16	Riac. dos Cavalos	1	1	1	1	2	2	1	1	1	2
17	Serra Vermelha	2	2	2	2	1	1	2	2	2	1
18	Tamanduá I	2	2	2	2	1	1	2	2	2	1
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	1	1	1	1	1	1	1	1	1	1
21	Campos	2	2	2	2	1	1	2	2	2	1
22	Cordeiro	1	1	1	1	1	1	1	1	1	1
23	Santo Antonio	2	2	2	2	1	1	2	2	2	1
24	São Domingos	2	2	2	2	1	1	2	2	2	1
25	Livramento	2	2	2	2	1	1	2	2	2	1
26	Namorados	2	2	2	2	1	1	2	2	2	1
27	Soledade	2	2	2	2	1	1	2	2	2	1
28	Antenor Navarro	1	1	1	1	2	2	1	1	1	2
29	Aparecida	1	1	1	1	2	2	1	1	1	2
30	Piancó	1	1	1	1	2	2	1	1	1	2
31	Emas	1	1	1	1	2	2	1	1	1	2
32	Patos	1	1	1	1	2	2	1	1	1	2
33	Serra Negra Norte	1	1	1	1	2	2	1	1	1	2
34	Faz. Alagamar	1	1	1	1	1	1	1	1	1	1
35	Pedro Velho	1	1	1	1	2	2	1	1	1	2
36	Caraúbas	1	1	1	1	2	2	1	1	1	2
37	Poço de Pedras	1	1	1	1	2	2	1	1	1	2
38	Guarita	1	1	1	1	2	2	1	1	1	2
39	Umburana/Sumé	2	2	2	2	1	1	2	2	2	1
40	Jatobá/Sumé	2	2	2	2	1	1	2	2	2	1
41	Gangorra/Sumé	2	2	2	2	1	1	2	2	2	1

Anexo C.22 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 3) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	2	2	2	2	2	2	2	2	2
2	Arrojado	1	1	1	1	1	1	1	1	1	1
3	Bartolomeu I	2	2	2	2	2	2	2	2	2	2
4	Cachoeira Alves	1	1	1	1	1	1	1	1	1	1
5	Chupadouro	2	2	2	2	2	2	2	2	2	2
6	Cochos	1	1	1	1	1	1	1	1	1	1
7	Emas	2	2	2	2	2	2	2	2	2	2
8	Engo Arcoverde	3	1	1	1	1	3	1	3	3	3
9	Engo Ávidos	3	3	3	3	3	3	3	3	3	3
10	Epitácio Pessoa	3	3	3	3	3	3	3	3	3	3
11	Farinha	3	3	3	3	3	3	3	3	3	3
12	Frutuoso II	2	2	2	2	2	2	2	2	2	2
13	Gamela	2	2	2	2	2	2	2	2	2	2
14	Jatobá I	3	3	3	3	3	3	3	3	3	3
15	Queimadas	2	2	2	2	2	2	2	2	2	2
16	Riac. dos Cavalos	3	1	1	1	1	3	1	3	3	3
17	Serra Vermelha	1	1	1	1	1	1	1	1	1	1
18	Tamanduá I	2	2	2	2	2	2	2	2	2	2
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	1	1	1	1	1	1	1	1	1	1
21	Campos	2	2	2	2	2	2	2	2	2	2
22	Cordeiro	1	1	1	1	1	1	1	1	1	1
23	Santo Antonio	2	2	2	2	2	2	2	2	2	2
24	São Domingos	1	1	1	1	1	1	1	1	1	1
25	Livramento	2	2	2	2	2	2	2	2	2	2
26	Namorados	2	2	2	2	2	2	2	2	2	2
27	Soledade	2	2	2	2	2	2	2	2	2	2
28	Antenor Navarro	3	3	3	3	3	3	3	3	3	3
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	1	1	1	1	1	1	1	1	1	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	1	1	1	1	1	1	1	1	1	1
35	Pedro Velho	1	1	1	1	1	1	1	1	1	1
36	Caraúbas	1	1	1	1	1	1	1	1	1	1
37	Poço de Pedras	1	1	1	1	1	1	1	1	1	1
38	Guarita	3	3	3	3	3	3	3	3	3	3
39	Umburana/Sumé	2	2	2	2	2	2	2	2	2	2
40	Jatobá/Sumé	2	2	2	2	2	2	2	2	2	2
41	Gangorra/Sumé	2	2	2	2	2	2	2	2	2	2

Anexo C.23 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 4) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	4	4	4	4	4	4	4	4	4	4
2	Arrojado	2	2	1	1	2	1	2	1	1	1
3	Bartolomeu I	3	4	4	3	3	3	4	4	4	3
4	Cachoeira Alves	1	3	1	1	1	1	3	1	1	1
5	Chupadouro	4	4	4	4	4	4	4	4	4	4
6	Cochos	3	3	3	3	3	3	3	3	3	3
7	Emas	4	4	4	4	4	4	4	4	4	4
8	Engo Arcoverde	2	2	1	1	2	2	2	1	1	2
9	Engo Ávidos	2	2	1	1	2	2	2	1	1	2
10	Epitácio Pessoa	2	2	2	2	2	2	2	2	2	2
11	Farinha	2	2	2	2	2	2	2	2	2	2
12	Frutuoso II	2	4	4	3	3	3	4	4	4	3
13	Gamela	3	4	4	4	4	4	4	4	4	4
14	Jatobá I	2	2	2	1	2	2	2	2	2	2
15	Queimadas	3	3	3	3	3	3	3	3	3	3
16	Riac. dos Cavalos	2	2	1	1	2	1	2	1	1	1
17	Serra Vermelha	3	3	3	3	3	3	3	3	3	3
18	Tamanduá I	4	4	4	4	4	4	4	4	4	4
19	Vazantes	3	3	3	3	3	3	3	3	3	3
20	Camalau	3	3	3	3	3	3	3	3	3	3
21	Campos	4	4	4	4	4	4	4	4	4	4
22	Cordeiro	3	3	3	3	3	3	3	3	3	3
23	Santo Antonio	3	3	4	3	3	3	4	4	4	3
24	São Domingos	3	3	3	3	3	3	3	3	3	3
25	Livramento	4	4	4	4	4	4	4	4	4	4
26	Namorados	4	4	4	4	4	4	4	4	4	4
27	Soledade	4	4	4	4	4	4	4	4	4	4
28	Antenor Navarro	2	2	2	2	2	2	2	2	2	2
29	Aparecida	1	1	1	1	2	1	1	1	1	1
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	2	2	2	1	2	1	2	2	2	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	3	3	3	3	3	3	3	3	3	3
35	Pedro Velho	1	3	3	3	3	3	3	3	3	3
36	Caraúbas	3	3	3	3	3	3	3	3	3	3
37	Poço de Pedras	1	3	3	3	3	3	3	3	3	3
38	Guarita	2	2	2	2	2	2	2	2	2	2
39	Umburana/Sumé	4	4	4	4	4	4	4	4	4	4
40	Jatobá/Sumé	4	4	4	4	4	4	4	4	4	4
41	Gangorra/Sumé	3	3	4	3	3	3	4	4	4	3

Anexo C.24 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 6) – Cenário II

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	5	5	5	5	5	5	4	5	5	5
2	Arrojado	2	2	2	2	2	2	2	2	2	2
3	Bartolomeu I	5	5	5	5	5	5	4	5	5	5
4	Cachoeira Alves	2	2	2	2	2	2	3	2	2	2
5	Chupadouro	5	5	6	5	5	5	5	5	6	6
6	Cochos	4	4	4	4	4	4	3	4	4	4
7	Emas	5	5	5	5	5	5	4	5	5	5
8	Engo Arcoverde	2	2	2	2	2	2	2	2	2	2
9	Engo Ávidos	2	2	2	3	2	2	2	2	2	2
10	Epitácio Pessoa	3	3	3	3	3	3	2	3	3	3
11	Farinha	3	3	3	3	3	3	2	3	3	3
12	Frutuoso II	5	5	5	5	5	5	4	5	5	5
13	Gamela	6	6	6	6	6	5	5	6	6	6
14	Jatobá I	3	2	2	3	2	2	2	2	2	2
15	Queimadas	4	4	5	4	4	4	4	5	4	4
16	Riac. dos Cavalos	2	2	2	3	2	2	2	2	2	2
17	Serra Vermelha	4	4	4	4	4	4	3	4	4	4
18	Tamanduá I	6	6	6	6	6	5	5	6	6	6
19	Vazantes	4	4	4	4	4	4	3	4	4	4
20	Camalau	4	4	4	4	4	4	3	4	4	4
21	Campos	5	5	6	5	5	5	5	5	6	6
22	Cordeiro	4	4	4	4	4	4	3	4	4	4
23	Santo Antonio	5	4	5	2	5	6	6	5	4	4
24	São Domingos	4	4	4	4	4	4	3	4	4	4
25	Livramento	6	6	6	6	6	5	5	6	6	6
26	Namorados	6	6	6	6	6	5	5	6	6	6
27	Soledade	5	5	5	5	5	5	4	5	5	5
28	Antenor Navarro	3	3	3	3	3	3	2	3	3	3
29	Aparecida	2	2	2	1	2	2	2	2	2	2
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	2	2	2	2	2	2	2	2	2	2
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	4	4	4	4	4	4	3	4	4	4
35	Pedro Velho	4	4	4	4	4	4	3	4	4	4
36	Caraúbas	4	4	4	4	4	4	3	4	4	4
37	Poço de Pedras	4	4	4	4	4	4	3	4	4	4
38	Guarita	3	3	3	3	3	3	2	3	3	3
39	Umburana/Sumé	5	5	6	5	5	5	5	5	6	6
40	Jatobá/Sumé	5	5	5	5	5	5	4	5	5	5
41	Gangorra/Sumé	5	5	5	5	5	5	4	5	5	5

Anexo C.25 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 2) - Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	1	2	1	1	1	2	2	2	2
2	Arrojado	1	1	1	2	2	1	1	1	1	1
3	Bartolomeu I	1	2	1	2	2	2	1	1	1	1
4	Cachoeira Alves	1	2	1	2	2	1	1	1	1	1
5	Chupadouro	1	2	2	2	2	2	2	2	1	1
6	Cochos	1	1	1	2	2	1	1	1	1	1
7	Emas	2	1	2	1	1	1	2	2	2	2
8	Engo Arcoverde	1	1	1	1	2	1	1	1	2	2
9	Engo Ávidos	1	1	1	1	2	1	1	1	1	1
10	Epitácio Pessoa	1	2	1	2	2	2	2	2	2	2
11	Farinha	1	1	1	1	2	1	1	1	2	2
12	Frutuoso II	2	1	2	1	1	1	2	2	1	1
13	Gamela	2	1	2	1	1	1	2	2	1	1
14	Jatobá I	1	1	1	2	2	1	1	1	2	2
15	Queimadas	2	1	2	1	1	1	2	2	2	2
16	Riac. dos Cavalos	1	1	1	1	2	1	1	1	1	1
17	Serra Vermelha	1	2	1	2	2	2	1	1	1	1
18	Tamanduá I	2	1	2	1	1	1	2	2	2	2
19	Vazantes	1	2	1	2	2	1	1	1	1	1
20	Camalau	2	1	2	1	1	1	2	2	2	2
21	Campos	2	1	2	1	1	1	2	2	2	2
22	Cordeiro	2	1	2	1	1	1	2	2	2	2
23	Santo Antonio	2	2	2	2	2	2	2	2	1	1
24	São Domingos	2	1	2	1	1	1	2	2	2	2
25	Livramento	1	2	2	2	2	2	2	2	1	1
26	Namorados	2	2	2	2	2	2	2	2	1	1
27	Soledade	2	1	2	1	1	1	2	2	2	2
28	Antenor Navarro	1	1	1	1	2	1	1	1	2	2
29	Aparecida	1	2	1	2	2	2	1	1	1	1
30	Piancó	1	2	1	2	2	2	1	1	1	1
31	Emas	2	1	2	1	1	1	2	2	2	2
32	Patos	1	1	1	1	2	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	2	1	1	1	1	1
34	Faz. Alagamar	2	1	2	1	1	1	2	2	2	2
35	Pedro Velho	2	1	2	1	1	1	2	2	2	2
36	Caraúbas	1	2	2	2	2	2	1	2	1	1
37	Poço de Pedras	1	1	1	2	2	1	1	1	2	2
38	Guarita	1	2	1	2	2	2	1	2	1	1
39	Umburana/Sumé	2	2	2	2	2	2	2	2	1	1
40	Jatobá/Sumé	2	1	2	2	1	1	2	2	2	2
41	Gangorra/Sumé	2	1	2	2	1	1	2	2	2	2

Anexo C.26 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 3) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	3	3	3	3	3	2	3	3	2	2
2	Arrojado	1	1	1	1	1	1	1	1	1	1
3	Bartolomeu I	1	1	1	1	1	1	1	1	1	1
4	Cachoeira Alves	1	1	1	1	1	1	1	1	1	1
5	Chupadouro	3	3	3	3	3	3	3	3	3	2
6	Cochos	1	1	1	1	1	1	1	1	1	1
7	Emas	2	3	2	1	3	2	2	2	2	1
8	Engo Arcoverde	1	1	1	1	1	1	1	1	1	1
9	Engo Ávidos	1	1	1	1	1	1	1	1	1	1
10	Epitácio Pessoa	1	2	1	2	2	3	1	1	3	3
11	Farinha	1	1	1	1	1	1	1	1	1	1
12	Frutuoso II	2	2	2	1	3	2	2	2	2	1
13	Gamela	2	2	2	1	3	2	2	2	2	1
14	Jatobá I	1	1	1	1	1	1	1	1	1	1
15	Queimadas	3	3	3	3	3	2	3	3	2	2
16	Riac. dos Cavalos	1	1	1	1	1	1	1	1	1	1
17	Serra Vermelha	1	1	1	1	1	1	1	1	1	1
18	Tamanduá I	3	3	3	3	3	2	3	3	2	2
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	3	3	3	3	3	2	3	3	2	2
21	Campos	2	3	2	1	3	2	2	2	2	1
22	Cordeiro	2	3	2	1	3	2	2	2	2	1
23	Santo Antonio	3	3	3	3	3	3	3	3	3	2
24	São Domingos	3	3	3	3	3	2	3	3	2	2
25	Livramento	3	3	3	3	3	3	3	3	3	2
26	Namorados	3	3	3	3	3	3	3	3	3	2
27	Soledade	2	3	2	1	3	2	2	2	2	1
28	Antenor Navarro	1	1	2	1	1	1	1	1	1	1
29	Aparecida	1	2	1	2	1	3	1	1	3	3
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	3	3	3	3	3	2	3	3	2	2
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	2	3	2	1	3	2	2	2	2	1
35	Pedro Velho	2	3	2	1	3	2	2	2	2	1
36	Caraúbas	1	2	3	2	2	3	3	1	3	3
37	Poço de Pedras	1	1	1	1	1	1	2	1	1	1
38	Guarita	1	2	1	2	2	3	1	1	3	3
39	Umburana/Sumé	3	3	3	3	3	3	3	3	3	2
40	Jatobá/Sumé	3	3	3	3	3	2	3	3	2	2
41	Gangorra/Sumé	3	3	3	3	3	2	3	3	2	2

Anexo C.27 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 4) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	4	4	4	4	4	4	4	4	4	4
2	Arrojado	1	1	1	1	1	1	1	1	1	1
3	Bartolomeu I	1	1	1	1	1	1	1	1	1	1
4	Cachoeira Alves	1	1	1	1	1	1	1	1	1	1
5	Chupadouro	4	4	4	4	4	4	4	4	4	4
6	Cochos	1	1	1	1	1	1	1	1	1	1
7	Emas	2	2	4	4	2	2	4	2	4	2
8	Engo Arcoverde	1	1	1	1	1	1	1	1	1	1
9	Engo Ávidos	1	1	1	1	1	1	1	1	1	1
10	Epitácio Pessoa	3	3	3	3	3	3	3	3	3	3
11	Farinha	1	1	1	1	1	1	1	1	1	1
12	Frutuoso II	2	2	4	2	2	2	4	2	4	2
13	Gamela	2	2	2	2	2	2	2	2	2	2
14	Jatobá I	1	1	1	1	1	1	1	1	1	1
15	Queimadas	4	2	4	4	4	4	4	4	4	4
16	Riac. dos Cavalos	1	1	1	1	1	1	1	1	1	1
17	Serra Vermelha	1	1	1	1	1	1	1	1	1	1
18	Tamanduá I	4	4	4	4	4	4	4	4	4	4
19	Vazantes	1	1	1	1	1	1	1	1	1	1
20	Camalau	4	2	4	4	4	4	4	4	4	4
21	Campos	2	2	2	2	2	2	2	2	2	2
22	Cordeiro	2	2	4	4	2	2	4	2	4	2
23	Santo Antonio	4	4	4	4	4	4	4	4	4	4
24	São Domingos	4	2	4	4	4	4	4	4	4	4
25	Livramento	4	4	4	4	4	4	4	4	4	4
26	Namorados	4	4	4	4	4	4	4	4	4	4
27	Soledade	2	2	2	2	2	2	2	2	2	2
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	3	3	3	1	3	1	3	1	3	3
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	4	2	4	4	4	4	4	4	4	4
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	2	2	2	2	2	2	2	2	2	2
35	Pedro Velho	2	2	4	2	2	2	4	2	4	2
36	Caraúbas	3	3	3	3	3	3	3	3	3	3
37	Poço de Pedras	1	1	2	2	2	2	2	2	2	1
38	Guarita	3	3	3	3	3	3	3	3	3	3
39	Umburana/Sumé	4	4	4	4	4	4	4	4	4	4
40	Jatobá/Sumé	4	4	4	4	4	4	4	4	4	4
41	Gangorra/Sumé	4	4	4	4	4	4	4	4	4	4

Anexo C.28 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 6) – Cenário III

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	5	5	5	6	5	5	5	5	5	5
2	Arrojado	1	2	2	2	2	2	2	2	2	1
3	Bartolomeu I	1	2	2	2	2	2	2	2	2	1
4	Cachoeira Alves	1	2	2	2	2	2	2	2	2	1
5	Chupadouro	6	6	6	6	6	6	6	6	6	6
6	Cochos	1	2	2	2	2	2	2	2	2	1
7	Emas	3	3	3	5	5	5	5	5	5	3
8	Engo Arcoverde	2	1	1	1	1	1	1	1	1	2
9	Engo Ávidos	1	2	2	2	2	2	2	2	2	1
10	Epitácio Pessoa	4	4	4	4	4	4	4	4	4	4
11	Farinha	1	1	2	1	1	1	1	1	1	1
12	Frutuoso II	3	3	3	5	3	5	3	3	3	3
13	Gamela	2	3	1	3	3	3	3	3	3	2
14	Jatobá I	1	2	2	2	2	2	2	2	2	1
15	Queimadas	5	5	5	5	5	5	5	5	5	5
16	Riac. dos Cavalos	1	1	2	1	1	1	1	1	1	1
17	Serra Vermelha	1	2	2	2	2	2	2	2	2	1
18	Tamanduá I	5	6	6	6	6	6	6	6	6	5
19	Vazantes	1	2	2	2	2	2	2	2	2	1
20	Camalau	5	5	5	5	5	5	5	5	5	5
21	Campos	2	3	1	3	3	3	3	3	3	2
22	Cordeiro	3	3	3	5	5	5	5	3	3	3
23	Santo Antonio	6	6	6	6	6	6	6	6	6	6
24	São Domingos	5	5	5	5	5	5	5	5	5	5
25	Livramento	6	6	6	6	6	6	6	6	6	6
26	Namorados	6	6	6	6	6	6	6	6	6	6
27	Soledade	2	3	1	3	3	3	3	3	3	2
28	Antenor Navarro	2	1	1	1	1	1	1	1	1	2
29	Aparecida	4	4	4	4	4	4	4	4	4	4
30	Piancó	1	2	2	2	2	2	2	2	2	1
31	Emas	5	5	5	5	5	5	5	5	5	5
32	Patos	1	1	2	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	2	1	1	1	1	1	1	1
34	Faz. Alagamar	2	3	1	3	3	3	3	3	3	2
35	Pedro Velho	3	3	3	5	3	5	3	3	3	3
36	Caraúbas	4	4	4	4	4	4	4	4	4	4
37	Poço de Pedras	2	1	1	1	1	1	1	1	1	2
38	Guarita	4	4	4	4	4	4	4	4	4	4
39	Umburana/Sumé	6	6	6	6	6	6	6	6	6	6
40	Jatobá/Sumé	5	6	6	6	6	6	6	6	6	5
41	Gangorra/Sumé	5	5	5	6	5	5	5	5	5	5

Anexo C.29 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 2) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	2	2	2	2	2	2	2	2	2	2
2	Arrojado	2	2	2	2	2	2	2	2	2	2
3	Bartolomeu I	2	2	2	2	2	2	2	2	2	2
4	Cachoeira Alves	2	2	2	2	2	2	2	2	2	2
5	Chupadouro	2	2	2	2	2	2	2	2	2	2
6	Cochos	2	2	2	2	2	2	2	2	2	2
7	Emas	2	2	2	2	2	2	2	2	2	2
8	Engo Arcoverde	2	2	2	2	2	2	2	2	2	2
9	Engo Ávidos	2	2	2	2	2	2	2	2	2	2
10	Epitácio Pessoa	1	1	1	1	1	1	1	1	1	1
11	Farinha	1	1	1	1	1	1	1	1	1	1
12	Frutuoso II	2	2	2	2	2	2	2	2	2	2
13	Gamela	2	2	2	2	2	2	2	2	2	2
14	Jatobá I	2	2	2	2	2	2	2	2	2	2
15	Queimadas	2	2	2	2	2	2	2	2	2	2
16	Riac. dos Cavalos	2	2	2	2	2	2	2	2	2	2
17	Serra Vermelha	2	2	2	2	2	2	2	2	2	2
18	Tamanduá I	2	2	2	2	2	2	2	2	2	2
19	Vazantes	2	2	2	2	2	2	2	2	2	2
20	Camalau	2	2	2	2	2	2	2	2	2	2
21	Campos	2	2	2	2	2	2	2	2	2	2
22	Cordeiro	1	1	1	1	1	1	1	1	1	1
23	Santo Antonio	2	2	2	2	2	2	2	2	2	2
24	São Domingos	2	2	2	2	2	2	2	2	2	2
25	Livramento	2	2	2	2	2	2	2	2	2	2
26	Namorados	2	2	2	2	2	2	2	2	2	2
27	Soledade	2	2	2	2	2	2	2	2	2	2
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	1	1	1	1	1	1	1	1	1
31	Emas	2	2	2	2	2	2	2	2	2	2
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	1	1	1	1	1	1	1
34	Faz. Alagamar	1	1	1	1	1	1	1	1	1	1
35	Pedro Velho	1	1	1	1	1	1	1	1	1	1
36	Caraúbas	1	1	1	1	1	1	1	1	1	1
37	Poço de Pedras	1	1	1	1	1	1	1	1	1	1
38	Guarita	1	1	1	1	1	1	1	1	1	1
39	Umburana/Sumé	2	2	2	2	2	2	2	2	2	2
40	Jatobá/Sumé	2	2	2	2	2	2	2	2	2	2
41	Gangorra/Sumé	2	2	2	2	2	2	2	2	2	2

Anexo C.30 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 3) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	3	3	3	3	3	3	3	3	3	3
2	Arrojado	2	2	2	2	2	2	2	2	2	2
3	Bartolomeu I	2	2	2	3	3	3	3	3	3	3
4	Cachoeira Alves	2	2	2	2	2	2	2	2	2	2
5	Chupadouro	2	2	2	3	3	3	3	3	3	3
6	Cochos	2	2	2	3	3	3	3	3	3	3
7	Emas	2	2	2	3	3	3	3	3	3	3
8	Engo Arcoverde	2	2	2	1	1	1	1	1	1	1
9	Engo Ávidos	2	2	2	1	1	1	1	1	1	1
10	Epitácio Pessoa	1	1	1	2	2	2	2	2	2	2
11	Farinha	2	2	2	1	1	1	1	1	1	1
12	Frutuoso II	2	2	2	3	3	3	3	3	3	3
13	Gamela	3	3	3	3	3	3	3	3	3	3
14	Jatobá I	2	2	2	1	1	1	1	1	1	1
15	Queimadas	2	3	3	3	3	3	3	3	3	3
16	Riac. dos Cavalos	2	2	2	1	1	1	1	1	1	1
17	Serra Vermelha	2	2	2	3	3	3	3	3	3	3
18	Tamanduá I	3	3	3	3	3	3	3	3	3	3
19	Vazantes	2	2	2	3	3	3	3	3	3	3
20	Camalau	3	3	3	3	3	3	3	3	3	3
21	Campos	3	3	3	3	3	3	3	3	3	3
22	Cordeiro	1	1	1	2	2	2	2	2	2	2
23	Santo Antonio	3	3	3	3	3	3	3	3	3	3
24	São Domingos	3	3	3	3	3	3	3	3	3	3
25	Livramento	3	3	3	3	3	3	3	3	3	3
26	Namorados	3	3	3	3	3	3	3	3	3	3
27	Soledade	3	3	3	3	3	3	3	3	3	3
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	1	1	2	2	2	2	2	2	2
31	Emas	2	2	2	1	1	1	1	1	1	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	1	1	2	2	2	2	2	2	2
34	Faz. Alagamar	1	1	1	2	2	2	2	2	2	2
35	Pedro Velho	1	1	1	2	2	2	2	2	2	2
36	Caraúbas	1	1	1	2	2	2	2	2	2	2
37	Poço de Pedras	1	1	1	2	2	2	2	2	2	2
38	Guarita	1	1	1	2	2	2	2	2	2	2
39	Umburana/Sumé	3	3	3	3	3	3	3	3	3	3
40	Jatobá/Sumé	3	3	3	3	3	3	3	3	3	3
41	Gangorra/Sumé	3	3	3	3	3	3	3	3	3	3

Anexo C.31 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 4) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	4	4	4	4	4	4	4	4	4	4
2	Arrojado	3	3	1	3	3	3	1	3	3	3
3	Bartolomeu I	3	3	3	3	3	3	3	3	3	3
4	Cachoeira Alves	3	3	1	3	3	3	1	3	3	3
5	Chupadouro	3	3	3	3	3	3	3	3	3	3
6	Cochos	3	3	3	3	3	3	3	3	3	3
7	Emas	3	3	3	3	3	3	3	3	3	3
8	Engo Arcoverde	3	3	3	3	3	3	3	3	3	3
9	Engo Ávidos	1	1	1	1	1	1	1	1	1	1
10	Epitácio Pessoa	2	2	2	2	2	2	2	2	2	2
11	Farinha	1	1	1	1	1	1	1	1	1	1
12	Frutuoso II	3	3	3	3	3	3	3	3	3	3
13	Gamela	4	4	4	4	4	4	4	4	4	4
14	Jatobá I	1	1	1	1	1	1	1	1	1	1
15	Queimadas	3	3	3	3	4	4	3	3	3	4
16	Riac. dos Cavalos	1	1	1	1	1	1	1	1	1	1
17	Serra Vermelha	3	3	3	3	3	3	3	3	3	3
18	Tamanduá I	4	4	4	4	4	4	4	4	4	4
19	Vazantes	3	3	3	3	3	3	3	3	3	3
20	Camalau	4	4	4	4	4	4	4	4	4	4
21	Campos	4	4	4	4	4	4	4	4	4	4
22	Cordeiro	2	2	2	2	2	2	2	2	2	2
23	Santo Antonio	4	4	4	4	4	4	4	4	4	4
24	São Domingos	4	4	4	4	4	4	4	4	4	4
25	Livramento	4	4	4	4	4	4	4	4	4	4
26	Namorados	4	4	4	4	4	4	4	4	4	4
27	Soledade	4	4	4	4	4	4	4	4	4	4
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	1	1
30	Piancó	1	2	2	1	2	1	2	1	1	1
31	Emas	1	1	1	1	1	1	1	1	1	1
32	Patos	1	1	1	1	1	1	1	1	1	1
33	Serra Negra Norte	1	2	2	1	2	1	2	1	1	1
34	Faz. Alagamar	2	2	2	2	2	2	2	2	2	2
35	Pedro Velho	2	2	2	2	2	2	2	2	2	2
36	Caraúbas	2	2	2	2	2	2	2	2	2	2
37	Poço de Pedras	2	2	2	2	2	2	2	2	2	2
38	Guarita	2	2	2	2	2	2	2	2	2	2
39	Umburana/Sumé	4	4	4	4	4	4	4	4	4	4
40	Jatobá/Sumé	4	4	4	4	4	4	4	4	4	4
41	Gangorra/Sumé	4	4	4	4	4	4	4	4	4	4

Anexo C.32 – Matriz de Simulações da Rede de Kohonen (Nº de Grupos = 6) – Cenário IV

	Bacias	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	Albino	6	6	6	6	6	6	6	6	6	6
2	Arrojado	3	3	5	3	4	4	3	4	5	3
3	Bartolomeu I	4	3	5	5	4	5	5	5	5	4
4	Cachoeira Alves	3	3	5	3	4	4	3	4	5	3
5	Chupadouro	4	3	5	5	4	5	5	5	5	4
6	Cochos	4	3	5	5	4	5	5	5	5	4
7	Emas	4	3	5	5	4	5	5	5	5	4
8	Engo Arcoverde	3	3	5	3	4	4	3	4	5	3
9	Engo Ávidos	3	1	4	3	1	4	3	4	1	3
10	Epitácio Pessoa	2	2	2	2	2	2	2	2	3	2
11	Farinha	1	1	4	3	1	1	1	4	1	1
12	Frutuoso II	4	3	5	5	4	5	5	5	5	4
13	Gamela	6	5	6	4	3	6	4	6	6	6
14	Jatobá I	3	1	4	3	1	4	3	4	1	3
15	Queimadas	5	4	6	6	5	5	3	5	6	5
16	Riac. dos Cavalos	3	1	4	3	1	4	3	4	1	3
17	Serra Vermelha	4	3	5	5	4	5	5	5	5	4
18	Tamanduá I	5	4	6	6	5	6	6	6	6	5
19	Vazantes	4	3	5	5	4	5	5	5	5	4
20	Camalau	6	5	3	4	3	3	4	6	4	6
21	Campos	6	5	6	4	3	6	4	6	6	6
22	Cordeiro	2	2	3	4	2	3	2	3	4	2
23	Santo Antonio	6	5	3	4	3	3	4	6	4	6
24	São Domingos	6	5	6	4	3	6	4	6	6	6
25	Livramento	5	4	6	6	5	6	6	6	6	5
26	Namorados	5	4	6	6	5	6	6	6	6	5
27	Soledade	6	5	6	4	3	6	4	6	6	6
28	Antenor Navarro	1	1	1	1	1	1	1	1	1	1
29	Aparecida	1	1	1	1	1	1	1	1	2	1
30	Piancó	1	2	1	1	1	1	1	1	2	1
31	Emas	3	1	4	3	1	4	3	4	1	3
32	Patos	1	1	1	1	1	1	1	1	2	1
33	Serra Negra Norte	1	2	1	1	1	1	1	1	2	1
34	Faz. Alagamar	2	2	3	4	2	3	2	3	4	2
35	Pedro Velho	2	2	3	2	2	3	2	3	4	2
36	Caraúbas	2	2	3	2	2	3	2	3	4	2
37	Poço de Pedras	2	2	3	2	2	3	2	3	4	2
38	Guarita	2	2	2	2	2	2	2	2	3	2
39	Umburana/Sumé	5	4	6	6	5	6	6	6	6	5
40	Jatobá/Sumé	6	6	6	6	6	6	6	6	6	6
41	Gangorra/Sumé	5	5	6	6	5	6	6	6	6	5

Anexo C.33 –**Simulações para escolher o tamanho da rede neural**

Simulações para decidir o melhor tamanho do mapa					
Configuração da rede	lattice: hexa	neigh:gaussian	training:size 1x2		
Cenário I k=16					
1º treinamento	2º treinamento	3º treinamento	4º treinamento	5º treinamento	
[3x4]	[4x5]	[4x6]	[4x7]	[5x5]	
QE: 0.487	QE: 0.443	QE: 0.424	QE: 0.405	QE: 0.417	
TE: 0.000	TE: 0.024	TE: 0.024	TE: 0.049	TE: 0.000	
6º treinamento	7º treinamento	8º treinamento	9º treinamento	10º treinamento	11º treinamento
[5x6]	[6x6]	[6x7]	[8x3]	[2x2]	[3x3]
QE: 0.397	QE: 0.387	QE: 0.359	QE: 0.413	QE: 0.542	QE: 0.494
TE: 0.000	TE: 0.000	TE: 0.024	TE: 0.000	TE: 0.000	TE: 0.000
Cenário II k=18					
1º treinamento	2º treinamento	3º treinamento	4º treinamento	5º treinamento	
[3x4]	[4x5]	[4x6]	[4x7]	[5x5]	
QE: 0.262	QE: 0.227	QE: 0.221	QE: 0.217	QE: 0.211	
TE: 0.000	TE: 0.073	TE: 0.000	TE: 0.073	TE: 0.000	
6º treinamento	7º treinamento	8º treinamento	9º treinamento	10º treinamento	11º treinamento
[5x6]	[6x6]	[6x7]	[8x3]	[2x2]	[3x3]
QE: 0.202	QE: 0.196	QE: 0.184	QE: 0.210	QE: 0.315	QE: 0.289
TE: 0.000	TE: 0.024	TE: 0.049	TE: 0.000	TE: 0.000	TE: 0.049
Cenário III CP = 05					
1º treinamento	2º treinamento	3º treinamento	4º treinamento	5º treinamento	
[3x4]	[4x5]	[4x6]	[4x7]	[5x5]	
QE: 1.612	QE: 1.427	QE: 1.369	QE: 1.324	QE: 1.334	
TE: 0.000	TE: 0.000	TE: 0.000	TE: 0.073	TE: 0.000	
6º treinamento	7º treinamento	8º treinamento	9º treinamento	10º treinamento	11º treinamento
[5x6]	[6x6]	[6x7]	[8x3]	[2x2]	[3x3]
QE: 1.274	QE: 1.161	QE: 1.149	QE: 1.344	QE: 1.879	QE: 1.686
TE: 0.049	TE: 0.000	TE: 0.024	TE: 0.049	TE: 0.000	TE: 0.000
Cenário IV Todos os atributos					
1º treinamento	2º treinamento	3º treinamento	4º treinamento	5º treinamento	
[3x4]	[4x5]	[4x6]	[4x7]	[5x5]	
QE: 0.918	QE: 0.850	QE: 0.828	QE: 0.801	QE: 0.820	
TE: 0.000	TE: 0.000	TE: 0.000	TE: 0.000	TE: 0.024	
6º treinamento	7º treinamento	8º treinamento	9º treinamento	10º treinamento	11º treinamento
[5x6]	[6x6]	[6x7]	[8x3]	[2x2]	[3x3]
QE: 0.767	QE: 0.770	QE: 0.723	QE: 0.825	QE: 1.023	QE: 0.947
TE: 0.000	TE: 0.000	TE: 0.000	TE: 0.024	TE: 0.000	TE: 0.000

ANEXO D: DADOS DAS BACIAS HIDROGRÁFICAS

Bacias	Características das bacias idrográficas																														
	A	Pr	L	Lt	Ld	Lm	Kc	Ke	Kf	Or	Rb	RL	Ra	Dd	Ct	IR	Le	SIN	Imax	Cmed	Lr	lr	Ip	IG	DS	P	E	L600	**SOLO 1**	**SOLO 2**	**SOLO 3**	**AE/A**
Albino	9,5	13,77	3,94	3,55	3,55	2,41	1,25	0,64	0,61	1,0	1,0	0,50	0,40	0,37	0,11	52,32	0,85	1,28	0,41	40,72	4,90	1,94	13,82	18,35	56,57	623,9	1644,4	22,20	0,00	0,00	0,86	0,03
Arrojado	30,1	35,77	12,59	12,02	56,55	2,92	1,65	0,37	0,23	4,0	3,0	0,34	0,65	1,54	0,11	522,69	1,56	2,04	6,79	40,86	15,37	2,39	24,48	18,84	114,28	712,4	1931,5	29,60	0,01	0,30	0,69	0,03
Bartolomeu I	59,5	36,05	9,28	13,56	92,73	6,41	1,31	0,58	0,69	4,0	4,0	0,58	0,19	1,56	0,08	342,98	1,92	1,75	5,80	40,98	13,49	4,41	27,85	12,81	98,83	678,6	1752,7	41,51	0,00	0,98	0,21	0,03
Cachoeira Alves	110,5	57,96	19,20	19,02	177,01	5,76	1,54	0,42	0,30	5,0	2,0	57,71	0,50	1,60	1,23	552,56	2,70	1,86	0,20	36,64	24,21	4,57	34,84	13,74	144,45	911,52	1931,5	33,04	0,00	0,67	0,33	0,02
Chupadouro	17,8	19,13	8,53	7,71	28,35	2,08	1,27	0,62	0,24	3,0	4,0	0,24	0,41	1,60	0,28	127,73	0,73	1,26	9,84	41,65	6,94	2,56	14,30	10,38	43,72	946,4	1931,5	29,31	0,00	0,38	0,61	0,06
Cochos	56,5	34,07	20,02	16,52	52,08	2,07	1,48	0,46	0,10	3,0	6,0	0,16	0,08	1,26	3,28	452,20	1,19	1,89	0,15	39,14	13,94	2,97	27,55	20,00	128,80	969,5	1875,5	37,00	0,00	1,00	0,00	0,02
Emas	35,0	23,62	12,71	10,50	26,36	2,75	1,12	0,80	0,22	3,0	5,0	0,64	0,17	0,75	0,17	354,08	1,55	1,86	0,21	40,94	8,83	3,97	31,60	35,09	207,57	687,6	1925,5	37,00	0,00	1,00	0,00	0,03
Engº Arcoverde	126,9	69,21	10,55	8,51	162,02	12,03	1,72	0,34	1,14	5,0	2,0	0,82	0,93	1,28	0,02	344,61	3,57	0,96	0,14	38,68	30,15	4,21	25,63	5,38	60,63	715,0	1903,8	38,96	0,03	0,98	0,00	0,05
Engº Ávidos	1009,5	178,00	51,05	48,27	1264,19	19,77	1,57	0,41	0,39	6,0	3,0	0,58	0,92	1,25	0,00	688,77	5,98	1,14	0,05	40,96	74,89	13,48	72,98	5,57	176,84	852,6	1853,5	34,01	0,00	0,73	0,28	0,03
Epitácio Pessoa	10659,0	747,50	183,10	182,82	7744,35	67,80	1,88	0,28	0,37	7,0	2,0	4,56	0,49	0,62	0,16	499,07	25,51	1,50	2,79	40,86	333,92	37,18	182,47	0,78	86,57	450,0	1968,0	40,87	0,09	0,71	0,13	0,01
Farinha	747,9	196,92	68,47	62,78	803,18	10,92	2,02	0,25	0,16	5,0	5,0	0,46	1,80	1,07	0,01	746,41	3,78	1,27	4,35	40,21	89,39	8,37	64,56	5,84	159,83	545,9	1922,1	31,95	0,01	0,75	0,11	0,01
Frutuoso II	19,1	19,09	7,75	7,38	23,45	2,47	1,22	0,67	0,32	3,0	3,0	1,99	0,32	1,22	1,15	483,69	0,95	1,47	0,33	38,73	6,56	2,92	28,90	48,65	212,87	921,5	1875,5	41,10	0,02	1,00	0,05	0,04
Gamela	12,8	15,91	6,01	5,62	5,62	2,13	1,25	0,64	0,35	1,0	1,0	0,50	0,40	0,44	0,08	193,36	0,76	1,33	85,11	35,15	5,62	2,27	26,78	74,88	267,77	771,1	1931,5	73,60	0,42	0,55	0,03	0,00
Jatobá I	94,0	53,77	25,53	25,17	134,53	3,68	1,55	0,41	0,14	5,0	2,0	0,61	0,86	1,43	0,03	944,40	1,32	1,41	28,07	40,85	22,52	4,17	33,15	14,77	143,22	730,0	1922,1	34,40	0,00	0,75	0,25	0,03
Queimadas	124,3	56,57	25,02	24,68	101,53	4,97	1,42	0,50	0,20	4,0	2,0	0,49	0,34	0,82	0,27	269,57	2,43	1,93	0,14	40,26	22,58	5,50	31,87	8,18	91,15	672,5	1834,4	37,00	0,00	1,00	0,00	0,02
Riacho Cavalos	161,4	87,12	24,71	23,78	207,97	6,53	1,92	0,27	0,26	4,0	6,0	0,26	0,72	1,29	0,04	644,27	1,82	1,07	0,08	38,49	39,13	4,13	33,36	7,61	96,66	670,4	1897,8	29,20	0,00	0,35	0,65	0,04
Serra Vermelha	55,7	37,90	15,58	15,39	84,32	3,58	1,42	0,49	0,23	3,0	10,0	0,21	0,07	1,51	0,93	438,69	1,53	1,69	0,08	40,57	15,13	3,68	29,14	14,89	111,18	678,6	1931,5	37,00	0,00	1,00	0,00	0,03
Tamanduá I	23,4	20,00	8,31	8,20	14,91	2,82	1,16	0,75	0,34	3,0	2,0	2,19	0,42	0,64	6,40	70,00	1,11	1,56	18,98	37,61	6,07	3,86	17,58	13,43	65,00	449,4	1805,5	37,00	0,00	1,00	0,00	0,02
Vazantes	137,0	61,86	23,08	22,50	194,43	5,94	1,48	0,46	0,26	4,0	4,0	0,96	0,20	1,42	1,03	595,93	2,96	1,94	0,04	40,61	25,29	5,42	41,97	13,61	159,32	737,1	1931,5	36,28	0,00	0,94	0,06	0,02
Camalau	1054,0	186,46	80,71	71,91	620,78	13,06	1,61	0,39	0,16	5,0	2,0	0,49	0,31	0,59	0,14	300,37	3,99	1,09	5,30	40,86	79,27	13,30	51,36	2,13	69,11	481,1	1715,5	36,46	0,01	0,88	0,12	0,01
Campos	181,2	63,76	32,52	32,18	86,57	5,57	1,33	0,57	0,17	3,0	2,0	0,84	0,32	0,48	0,09	224,61	2,31	1,64	15,31	37,89	24,16	7,50	51,15	6,57	88,42	357,8	1715,5	56,64	0,23	0,72	0,05	0,01
Cordeiro	1665,8	228,50	87,56	78,76	990,84	19,02	1,57	0,41	0,22	5,0	2,0	0,42	0,22	0,59	0,22	428,27	6,01	1,14	7,22	40,93	96,11	17,33	228,18	4,68	190,83	445,3	1715,5	35,33	0,01	0,89	0,03	0,00
Santo Antonio	340,6	93,83	25,54	25,39	365,37	13,34	1,42	0,49	0,52	5,0	2,0	47,16	0,50	1,07	0,64	300,35	4,34	1,29	5,10	38,29	37,50	9,08	935,42	2,90	53,53	453,7	1667,9	41,37	0,10	0,76	0,00	0,00
São Domingos	65,5	45,90	23,18	22,43	28,02	2,83	1,59	0,40	0,12	2,0	3,0	0,38	0,16	0,43	0,06	102,66	1,10	1,50	10,06	39,60	19,41	3,37	259,38	7,12	57,61	375,4	1616,7	34,36	0,00	0,78	0,22	0,01
Livramento	37,0	24,00	4,04	3,69	5,73	1,51	1,06	0,90	0,37	2,0	2,0	129,00	1,00	0,94	0,49	18,77	0,58	1,40	3,79	40,86	9,30	3,98	7,26	9,60	23,72	507,8	1805,5	37,00	0,00	1,00	0,00	0,00
Namorados	14,2	16,05	7,62	6,94	12,42	1,86	1,19	0,70	0,24	2,0	3,0	0,75	0,20	0,88	0,28	87,52	0,79	1,55	17,86	40,86	5,28	2,69	15,44	16,24	61,19	466,6	1913,9	37,00	0,00	1,00	0,00	0,01
Soledade	313,1	81,21	35,86	35,63	208,31	8,73	1,29	0,61	0,24	4,0	2,0	0,11	0,17	0,67	0,18	46,57	3,21	1,46	2,87	40,86	29,81	10,50	34,57	4,15	73,39	343,8	1610,1	61,45	0,43	0,15	0,00	0,01
Antenor Navarro	1720,0	288,49	82,86	81,74	1665,82	17,89	2,10	0,23	0,22	6,0	5,0	0,47	1,16	1,12	0,00	715,74	7,32	1,61	0,05	48,90	131,99	11,23	68,06	2,80	107,66	561,1	1909,6	51,37	0,20	0,51	0,30	0,01
Aparecida	3720,0	374,14	147,98	146,86	4019,95	23,38	1,78	0,32	0,16	7,0	5,0	0,70	0,65	1,16	0,00	784,18	11,92	2,02	0,04	48,85	164,74	21,00	90,90	1,83	107,44	794,9	1912,4	40,60	0,09	0,55	0,36	0,01
Piancó	4710,0	381,59	145,86	145,39	7056,87	31,62	1,57	0,40	0,22	7,0	5,0	1,42	0,15	1,53	0,00	1409,30	11,80	1,49	0,03	48,93	160,75	28,69	135,24	2,92	198,58	863,0	1830,8	36,40	0,01	0,85	0,15	0,01
Emas	530,0	135,34	46,95	45,29	511,09	12,80	1,55	0,42	0,27	5,0	3,0	5,78	0,69	0,85	0,01	636,00	4,46	1,35	0,07	48,99	56,56	10,63	67,21	6,89	169,00	1,324,1	1932,2	35,08	0,00	0,89	0,08	0,03
Patos	1850,0	270,79	80,28	78,08	1843,53	20,84	1,85	0,29	0,26	6,0	4,0	9,34	0,17	1,10	0,00	1072,18	6,40	1,20	0,03	283,22	120,56	13,88	221,54	4,25	173,72	969,5	1778,1	32,65	0,03	0,65	0,17	0,01
Serra Negra Norte	3330,0	382,65	130,73	129,33	3589,42	22,80	1,96	0,26	0,17	6,0	4,0	0,42	0,16	1,20	0,00	1248,92	7,29	1,27	0,03	48,98	172,71	17,26	104,25	3,50	191,15	843,3	1799,1	34,51	0,02	0,78	0,11	0,01
Faz, Alagamar	2270,0	244,38	125,85	125,49	1362,88	16,90	1,48	0,45	0,13	5,0	1,5	0,57	0,41	0,64	0,00	401,19	6,72	1,59	0,05	48,96	100,08	21,25	95,66	2,91	134,06	545,2	1580,6	46,38	0,13	0,79	0,06	0,00
Pedro Velho	3590,0	345,18	193,78	193,78	2466,52	17,81	1,65	0,37	0,09	5,0	7,0	0,15	0,11	0,71	0,00	512,33	7,32	1,64	0,11	48,96	148,05	23,31	117,69	3,48	204,55	769,7	1645,3	43,55	0,15	0,70	0,11	0,00
Caraúbas	5120,0	383,89	110,14	109,21	2842,45	46,24	1,51	0,44	0,42	6,0	8,0	0,89	0,29	0,56	0,00	419,21	14,83	1,27	0,02	49,00	158,44	32,14	117,83	1,76	125,41	601,0	1682,6	36,14	0,02	0,85	0,09	0,01
Poço de Pedras	3140,0	347,72	130,31	129,08	2371,53	24,66	1,72	0,34	0,19	6,0	6,0	0,83	0,11	0,74	0,00	439,79	9,26	1,49	0,02	48,98	151,40	21,23	84,56	1,49	84,35	602,2	1726,2	50,08	0,30	0,44	0,13	0,00
Guarita	17220,0	892,31	313,14	312,21	10829,32	56,21	1,88	0,28	0,18	7,0	1,0	0,83	1,32	0,62	0,00	705,70	20,22	1,43	0,06	49,00	398,86	44,13	169,84	1,10	146,15	578,4	1652,1	42,06	0,11	0,69	0,14	0,01
Umburana/Sumé	10,4	11,41	4,90	4,58	6,54	1,62	1,13	0,78	0,33	2,0	2,0	1,48	0,24	0,82	0,38	64,99	0,49	1,14	0,27	41,29	3,09	2,57	14,26	24,17	68,14	551,4	1900,8	36,16	0,00	0,93	0,07	0,00
Jatobá/Sumé	26,8	22,09	9,84	9,67	14,81	2,56	1,23	0,66	0,26	2,0	3,0	0,99	0,24	0,59	0,16	125,05	0,74	1,13	0,14	48,28	7,68	3,28	22,72	16,99	85,36	681,8	1900,8	31,18	0,06	0,19	0,75	0,00
Gangorra/Sumé	137,4	57,89	29,20	29,03	82,58	4,60	1,40	0,51	0,16	4,0	2,0	0,50	0,38	0,61	0,02	172,64	1,47	1,27	0,06	48,95	22,86	5,88	34,12	6,90	80,02	595,9	1900,8	35,25	0,01	0,80	0,19	0,00