UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ
***CAMPUS* DE CASCAVEL**
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

QUANTIS MENSAIS DE PRECIPITAÇÃO NO ESTADO DO PARANÁ UTILIZANDO TÉCNICAS MULTIVARIADAS DE AGRUPAMENTO

WAGNER ALESSANDRO PANSERA

CASCAVEL - Paraná - Brasil
Fevereiro – 2010.

WAGNER ALESSANDRO PANSERA

QUANTIS MENSAIS DE PRECIPITAÇÃO NO ESTADO DO PARANÁ UTILIZANDO TÉCNICAS MULTIVARIADAS DE AGRUPAMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Agrícola, em cumprimento parcial aos requisitos para obtenção do título de Mestre em Engenharia Agrícola, área de concentração em Recursos Hídricos e Saneamento Ambiental.

Orientador: Profº Dr. Benedito Martins Gomes.

CASCAVEL - Paraná - Brasil

Fevereiro – 2010.

Pansera, Wagner Alessandro
P196 Quantis mensais de precipitação no Estado do Paraná utilizando técnicas multivariadas de agrupamento. / Wagner Alessandro Pansera. – Cascavel, 2010.
74 f.

Orientador: Profº Dr. Benedito Martins Gomes.
Dissertação (Mestrado) – Universidade Estadual do Oeste do Paraná – Campus de Cascavel.

1. Recursos hídricos – Gestão. 2. Sistema hidrológico – Modelo probabilístico 3. Análise de freqüência regional - Metodologia. 4. Medida de heterogeneidade. 5. Momentos-L - Distribuição de probabilidade I. Gomes, Benedito Martins. II. Título.

CDD – 628.16098162
333.91098162

Ficha Catalográfica elaborada pelo Sistema de Bibliotecas da Unioeste (Sandra Regina Mendonça CRB – 9/1090)

WAGNER ALESSANDRO PANSERA

QUANTIS MENSAIS DE PRECIPITAÇÃO NO ESTADO DO PARANÁ UTILIZANDO TÉCNICAS MULTIVARIADAS DE AGRUPAMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Mestre em Engenharia Agrícola, área de concentração Recursos Hídricos e Saneamento Ambiental, **aprovada** pela seguinte banca examinadora:

Orientador: Professor Dr. Benedito Martins Gomes
Centro de Ciências Exatas e Tecnológicas, UNIOESTE - Cascavel

Professor Dr. Marcio Antonio Vilas Boas
Centro de Ciências Exatas e Tecnológicas, UNIOESTE - Cascavel

Professor Dr. Ricardo Nagamine Costanzi
Universidade Tecnológica Federal do Paraná, UTFPR - Londrina

Cascavel, fevereiro de 2010.

BIOGRAFIA

Data de nascimento: 17/04/1985

Naturalidade: Cascavel – PR

Graduação: Bacharelado em Engenharia Agrícola – Unioeste – 2007

Pós-graduação *strictu sensu*: Mestrado Engenharia Agrícola: área de concentração em Recursos Hídricos e Saneamento Ambiental – UNIOESTE.

À minha mãe Ivete,

DEDICO.

AGRADECIMENTOS

À Universidade Estadual do Oeste do Paraná (UNIOESTE), *campus* de Cascavel, em especial ao Programa de Pós-graduação em Engenharia Agrícola, pelo apoio e pela oportunidade de realização do curso.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela disponibilização da bolsa de estudos.

Ao Professor Benedito Martins Gomes, pela orientação, compreensão, incentivo, amizade e confiança.

Aos Professores, Marcio Antonio Villas Boas e Silvio César Sampaio, pela dedicação e transposição dos conhecimentos durante a realização desta dissertação.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

QUANTIS MENSAIS DE PRECIPITAÇÃO NO ESTADO DO PARANÁ UTILIZANDO TÉCNICAS MULTIVARIADAS DE AGRUPAMENTO

O conhecimento do comportamento espacial e temporal da precipitação pluvial e sua frequência são vitais para o planejamento, operação e dimensionamento de obras hidráulicas. Para este fim, são construídos modelos probabilísticos baseados nas observações efetuadas nos locais de interesse com a finalidade de obter os riscos de ocorrência futura. No entanto, uma rede de monitoramento hidrológico não apresenta dados de todos os locais de interesse, sendo necessário recorrer a procedimentos que possibilitem o transporte desta informação. A regionalização hidrológica é um procedimento estatístico que permite, dentro de uma região homogênea, estimar probabilidades de ocorrência, pelos dados das estações vizinhas, no local de interesse. O problema está em como determinar uma região homogênea. A solução proposta neste trabalho foi a utilização de metodologias de agrupamento multivariados, k-médias e hierárquicos, validados por índices de qualidade de agrupamento. Foram usadas 227 estações localizadas no estado do Paraná com dados mensais referentes ao período de 1976-2006. Os agrupamentos obtidos foram submetidos às medidas de discordância e heterogeneidade para avaliar a homogeneidade dos grupos. A metodologia de agrupamento que obteve o melhor desempenho foi a metodologia híbrida entre k-médias e *ward*, obtendo erros de, no máximo, 10% na estimativa de quantis regionais adimensionais.

Palavras-chave: medida de heterogeneidade, momentos-L, distribuição de probabilidade.

ABSTRACT

RAINFALL MONTHLY QUANTILE IN PARANÁ-BR USING MULTIVARIATE CLUSTERING TECHNIQUES

The rainfall is the only way of entry of water in river basins, therefore the knowledge of their spatial and temporal behavior as well as their frequency is of vital importance for the planning, operation and design of hydraulic works. To this end, probabilistic models are built based on observations made in the place of interest in order to obtain future risks. It often happens that a network of hydrological monitoring data does not show the place of interest is necessary to use procedures that allow the transport of this information. Hydrological regionalization is a statistical procedure that allows, within a homogeneous region, estimate probabilities of occurrence, by data from neighboring stations, in site of interest. The problem is how to define homogeneous region. The solution proposed in this paper was the use of multivariate clustering methods, k-means and hierarquical cluster, validated by quality index. The clusters were then subjected to measures of discordancy and heterogeneity to evaluate homogeneity of groups. The clustering methodology that obtained the best performance was a hybrid approach between k-means and *ward*, getting errors of at most 10% in the estimation of quantiles regional dimensionless.

Keywords: heterogeneity measure, L-moments, probability distribution

1 INTRODUÇÃO

A precipitação pluviométrica é o principal elo entre a fase atmosférica e a fase terrestre do ciclo hidrológico, constituindo-se na entrada do sistema hidrológico e, por consequência, na principal forma de entrada de água em uma bacia hidrográfica. A disponibilidade de precipitação em uma bacia é fator determinante para se quantificar, dentre outras coisas, a necessidade de irrigação e o abastecimento doméstico e industrial. A determinação da intensidade de precipitação é importantíssima em estudos que visem o controle de enchentes e a minimização da ocorrência de erosão hídrica. As características principais de interesse são o total precipitado, a duração da precipitação e suas distribuições temporal e espacial.

Quando é feito um estudo para planejamento de longo prazo do uso de uma ou mais bacias hidrográficas, a precipitação é um dado básico, pois não sofre influência direta da ação antrópica provocada no meio. Deve-se monitorar a ocorrência das precipitações para melhorar o conhecimento hidrológico da região de forma a gerar dados necessários para projetos e obras hidráulicas. Estes fornecem subsídios ao planejamento e gerenciamento racional de recursos hídricos, tais como outorga, cobrança, enquadramento dos corpos hídricos e estudos de previsão. Nos projetos de drenagem e construção de reservatórios de regularização (barragens), entre outros, os dados de precipitação serão muitas vezes necessários para o dimensionamento das obras e conduzirão a resultados tanto mais seguros quanto melhor for sua definição.

Entretanto, como outros fenômenos naturais, a precipitação está sujeita há uma componente aleatória, tanto espacial quanto temporal, não sendo possível saber exatamente quando, onde e como ela irá ocorrer. Por essa razão, foram criados modelos probabilísticos para estimar ocorrências futuras baseados nas observações de eventos passados.

O processo de modelagem supõe a existência de amostras de eventos que possibilitem o ajuste dos parâmetros de modo a representar as condições locais. Ocorre que, muitas vezes, a necessidade de informações recai em locais com ausência de dados. Uma rede hidrológica raramente cobre todos os locais de interesse em uma bacia hidrográfica, o que gera lacunas espaciais que precisam ser preenchidas. Nessas situações, um procedimento usual é a aplicação de estudos de regionalização, também chamados de análise de frequência regional.

Quando existem séries observadas em várias estações de monitoramento pluviométrico distribuídas no espaço e há similaridade nas frequências observadas, o

conjunto pode ser reunido em uma região dita homogênea, e ser tratado conjuntamente, pela análise regional de frequências. Dessa forma, resolve-se o problema com tamanho e consistências locais dos dados, ou seja, de uma única estação. Assim, as estimativas de probabilidades obtidas pela análise regional poderão ser mais robustas que as produzidas pela análise de frequência local.

A metodologia para análise regional proposta por Hosking e Wallis (1997), que utiliza os princípios do *index-flood*, tem sido bastante difundida. Seu principal atrativo está na utilização de momentos-L para obtenção da distribuição regional de frequências e para a construção de estatísticas auxiliares, com destaque para a medida de heterogeneidade H, que inseriu maior objetividade no julgamento da homogeneidade das regiões.

A referida medida de heterogeneidade, porém, depende do conhecimento dos momentos-L, o que limita a sua aplicação. Seria desejável um recurso mais generalizável para a análise, pois a determinação dos parâmetros da distribuição regional pode ser executada por diferentes métodos. Uma solução para este problema são as técnicas multivariadas de agrupamento que possibilitam classificar as estações de monitoramento pluviométrico baseadas em uma distância de similaridade.

Em geral, os métodos de agrupamento, apresentam como deficiência a necessidade de definição preliminar do número de classes. Usualmente, o procedimento adotado é iterativo, efetuando-se os agrupamentos para vários números de classes, e avaliando-se a melhor alternativa por meio de um índice de validade. No entanto, existem muitas alternativas de índices de validade disponíveis, sendo desejável que pelo menos certo número destes índices aponte para uma mesma solução, atribuindo-lhe maior confiabilidade. Alguma forma de conciliação ou investigação conjunta destes índices, associada a outros recursos, que considerem os objetivos específicos do agrupamento para a definição de regiões homogêneas, poderia auxiliar numa solução mais conclusiva.

Definidas as regiões homogêneas pelas técnicas de agrupamento e seus índices de validade é interessante que estas sejam submetidas ao teste de heterogeneidade H. Dessa forma, estar-se-á averiguando a confiabilidade das metodologias empregadas na obtenção do resultado final dos grupos.

2 OBJETIVOS

2.1 Objetivo Geral

Estimar quantis regionais de precipitação pluviométrica no estado do Paraná utilizando técnicas multivariadas de agrupamento.

2.2 Objetivos Específicos

Testar as técnicas hierárquicas e k-médias e a forma conjunta entre as duas na formação de grupos de estações homogêneas utilizando como variável classificatória as séries históricas e verificar a utilização dos índices de qualidade de agrupamento.

Verificar o efeito das estações discordantes e tamanhos de grupo na medida de heterogeneidade.

Testar as distribuições gama e Pearson tipo III na estimativa de quantis regionais.

3 REVISÃO BIBLIOGRÁFICA

3.1 Gestão de Recursos Hídricos

Com o acelerado crescimento populacional no mundo houve um aumento da demanda de água, o que ocasionou problemas de escassez desse recurso, além de sua degradação em várias regiões do planeta. A água é um recurso natural renovável dos processos físicos do ciclo hidrológico. Para acompanhamento, análise e gerenciamento dos recursos hídricos é fundamental a medição constante dos principais elementos que controlam o ciclo hidrológico, para a determinação da água disponível (BARBOSA; VALERIANO; SCOFIELD, 2005).

Para proteger os mananciais e preservar o seu papel ecológico e social, informações confiáveis sobre a sua qualidade e quantidade, em cada bacia hidrográfica, são extremamente importantes para o gerenciamento e planejamento adequado à sua utilização, devendo fazer-se acompanhar de uma relação de princípios básicos que configurem uma política de gestão das águas capaz de atender aos interesses do País, do Estado, do Município como, também, da população residente na bacia hidrográfica (FIGUEIREDO; RUBERT, 2001).

A incerteza resultante da escassez de água no mundo vem acarretando a necessidade de se introduzirem práticas mais flexíveis de gestão desse recurso, como descentralização, integração, participação e financiamento compartilhado, pois somente dessa forma a preocupação com a sustentabilidade será incorporada tanto pelas políticas públicas como pelas ações de empresários e cidadãos (LUCHINI; SOUZA; PINTO, 2003).

Descentralização na política de recursos hídricos significa a institucionalização, em nível local, de condições institucionais, técnicas, financeiras e organizacionais para a implementação das tarefas de gestão, conforme atribuições designadas na lei de recursos hídricos, garantindo continuidade no fluxo da oferta dos bens e serviços. O conceito de local refere-se aqui à bacia hidrográfica como unidade de planejamento e gestão – princípio estabelecido na Lei Federal 9.433/97 e leis dos estados da federação – em referência ao fenômeno geomorfológico e geográfico de área de drenagem que forma uma bacia, e condiciona a sua gestão e planejamento, seja no que concerne à quantidade ou à qualidade de suas águas (PEREIRA; JOHNSSON, 2005).

Também é definido na lei um conjunto de instrumentos considerados essenciais à boa gestão do uso da água: os planos de recursos hídricos, que são planos diretores que visam a fundamentar e orientar a implementação da Política Nacional de Recursos Hídricos

e o gerenciamento dos recursos hídricos; a outorga de direito de uso dos recursos hídricos, instrumento através do qual o usuário assegura, por prazo determinado, o seu direito ao uso desse recurso; a cobrança pelo uso dos recursos hídricos, instrumento capaz de promover as condições de equilíbrio entre as forças de oferta (disponibilidade de água) e as de demanda, promovendo, em consequência, a harmonia entre os usuários; o enquadramento dos corpos d'água em classes de uso, que constitui, de certa forma, uma classificação que permite a destinação de volumes de água de determinado padrão de qualidade a usos cuja exigência seja compatível com esse padrão; e o Sistema Nacional de Informações sobre Recursos Hídricos, conjunto de elementos organizados sob a forma de banco de dados, que auxilia no gerenciamento e planejamento do uso dos recursos hídricos (LUCHINI; SOUZA; PINTO, 2003).

O sistema de informação sobre recursos hídricos tem como objetivo principal o de produzir, sistematizar e disponibilizar dados e informações que caracterizem as condições hídricas da bacia em termos de quantidade e qualidade da água nos diversos usos, e em termos das condições do ecossistema, traduzido pelas pressões antrópicas nela existentes. Estas últimas assumem diversas formas possíveis de caracterização, por mapas de uso e ocupação do solo, declividade, cobertura vegetal e cargas pontuais, referentes a captações e lançamentos em diferentes pontos da rede hidrográfica, expressas no cadastro de usuários da água na bacia (PEREIRA; JOHNSSON, 2005).

O surgimento dos sistemas de informação sobre recursos hídricos ocorreu por força da necessidade de dinamizar o processo de gestão participativa. Dois termos de uma equação a ser resolvida surgiram: a crescente complexidade dos múltiplos usos da água, forçando a melhoria da dinâmica gerencial e a oferta de serviços gerada pelo desenvolvimento tecnológico como o Sistema de Informações Geográficas (SIG), o sensoriamento remoto, a telemetria, o desenvolvimento de modelos matemáticos, entre outros. A questão continua sendo esta: como usar os recursos tecnológicos para, de fato, contribuir no esclarecimento da complexidade dos usos da água e da dinâmica ambiental de uma bacia hidrográfica (FISTAROL; FRANK; REFOSCO, 2004).

As condições climáticas são determinadas pelo acompanhamento da precipitação, temperatura, umidade e vento, e as condições hídricas do solo são obtidas pelo balanço hídrico que permite determinar os déficits e o grau de *stress* nas comunidades vegetais, juntamente com as medidas da altura das águas nos rios, possibilitando uma visão quantitativa do potencial hidrológico existente a cada mês, disponíveis para as diversas atividades humanas (SILVA; D'ANGIOLELLA, 2001).

Dentre as atividades econômicas, a que tem maior dependência das condições do tempo e do clima é a agricultura. As condições atmosféricas afetam todas as etapas das atividades agrícolas, que vão desde o preparo do solo para a semeadura até a colheita e,

em muitos casos, transpondo as barreiras da unidade produtora, afetando o transporte, o preparo e o armazenamento dos produtos (CORAL *et al*., 2005). Dessa forma, o monitoramento e análise da influência dos elementos climáticos nas diversas fases do desenvolvimento de uma cultura podem contribuir para a redução de possíveis prejuízos provenientes de condições meteorológicas adversas.

Os parâmetros climáticos influenciam no tipo de cultura a ser cultivada e como ela será manejada, como a época propicia para a semeadura, poda, colheita, rotatividade de cultura, etc. Quanto à disponibilidade de água, as variáveis evapotranspiração e chuvas, através do balanço hídrico pode se saber quais são os períodos críticos e a conveniência ou não do uso de sistemas de irrigação, dimensionado e operado segundo dados agroclimatológicos da região (COSTA *et al*., 2006).

Numa visão histórica, sabe-se que países em desenvolvimento, como o Brasil, dependem fortemente da agricultura, sofrem com as condições extremas de precipitação que provocam grandes perdas de produção agrícola prejudicando sensivelmente toda a área econômica. Esses eventos meteorológicos extremos associados aos efeitos econômicos causam graves consequências na sociedade, por serem provocados por anomalias extensas e altamente prejudiciais. Uma forma de minimizar esses problemas é dispor de diagnóstico para tais fatos. Por isso, a necessidade de se conhecer a época de plantio, analisando todo o ciclo da cultura e procurando prever as condições ambientais em todas as suas fases fenológicas (MARTINS; PINHO; GONÇALVES, 2008).

3.2 Regionalização Hidrológica

A escassez de dados hidrológicos inviabiliza o uso de dados locais em análise de frequência para estimar quantis. Para resolver este problema, hidrologistas calculam quantis de um local de interesse baseados em bacias hidrográficas vizinhas, que são semelhantes ao local de interesse. Um grupo de bacias hidrográficas com resposta hidrológica semelhante constitui uma região homogênea e o procedimento de identificação de regiões homogêneas é tradicionalmente chamado de regionalização hidrológica (RAO; SRINIVAS, 2006b).

A análise local de frequência de variáveis hidrológicas dispõe de um conjunto de técnicas de inferência estatística e de modelos probabilísticos, que têm sido objeto de frequente investigação, visando, principalmente, à obtenção de estimativas cada vez mais eficientes e confiáveis. Entretanto, a inexistência de amostras suficientemente longas impõe um limite superior ao grau de sofisticação estatística a ser empregado na análise local de frequência. Em linhas gerais, a análise regional de frequência utiliza um grande conjunto de

dados espacialmente disseminados de certa variável, como por exemplo, vazões e precipitações observadas em pontos distintos de uma região considerada homogênea, do ponto de vista estatístico ou dos processos físicos em foco, para estimar os quantis associados a diferentes probabilidades de excedência, para certo local dentro dessa região. A análise de frequência regional pode ser usada para aumentar a confiabilidade dos quantis estimados para um ponto já monitorado, bem como para estimar os quantis em locais que não possuem uma coleta sistemática de informações. Em geral, essa última aplicação da análise de frequência regional é a mais comum (NAGHETTINI; PINTO, 2007).

Um importante requisito para análise de frequência regional é a identificação de regiões que são usadas para transferência de informação hidrológica. Neste contexto, uma região, significa um conjunto de bacias hidrográficas, não havendo necessidade de contiguidade geográfica, que podem ser consideradas similares em termos de resposta hidrológica. O objetivo do processo de regionalização pode ser visto como a identificação de agrupamentos de regiões que são suficientemente similares para justificar a combinação e transferência de informação hidrológica de locais dentro da região (BURN, 1997).

Melo Junior *et al*. (2006) determinaram regiões homogêneas quanto à distribuição de frequência de chuvas com auxílio da análise dos componentes principais e técnicas de agrupamentos no leste do Estado de Minas Gerais. A metodologia hierárquica de *ward* proporcionou a melhor distribuição espacial dos grupos, indicando três e cinco grupos para os critérios de classificação, respectivamente, meso e macroescala hidroclimática.

Keller Filho, Assad e Lima (2005) delimitaram regiões homogêneas quanto à distribuição de probabilidades de chuva. A delimitação das regiões foi feita mediante aplicação da análise de agrupamento hierárquica de *ward*, com variáveis classificatórias definidas pela proporção de pêntadas secas e por medidas de posição, escala e forma das distribuição de frequências da quantidade de chuva. A análise de agrupamento permitiu identificar 25 zonas pluviometricamente homogêneas em todo o território brasileiro.

Porém, a simples delimitação de regiões homogêneas por metodologias de classificação não garante necessariamente que a região seja homogênea. Para verificar a veracidade da suposição de homogeneidade, Hosking e Wallis (1997) desenvolveram dois procedimentos: a medida de discordância e o teste de heterogeneidade.

O objetivo principal da medida de discordância é identificar locais em que os quocientes-L amostrais são discrepantes dos demais (CANNARAZZO *et al*., 2009). O teste de heterogeneidade avalia a homogeneidade centrando-se sobre três medidas para diferentes ordens dos quocientes-L amostrais. O conceito básico do teste é determinar a variabilidade amostral dos quocientes-L e compará-la com a variação que seria esperada em um grupo homogêneo. O valor médio esperado e o desvio padrão de tais medidas de dispersão para um grupo homogêneo são avaliados por repetidas simulações, pela geração

de grupos com o mesmo tamanho de registro dos dados observados (CASTELLARIN; BURN; BRATH, 2008).

Rao e Srinivas (2006a) usaram três algoritmos híbridos, nos quais aplicavam um processo de agrupamento particional para refinar os resultados obtidos das técnicas de agrupamento hierárquicas. Foram usados os algoritmos hierárquicos de ligação simples, ligação completa e *ward*, enquanto o algoritmo particional utilizado foi o k-médias. O método híbrido entre k-médias e *ward* proporcionou os melhores resultados. Rao e Srinivas (2006b) fizeram uso do algoritmo de agrupamento particional *fuzzy* c-médias na obtenção de regiões estatisticamente homogêneas para análise de frequência regional de vazão. Os autores em questão usaram o teste de homogeneidade e a medida de discordância para validar os resultados obtidos pelos algoritmos classificatórios em bacias hidrográficas de Indiana, EUA.

Rahnama e Rostami (2007) usaram a metodologia dos momentos-L para regionalização de vazão na bacia hidrográfica do rio Halil no Irã. Para identificar as regiões homogêneas utilizaram-se da metodologia de agrupamento hierárquico de *ward*. Os grupos obtidos foram então submetidos ao teste de heterogeneidade e à medida de discordância. Selecionaram a distribuição mais apropriada e estimaram a raiz quadrática do erro médio entre a regionalização e os dados observados, bem como, na estimativa local. Os resultados demonstraram bom ajuste entre os dados estimados e os dados observados.

Cannarazzo *et al*. (2009) iniciando com 105 estações hidrométricas, selecionaram 57 usando a medida de discordância proposta por Hosking e Wallis (1997). Foi aplicado o teste de Mantel para descobrir quais parâmetros físicos e morfológicos se correlacionavam melhor com a vazão usando uma abordagem de matriz de distâncias. Estes parâmetros foram usados na metodologia hierárquica de *ward* e foi aplicado o teste de heterogeneidade para verificar a homogeneidade dos grupos obtidos. Usando a distribuição log-normal de três parâmetros estimaram quantis adimensionais e, por meio de regressão múltipla, obtiveram expressões que relacionam a vazão média anual para algumas características climáticas e morfológicas da bacia.

Yang *et al*. (2010) estudaram a bacia hidrográfica do rio Pérola utilizando os algoritmos hierárquicos de ligação média e *ward*, seguidos do teste de heterogeneidade e medida de discordância, usando características topográficas da bacia em estudo, obtendo seis regiões homogêneas. Estudaram o ajuste de varias distribuições de probabilidades com três parâmetros, por fim, mostraram por simulações de Monte Carlo que, para períodos de retorno menores que 100 anos, as estimativas dos quantis são confiáveis.

Fowler e Kilsby (2003) utilizaram a regionalização hidrológica para identificar mudanças na precipitação máxima anual de 1, 2, 5 e 10 dias em nove regiões hidrológicas do Reino Unido, no período de 1961 a 2000. Os autores observaram pequenas mudanças para 1 e 2 dias, porém para 5 e 10 dias as mudanças foram significativas.

Modarres (2008) usando os momentos-L regionalizou a velocidade do vento para o árido e semiárido do Irã usando a distribuição logística generalizada. O autor conseguiu identificar regiões mais suscetíveis à erosão eólica e observou que não houve linearidade entre período de retorno e energia do vento.

Bocchiola, Medagliani e Rosso (2006) fizeram uso da análise de frequência regional para estudar a camada de neve de três dias. Usando 40 estações dos Alpes italianos centrais utilizaram a medida de discordância e o teste de heterogeneidade para aferir a homogeneidade da região e, então, estimaram os quantis regionais para período de retorno de 300 anos, obtendo o perigo de avalanche.

Gellens (2002) estudou precipitações extremas com durações de 1, 2, 4, 5, 7, 10, 15, 20, 25 e 30 dias de 165 estações climatológicas na Bélgica, para estimar a curva de crescimento regional usando distribuição de valores extremos, com o objetivo de determinar quantis locais. Testou as durações segundo a medida de heterogeneidade para verificar sazonalidades verão-inverno. Por fim, obteve os quantis adimensionais para diferentes períodos de retorno.

Kjeldsen, Smithers e Schulze (2001) estudaram precipitações com menos de 24 h de duração usando estações com menos de 10 anos de registros. Usaram a metodologia hierárquica de *ward* com variáveis classificatórias locais, tais como, coordenadas geográficas e sazonalidade. Testaram a homogeneidade do grupo e estimaram a curva regional de precipitação com menos de 24 h.

Goel *et al*. (2004) regionalizaram ventos extremos em Ontário, Canadá. Por meio de um procedimento heurístico, que visava minimizar a medida de heterogeneidade e maximizar o número de estações, obtiveram os agrupamentos, identificaram a distribuição regional e estimaram os quantis com erros menores de 15%.

3.3 Análise Multivariada

No gerenciamento ambiental é necessária uma visão holística. Uma das maneiras viáveis para solucionar este problema são os métodos multidimensionais. Quando se deseja informações a cerca de um grupo de variáveis ou, às vezes, do conjunto total dos dados de uma região, é usual recorrer-se à análise multivariada. Esta técnica estatística é também usada para reduzir ao máximo o número de variáveis envolvidas em um problema com uma pequena perda de informações.

Segundo Lopes (2006), os principais objetivos desta técnica são:

- Reduzir a dimensão de interpretação de uma matriz de dados;

- Investigar o comportamento espacial e temporal das variáveis consideradas;
- Obter grupos homogêneos das variáveis.

Os métodos partem de uma matriz de dados $X_{(nxp)}$ cujas linhas correspondem a “n” unidades (indivíduos) e fornecem “p” características (valores numéricos) cada. No caso de estudos climatológicos, essas unidades podem ser estações meteorológicas; as características seriam dados meteorológicos como pressão, temperatura, vento, umidade relativa, etc., ordenados ou não segundo sequências cronológicas. No caso de uma única variável, estes dados podem constituir uma sequência cronológica, em que cada linha “i” representa o valor da variável para um mês “i” no conjunto “p” locais. A coluna “j” forma uma série temporal da variável em estudo para a j-ésima (coluna) (LOPES, 2006, MOITA; MOITA, 1998).

A matriz de dados é organizada da seguinte forma:

$$X_{(nxp)} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \Leftrightarrow X_{(nxp)} = \begin{pmatrix} X_1 & X_2 ..... & X_p \end{pmatrix}$$

Observe-se que a matriz $X_{(nxp)}$ pode ser interpretada como um ordenamento de “p” vetores colunas (cada vetor equivalente à série temporal de cada variável utilizada no trabalho) ou de “n” vetores linha (cada vetor indicando valores das variáveis meteorológicas de uma determinada rede de estações). O primeiro caso descreve, principalmente, o comportamento temporal de uma rede, enquanto que o segundo ilustra a situação espacial das variáveis em cada época (LOPES, 2006).

A decisão sobre quais variáveis são importantes é feita, geralmente, com base na intuição ou na experiência, ou seja, em critérios que são mais subjetivos que objetivos. A redução de variáveis por meio de critérios objetivos, permitindo a construção de gráficos bidimensionais, contendo maior informação estatística, pode ser conseguida pela análise dos componentes principais. Também é possível construir agrupamentos entre as amostras de acordo com suas similaridades, utilizando todas as variáveis disponíveis, e representá-los de maneira bidimensional por um dendrograma. A análise de componentes principais e de agrupamento hierárquico são técnicas de estatística multivariada complementares (MOITA; MOITA, 1998).

3.3.1 Análise de agrupamentos

Regiões geograficamente contíguas com base em fronteiras geográficas, políticas, administrativas ou fisiográficas foram usadas por um longo tempo em hidrologia para análise de frequência regional. No entanto, esta prática não é aceitável, porque as regiões delimitadas não garantem a homogeneidade hidrológica. Consequentemente, foram desenvolvidos para a regionalização diversos métodos que consideram a semelhança entre locais, em um espaço multidimensional dos atributos relacionados às bacias hidrográficas, tais como características fisiográficas, atributos da localização geográfica e estatísticas locais da variável hidrológica (RAO; SRINIVAS, 2006a).

Recentemente, o aumento da conscientização sobre o uso de dados hidroclimáticos parece ter solicitado a várias agências para trabalhar na criação de bancos de dados em uma multiplicidade de variáveis que influenciam os processos hidrológicos. Para usar efetivamente os arquivos de dados em estudos de regionalização, há a necessidade de desenvolver abordagens para identificar e interpretar os padrões inerentes aos dados hidrológicos. Para esta tarefa, os algoritmos de agrupamentos, que são considerados eficazes em reconhecer a distribuição dos padrões em grandes e pequenos conjuntos de dados, parecem promissores (RAO; SRINIVAS, 2006b).

Análise de agrupamentos é o nome para um grupo de técnicas multivariadas cuja finalidade primária é agregar objetos com base nas características que eles possuem. A análise de agrupamentos classifica objetos de modo que cada objeto é muito semelhante aos outros no agrupamento, em relação a algum critério de seleção pré-determinado. Os agrupamentos resultantes de objetos devem então exibir elevada homogeneidade interna (dentro dos agrupamentos) e elevada heterogeneidade externa (entre agrupamentos). Assim, se a classificação for bem sucedida, os objetos dentro dos agrupamentos estarão próximos quando representados graficamente e diferentes agrupamentos estarão distantes (HAIR JR. *et al*., 2005).

A maior parte dos métodos de agrupamentos requer que a matriz de proximidades entre objetos seja previamente obtida. A proximidade é o termo utilizado para indicar similaridade ou dissimilaridade, que é medida de pelas distâncias. A matriz de proximidade refere-se a uma matriz nxn de coeficientes de proximidades entre os objetos. Assim, na i-ésima linha dessa matriz encontram-se os coeficientes de proximidade entre o i-ésimo objeto e cada um dos demais, incluindo ele mesmo. Existem muitos tipos de medidas de proximidade, sejam elas coeficientes de similaridade ou dissimilaridade que, por sua vez, dependem do tipo de variáveis que está sendo considerado ou utilizado na análise de agrupamentos. Essas medidas são calculadas a partir da matriz de dados (FERREIRA, 2008).

Muitos algoritmos têm sido propostos para análise de agrupamentos. Primeiro, há técnicas hierárquicas que começam com o cálculo das distâncias de cada objeto a todos os outros objetos. Grupos são então formados por um processo de aglomeração ou divisão. A segunda abordagem para análise de agrupamentos envolve partição, com objetos podendo se mover para dentro e para fora de grupos em diferentes estágios da análise. Há muitas variações nos algoritmos usados, mas a abordagem básica envolve primeiro escolher centros de grupos, mais ou menos arbitrários, e alocar os objetos. O processo é repetido até que atenda uma regra de parada (MANLY, 2008).

A técnica de agrupamento hierárquico interliga as amostras por suas associações, produzindo um dendrograma no qual as amostras semelhantes, segundo as variáveis escolhidas, são agrupadas entre si. A suposição básica de sua interpretação é esta: quanto menor a distância entre os pontos, maior a semelhança entre as amostras. Os dendrogramas são especialmente úteis na visualização de semelhanças entre amostras ou objetos representados por pontos em espaço com dimensão maior do que três, em que a representação de gráficos convencionais não é possível (MOITA; MOITA, 1998).

Os métodos não hierárquicos são métodos que têm como objetivo encontrar diretamente uma partição de n elementos em k grupos (clusters), de modo que a partição satisfaça dois requisitos básicos: “coesão” interna (ou “semelhança” interna) e isolamento (ou separação) dos grupos formados. Para se buscar a “melhor” partição de ordem k, algum critério de qualidade da partição deve ser empregado. É possível, computacionalmente, criar todas as partições possíveis de ordem k e, a partir do conhecimento dessas partições, decidir qual seria a mais adequada. Deste modo, são necessários processos que investiguem algumas das partições possíveis com o objetivo de encontrar a partição “quase ótima” (MINGOTI, 2005).

Nos métodos hierárquicos, em geral, o número de grupos k não é conhecido e uma vez que um objeto é aglomerado, ele nunca é realocado. Nos métodos não-hierárquicos o processo é aplicado à matriz de dados e não à matriz de dissimilaridade. Além disso, deve-se conhecer, a priori, o número de grupos k. Os objetos são aglomerados aos k grupos utilizando-se uma função objetivo como critério, cessando de realocá-los quando uma regra de parada pré-especificada for contemplada. Esse método é muitas vezes denominado método de otimização e não partição, função desta característica (FERREIRA, 2008).

Os algoritmos que representam as técnicas de agrupamentos hierárquicos aglomerativos incluem: (1) ligação simples ou vizinho mais próximo, (2) ligação completa ou vizinho mais distante, (3) ligação média, (4) centróide e (5) algoritmo de *ward* (RAO; SRINIVAS, 2006b).

Existem muitos métodos não-hierárquicos baseados em misturas de distribuição, estimação de densidade e partição. Esse último é o mais usado, sendo que o método das k-médias, um tipo de método de partição, é o mais popular (FERREIRA, 2008).

4 MATERIAL E MÉTODOS

4.1 Fluxograma de Pesquisa

A Figura 1 mostra em qual etapa foi aplicado cada método utilizado na pesquisa.

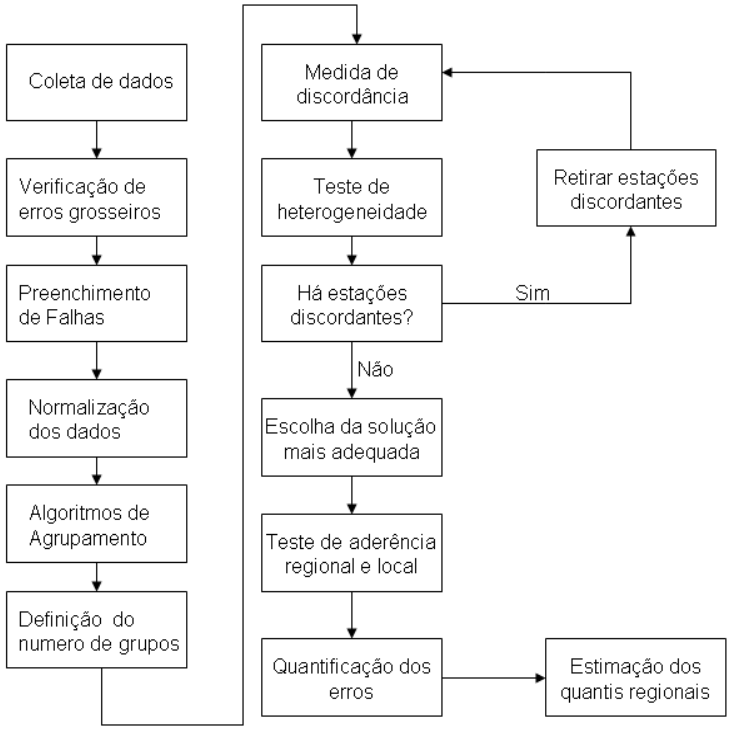

Figura 1 - Fluxograma da metodologia de pesquisa.

4.2 Coleta dos Dados

A metodologia foi empregada em 227 estações pluviométricas pertencentes à Superintendência de Desenvolvimento de Recursos Hídricos e Saneamento Ambiental (SUDERHSA) do estado do Paraná. Os dados foram obtidos no *site* da Agência Nacional das Águas (ANA), via sistema de informações HIDROWEB.

As estações foram selecionadas seguindo os seguintes critérios: (1) dados referentes ao período de 1976 a 2006, (2) menos de 18 falhas, (3) não possuir uma sequência de falhas superior a quatro consecutivas.

Os critérios acima utilizados ajudaram a obter estações com o mesmo tamanho de série e poucas falhas. Este procedimento foi adotado para testar as metodologias de agrupamento na obtenção de séries pluviométricas homogêneas que pudessem ser descritas pela mesma distribuição de probabilidades.

A Figura 2 mostra a distribuição espacial das estações selecionadas para esta pesquisa.

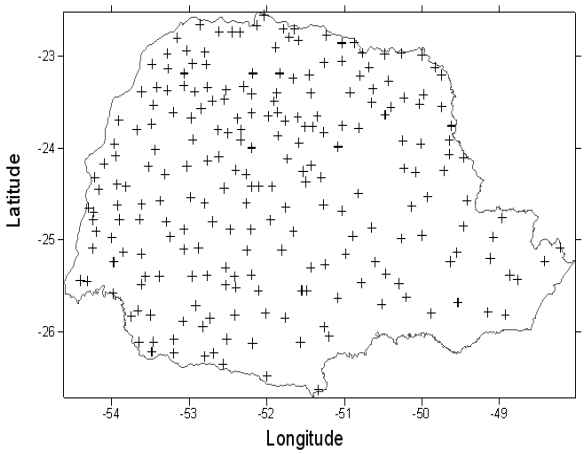

Figura 2 - Distribuição espacial das estações utilizadas neste estudo.

4.3 Preenchimento de Falhas

As estações foram submetidas a uma avaliação preliminar para identificar erros grosseiros de leitura para, posteriormente, ser efetuado o preenchimento das falhas.

Como as estações pluviométricas não apresentam séries continuas de dados, por defeitos no aparelho ou ausência de observador, há necessidade do preenchimento de falhas. Para o preenchimento de falhas em séries pluviométricas, utilizou-se o método da ponderação regional com base em regressões lineares, que consiste em estabelecer regressões lineares entre os postos com dados a serem preenchidos, Y_c, e cada um dos postos vizinhos: X_1, X_2,..., X_n. De cada uma das regressões lineares efetuadas, obteve-se o coeficiente de correlação r, sendo o preenchimento realizado pela Equação 1.

$$Y_C = \frac{r_{yx_1} X_1 + r_{yx_2} X_2 + ... + r_{yx_n} X_n}{r_{yx_1} + r_{yx_2} + ... + r_{yx_n}} \qquad \text{Eq. (1)}$$

em que: r_{yx_j} é o coeficiente de correlação entre os postos citados e n representa o número total de postos vizinhos considerados.

Para aplicação do método adota-se como critério mínimo a obtenção de coeficiente de determinação superior a 0,7 e a existência de pelo menos oito pares de eventos entre as estações para a realização da regressão. Nos casos em que a estação em análise apresenta boa correlação, com apenas uma estação de apoio, utiliza-se, para o preenchimento dos dados de chuva, o preenchimento de regressão linear simples.

4.4 Análise de Agrupamentos

Foi organizada uma matriz de 227 x 372, em que as linhas eram as estações pluviométricas e as colunas os totais mensais em ordem cronológica.

Os dados desta matriz foram transformados para ficarem no intervalo de [0,1], conforme a Equação 2.

$$x_{ij} = \frac{y_{ij} - y_{i(\min)}}{y_{i(\max)} - y_{i(\min)}} \qquad \text{Eq. (2)}$$

Na Equação 2, x_{ij} é o elemento escalonado i da variável j, y_{ij} o valor real do elemento i da variável j e $y_{i(min)}$ e $y_{i(max)}$ são os valores máximo e mínimo da variável j.

A medida de dissimilaridade utilizada foi a distância euclidiana, a qual corresponde à distância geométrica tomada em um espaço de p dimensões. Sendo X_{ij} a observação transformada da i-ésima estação pluviométrica (i = 1, 2,…,n), com referência à j-ésima variável ou frequência absoluta em cada classe (j = 1, 2, …, p) estudada, define-se a distância euclidiana entre dois postos i e i', por meio da Equação 3.

$$d_{ii'} = \sum_{j=1}^{p} \left\| x_{ij} - x_{i'j} \right\|^2 \qquad \text{Eq. (3)}$$

4.4.1 Agrupamentos hierárquicos

Agrupamento hierárquico é uma forma de classificar e agrupar dados, criando uma "árvore de grupos". A árvore não é um único conjunto de grupos, mas uma hierarquia multinível, em que grupos de um nível são unidos como grupos do nível imediatamente superior. O procedimento de agrupamento hierárquico geral é um método recursivo definido pelo seguinte algoritmo iterativo (CORTÉS; PALMA; WILSON, 2007):

(1) As distâncias entre todos os vetores que formam o conjunto de dados são calculadas utilizando-se uma dada métrica. Assumindo que existem n pontos representados por vetores, n-1 distâncias entre eles precisam ser calculadas.

(2) O par de pontos com a menor distância (ou seja, mais próximos ou similares) é agrupado. Para a próxima iteração, o grupo é representado como único ponto de dados, substituindo os que foram agrupados. Existem vários métodos de agrupamento, sendo os mais comuns: ligação simples, ligação completa, ligação centroide, ligação média e *ward*.

(3) O novo vetor é considerado como um único ponto de dados no conjunto, em vez do par de pontos previamente agrupados, que não são considerados na próxima iteração. No próximo passo, um novo cálculo de todas as distâncias entre os pontos, reduzindo de n para n-1, restando n-2 distâncias. Depois de n-1 iterações, todos os dados são agrupados em um único ponto.

Na ligação simples é usada a menor distância entre dois grupos do conjunto. Ao contrário, na ligação completa é usada a maior distância de dois grupos no conjunto. A metodologia da média pode ser vista como um equilíbrio entre as duas anteriores, pois a similaridade é mensurada pela média de similaridade entre os objetos de dois grupos diferentes (WU; XIONG; CHEN, 2009).

No método de ligação pelo centroide é usada a distância entre centroides de dois grupos, usualmente calculada como a média aritmética (CORTÉS; PALMA; WILSON, 2007).

O método de *ward* usa a soma incremental de quadrados, ou seja, o aumento no total da soma quadrada interna devido à junção de dois grupos. A soma quadrada interna de um grupo é definida como a soma das distâncias entre todos os objetos no grupo e seus centroides (OURDA *et al*., 2008).

Na Figura 3 está a representação gráfica e analítica das metodologias de agrupamento hierárquicos.

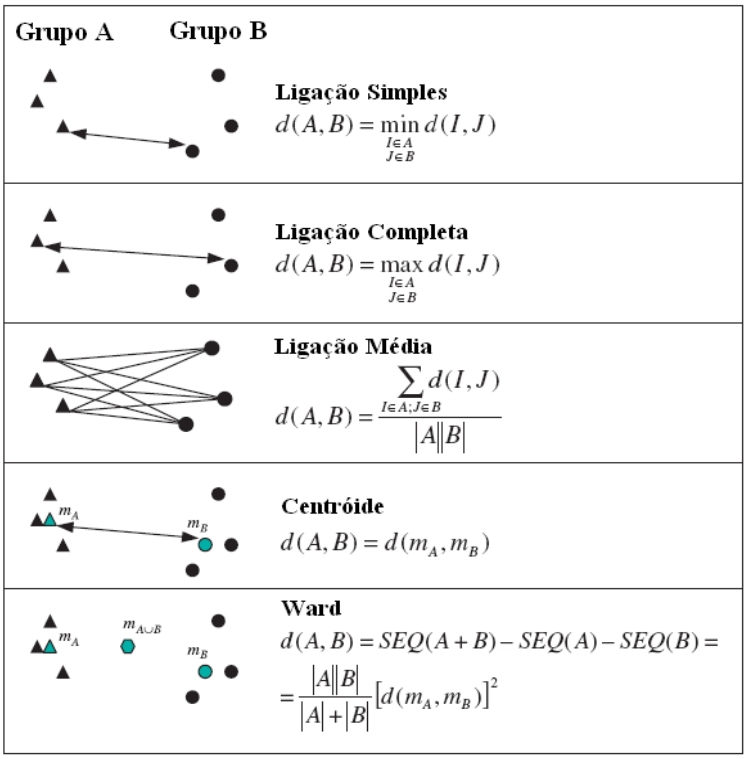

Figura 3 - Representação gráfica e analítica das metodologias de agrupamento hierárquico.

Fonte: Barreto *et al*. (2007).

Nas equações expressas na Figura 3, I e J representam objetos dos grupos, $d(\ )$ é a função de distância, $|\ |$ representa o número de objetos do grupo, $SEQ(\ )$ é a de soma de quadrados e m representa o centroide do grupo.

4.4.2 Algoritmo k-médias

O procedimento k-médias, sendo um algoritmo particional, é baseado no critério do erro quadrado. O objetivo geral é obter uma partição na qual, para um número fixado de grupos, sejam minimizados os erros quadrados (GARCÍA; GONZÁLEZ, 2004), conforme Equação 4.

$$F_{OB} = \sum_{k=1}^{N} \sum_{x \in Q_k} \|x - c_k\|^2 \qquad \text{Eq. (4)}$$

em que: F_{ob} é a função objetivo, N é o número de grupos, c_k é o centro do agrupamento k e x é um objeto que pertence ao grupo Q_k.

O algoritmo k-médias pode ser descrito nos seguintes passos, segundo Guldemir e Sengur (2006):

Passo 1: escolha N grupos iniciais e distribua aleatoriamente seus centros z_1, z_2,..., z_k, nos n pontos $\{X_1, X_2,..., X_n\}$.

Passo 2: atribua o ponto X_i, i=1,2,...,n para o grupo k_j, j=1,2,...,N se $\|X_i - z_j\| < \|X_i - z_p\|$, p=1,2,...,N e j $\neq$ p.

Passo 3: calcule o novo centro do agrupamento, conforme Equação 5.

$$z_i^{novo} = \frac{1}{n_i} \sum_{X_j \in k_i} X_j \qquad \text{Eq. (5)}$$

em que: n_i é o número de elementos pertencentes ao cluster k_i.

Passo 4: se $\|z_i^{novo} - z_i\| < \varepsilon$ ou $z_i^{novo} = z_i$ termine o procedimento, caso contrário continue do passo 2.

A Figura 4 ilustra o algoritmo k-médias descrito nos passos 1 a 4.

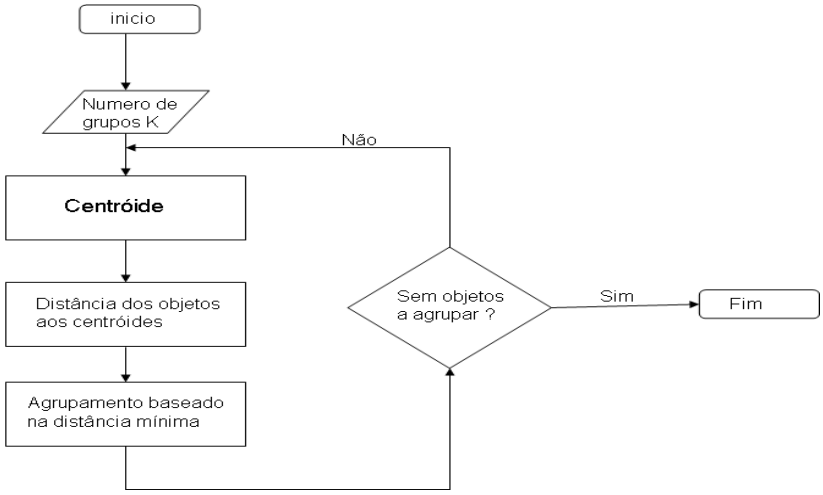

Figura 4 - Representação do algoritmo k-médias na forma de fluxograma.

Na Figura 4, mostra-se o algoritmo k-médias em forma de fluxograma, em que, enquanto houver elementos na base sem pertencerem a nenhum grupo, faz-se a comparação entre os elementos (nova informação comparada aos centroides dos grupos existentes) para poder incluir os novos.

Neste trabalho, além do algoritmo acima, foi utilizada uma forma híbrida, em que os centroides iniciais do algoritmo k-médias eram aqueles obtidos pelas metodologias hierárquicas (CHENG; LIAO, 2009).

4.4.3 Validação dos agrupamentos

A correlação cofenética é uma medida de validação utilizada, principalmente, nos métodos de agrupamento hierárquicos. A ideia básica é realizar uma comparação entre as distâncias efetivamente observadas entre os objetos e as distâncias previstas a partir do processo de agrupamento. A correlação cofenética mede o grau de ajuste entre a matriz de dissimilaridade original (matriz D) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C) Beaver e Palazoğlu (2006), conforme expressão 6.

$$CC = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(c_{ij}-\bar{c}\right)\left(d_{ij}-\bar{d}\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(c_{ij}-\bar{c}\right)^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}-\bar{d}\right)^2}} \qquad \text{Eq. (6)}$$

em que: c_{ij} é o valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética; d_{ij} é o valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz de dissimilaridade;

$$\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}c_{ij} \qquad \text{Eq. (7)}$$

$$\bar{d} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}d_{ij} \qquad \text{Eq. (8)}$$

Nota-se que essa correlação equivale à correlação de Pearson entre a matriz de dissimilaridade original e aquela obtida após a construção do dendrograma. Assim, quanto mais próximo de 1, menor será a distorção provocada pelo agrupamento dos indivíduos com os métodos.

O índice de Davies-Bouldin é baseado na ideia de que uma boa partição é aquela para a qual as medidas de separação entre grupos e a densidade dos grupos apresentam valores altos. O índice é uma função da razão entre a soma da dispersão intragrupos e a separação entre grupos (AMIN-NASERI; SOROUSH, 2008), conforme equações 9, 10, 11 e 12.

$$\bar{R} = \frac{1}{k}\sum_{i=1}^{k}\max_{\{i\neq j\}}\left\{R_{ij}\right\} \qquad \text{Eq. (9)}$$

$$R_{ij} = R\left(s_i^q, s_j^q, d_{ij}\right) = \frac{s_i^q + s_j^q}{d_{ij}} \qquad \text{Eq. (10)}$$

$$s_i^q = \frac{1}{n_i} \sum_{j \in A_i} \left\| x_j - r_i \right\|_q \quad \text{Eq. (11)}$$

$$s_j^q = \frac{1}{n_j} \sum_{j \in A_j} \left\| x_i - r_j \right\|_q \quad \text{Eq. (12)}$$

em que: s_i^q e s_j^q são as dispersões das classes C_i e C_j, respectivamente; r_i e r_j e A_i e A_j são respectivamente os centros de gravidade e os conjuntos dos índices dos elementos das classes C_i e C_j e, finalmente, d_{ij} é a dissemelhança entre os centros das classes C_i e C_j.

Neste trabalho adotou-se q=1 para o cálculo da dispersão dos pontos dentro do agrupamento e a distância euclidiana (d_{ij}) entre centroides, conforme Pakhira, Bandyopadhyay e Maulik (2004).

O índice de Dunn tenta identificar partições de classes compactas e isoladas. O número de classes que maximiza o valor do índice é tomado como o valor de k, adequado ao conjunto de dados. O índice é dado pela Equação 13.

$$ID = \min_{1 \le i \le K} \left\{ \min_{1 \le i \le K, j \ne i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le i \le K} \Delta(C_k)} \right\} \right\} \quad \text{Eq. (13)}$$

em que: $\delta(C_i, C_j)$ representa a distância entre os agrupamentos C_i e C_j, calculada pela Equação 14; $\Delta(C_k)$ representa o diâmetro do grupo C_k, dado pela Equação 15. O valor de k para os quais ID é maximizado é tomado como o número ideal de agrupamentos.

$$\delta(C_i, C_j) = \max_{y_i \in C_i, y_j \in C_j} \{ d(y_i, y_j) \} \quad \text{Eq. (14)}$$

$$\Delta(C_k) = \max_{y_i, y_j \in C_k} \{ d(y_i, y_j) \} \quad \text{Eq. (15)}$$

em que: $d(y_i, y_j)$ é a distância euclidiana entre os objetos y_i e y_j.

Além dos índices de qualidade de agrupamentos foram estabelecidas as seguintes regras na formação dos grupos:

(1) $k_{\min} = 2$;

(2) $k_{\max} \le \sqrt{227} \Rightarrow k_{\max} \le 15$;

(3) número de estações em cada grupo maior ou igual cinco.

4.5 Regionalização

A metodologia de análise de frequência regional da precipitação é baseada em Hosking e Wallis (1997), Nanegheti e Pinto (2007) e Saf (2010). Os quantis adimensionais das regiões homogêneas são estimados, conforme Equação 16.

$$P(F)_i = \mu_i p(F) \qquad \text{Eq. (16)}$$

em que: μ é o fator de adimensionalização, $p(F)$ representa o quantil regional adimensionalisado de não exceder a probabilidade F e $P(F)$ é estimativa do quantil local i.

As séries de um local da região homogênea são adimensionalzidos pela suas respectivas médias, conforme equações 17 e 18.

$$\hat{\mu}_i = \frac{\sum P_i}{n} \qquad \text{Eq. (17)}$$

$$p_{ij} = \frac{P_{ij}}{\hat{\mu}_i},\ j = 1,2,...,n_i \text{ e } i = 1,2,...,N \qquad \text{Eq. (18)}$$

em que: $\hat{\mu}$ é estimador do fator de adimensionalização, P representa a série de valores de precipitação pluviométrica do local i, p_{ij} representa a séries de valores adimensionais e n é número de observações do local i.

Dessa forma, os dados padronizados formam a base para estimar a curva regional de quantis adimensionais. A forma genérica da distribuição é conhecida, a menos dos seus parâmetros, que são próprios da distribuição. Para contornar este problema Hosking e Wallis (1997) propõem que os parâmetros da curva regional de quantis adimensionais sejam obtidos pela ponderação dos parâmetros locais estimados para cada local i, pelos respectivos tamanhos das amostras, conforme Equação 19.

$$\hat{\theta}_k^R = \frac{\sum_1^N n_i \hat{\theta}_k^i}{\sum_1^N n_i} \qquad \text{Eq. (19)}$$

em que: N representa o número de estações, $\hat{\theta}_k^i$ é o momento-L de interesse e n_i representa o tamanho da série.

Substituindo estas estimativas em $q(F)$ produz a estimativa dos quantis regionais, conforme Equação 20.

$$p(F) = p\left(F;\theta_1^R,...,\theta_p^R\right) \quad \text{Eq. (20)}$$

4.5.1 Medida de discordância

Os quocientes de momentos-L de um local i (CV-L, assimetria-L e curtose-L) são considerados como um ponto em um espaço tridimensional, conforme Equação 21.

$$u_i = \left(t^i, t_3^i, t_4^i\right)^T \quad \text{Eq. (21)}$$

em que: t, t_3 e t_4 denotam CV-L, assimetria-L e curtose-L, respectivamente, e o símbolo T indica matriz transposta.

Com os momentos-L locais são então calculados os momentos-L regionais pela média aritmética, conforme Equação 22.

$$\bar{u} = \frac{\sum_1^N u_i}{N} = \left(t^R, t_3^R, t_4^R\right)^T \quad \text{Eq. (22)}$$

A matriz de covariância amostral S é definida pela Equação 23.

$$S = \frac{1}{N-1}\sum_1^N \left(u_i - \bar{u}\right)\left(u_i - \bar{u}\right)^T \quad \text{Eq. (23)}$$

Por fim, segundo Hosking e Wallis (1997), calcula-se a medida de discordância conforme Equação 24.

$$D_i = \frac{N}{3(N-1)}\left(u_i - \bar{u}\right)^T S^{-1}\left(u_i - \bar{u}\right) \quad \text{Eq. (24)}$$

Os valores críticos de D_i em função do número de locais estão descritos na Tabela 1.

Tabela 1 - Valores críticos da medida de discordância

N° de postos na região	$D_{critico}$	N° de postos na região	$D_{critico}$
5	1,333	11	2,632
6	1,648	12	2,757
7	1,917	13	2,869
8	2,140	14	2,971
9	2,329	≥15	3,000
10	2,491		

Fonte: Hosking e Wallis (1997).

4.5.2 Medida de heterogeneidade

A medida de heterogeneidade é baseada nos momentos-L e na teoria que todas as estações de uma região têm a mesma população de momentos-L e avalia se a região estudada é homogênea ou não.

Hosking e Wallis (1997) recomendam que a medida de heterogeneidade, denotada por H, baseie-se preferencialmente no cálculo da dispersão de t, ou seja, o CV-L para as regiões proposta e simulada. Efetua-se o cálculo do desvio padrão ponderado V dos CV-L´s das amostras observadas, conforme Equação 25.

$$V = \left[\frac{\sum_{1}^{N} n_i \left(t^i - t^R \right)^2}{\sum_{1}^{N} n_i} \right]^{0,5} \qquad \text{Eq. (25)}$$

Em seguida, para a simulação da região homogênea é utilizada a distribuição Kappa de quatro parâmetros. Essa distribuição é definida pelos parâmetros ξ, α, k e h e inclui, como casos particulares, as distribuições Logística, Generalizada de Valores Extremos e Generalizada de Pareto, sendo, portanto, teoricamente capaz de representar variáveis hidrológicas e hidrometeorológicas.

A distribuição Kappa de quatro parâmetros pode ser descrita pelas equações 26, 27 e 28.

$$f(x) = \alpha^{-1} \left\{ 1 - \frac{k(x-\xi)}{\alpha} \right\}^{\frac{1}{k-1}} \{F(x)\}^{1-h} \qquad \text{Eq. (26)}$$

$$F(x) = \left[1 - h \left\{ 1 - \frac{k(x-\xi)}{\alpha} \right\}^{\frac{1}{k}} \right]^{\frac{1}{h}} \qquad \text{Eq. (27)}$$

$$x(F) = \xi + \frac{\alpha}{k} \left\{ 1 - \left(\frac{1 - F^h}{h} \right)^k \right\} \qquad \text{Eq. (28)}$$

Os momentos-L da distribuição Kappa são definidos para $h \geq 0$ e $k > -1$ ou para $h < 0$ e $-1 < k < -h^{-1}$, calculados pelas equações 28 a 32.

$$\lambda_1 = \xi + \frac{\alpha(1 - g_1)}{k} \qquad \text{Eq. (29)}$$

$$\lambda_2 = \frac{\alpha(g_1 - g_2)}{k} \qquad \text{Eq. (30)}$$

$$\tau_3 = \frac{-g_1 + 3g_2 - 2g_3}{g_1 - g_2} \qquad \text{Eq. (31)}$$

$$\tau_4 = \frac{-g_1 + 6g_2 - 10g_3 + 5g_4}{g_1 - g_2} \qquad \text{Eq. (32)}$$

$$g_r = \frac{r\Gamma(1+k)\Gamma\left(\frac{r}{h}\right)}{h^{1+k}\Gamma\left(1+k+\frac{r}{h}\right)} \text{ se } h > 0 \qquad \text{Eq. (33a)}$$

$$g_r = \frac{r\Gamma(1+k)\Gamma\left(-k-\frac{r}{h}\right)}{(-h)^{1+k}\Gamma\left(1-\frac{r}{h}\right)} \text{ se } h < 0 \qquad \text{Eq. (33b)}$$

em que: $\Gamma(\ )$ representa a função gama.

Os parâmetros da população Kappa são estimados de modo a reproduzir os quocientes de momentos-L regionais. Com os parâmetros populacionais, são simuladas N_{SIM} regiões homogêneas, sem correlação cruzada e/ou serial, contendo $N_{amostras}$ individuais, cada qual com n_i valores da variável normalizada. Em seguida, as estatísticas V_i (i=1, 2,..., N_{SIM}) são calculadas para todas as simulações de regiões homogêneas, pela Equação 25.

Após as simulações é calculada a média aritmética e o desvio padrão para mensurar a dispersão esperada para uma região homogênea.

$$\mu_v = \frac{\sum_{1}^{N_{SIM}} V_i}{N_{SIM}} \qquad \text{Eq. (34)}$$

$$\sigma_v = \sqrt{\frac{\sum_{1}^{N_{SIM}} (V_i - \mu_v)}{N_{SIM} - 1}} \qquad \text{Eq. (35)}$$

A medida de heterogeneidade H estabelece uma comparação entre a dispersão observada e a dispersão simulada, conforme Equação 36.

$$H = \frac{V - \mu_v}{\sigma_v} \qquad \text{Eq. (36)}$$

Existem mais duas medidas de heterogeneidade. A primeira, V_2, mede a dispersão dos momentos-L amostrais baseada em CV-L e assimetria-L. A segunda medida, V_3, mede a dispersão dos momentos-L amostrais baseada em assimetria-L e curtose-L. Ambas as medidas são descritas nas equações 37 e 38.

$$V_2 = \left\{ \frac{\sum_1^N n_i \left[\left(t^i - t^R \right) + \left(t_3^i - t_3^R \right) \right]^{0,5}}{\sum_1^N n_i} \right\}^{0,5} \qquad \text{Eq. (37)}$$

$$V_3 = \left\{ \frac{\sum_1^N n_i \left[\left(t_3^i - t_3^R \right) + \left(t_4^i - t_4^R \right) \right]^{0,5}}{\sum_1^N n_i} \right\}^{0,5} \qquad \text{Eq. (38)}$$

As medidas de heterogeneidade H_2 e H_3 são calculadas pela Equação 25. De acordo com o teste de significância, proposto por Hosking e Wallis (1997), se $H < 1$ considera-se a região como “aceitavelmente homogênea”, se $1 < H < 2$, a região é “possivelmente heterogênea” e, finalmente, se $H > 2$, a região deve ser classificada como “definitivamente heterogênea”.

4.6 Estimação de Quantis

Em geral, a seleção da “melhor” distribuição de probabilidades baseia-se na qualidade e consistência de seu ajuste aos dados disponíveis. Entretanto, Hosking e Wallis (1997) ponderam que o objetivo da análise regional de frequência não é o de ajustar uma distribuição a uma amostra em particular. De fato, o que se objetiva é a obtenção de estimativas de quantis de uma distribuição de probabilidades da qual se espera serem extraídos futuros valores amostrais. Em outras palavras, o que se preconiza é a seleção, entre diversas candidatas, da distribuição mais robusta, ou seja, aquela mais capaz de produzir boas estimativas de quantis, mesmo que os futuros valores amostrais possam ser extraídos de outra distribuição, algo diferente da que foi ajustada. Foram usadas as

distribuições Gama e Pearson tipo III para modelar a distribuição de frequência da precipitação mensal. A distribuição Gama é dada pela Equação 39.

$$f(x) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{Eq. (39)}$$

em que: $f(x)$ função densidade de probabilidade, β parâmetro de escala, α parâmetro de forma e $\Gamma(\ )$ função gama representada na Equação 40.

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad \text{Eq. (40)}$$

A função de densidade de probabilidade de uma distribuição Pearson tipo III é dada pela Equação 41.

$$f(x) = \frac{1}{\beta\Gamma(\alpha)} \left(\frac{x-\gamma}{\beta} \right)^{\alpha-1} \exp\left(-\frac{x-\gamma}{\beta} \right) \quad \text{Eq. (41)}$$

em que: $f(x)$ função densidade de probabilidade, β parâmetro de escala, α parâmetro de forma, γ parâmetro de posição e $\Gamma(\ )$ função gama representada na Equação 40.

Os parâmetros das distribuições de probabilidade foram estimados utilizando-se a teoria dos momentos-L. A estimação dos momentos-L, a partir de uma amostra de tamanho finito n, inicia-se com a ordenação de seus elementos constituintes em ordem crescente, ou seja, $x_1 \leq x_2 \leq ... \leq x_n$. Os estimadores não-enviesados dos momentos ponderados de probabilidade são dados pelas equações 41 e 42.

$$a_r = \frac{1}{N} \sum_{i}^{N} \frac{\binom{N-i}{r}}{\binom{N-1}{r}} x_i \quad \text{Eq. (41)}$$

$$b_r = \sum_{i}^{r} \binom{r}{i} (-1)^i a_i \quad \text{Eq. (42)}$$

Os estimadores não-enviesados dos momentos-L amostrais são definidos pelas equações 43 a 46.

$$l_1 = b_0 \quad \text{Eq. (43)}$$

$$l_2 = 2b_1 - b_0 \quad \text{Eq. (44)}$$

$$l_3 = 6b_2 - 6b_1 + b_0 \quad \text{Eq. (45)}$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \quad \text{Eq. (46)}$$

Da mesma forma, os quocientes de momentos-L amostrais são dados pelas equações 47 a 49.

$$t = \frac{l_2}{l_1} \quad \text{Eq. (47)}$$

$$t_3 = \frac{l_3}{l_2} \quad \text{Eq. (48)}$$

$$t_4 = \frac{l_4}{l_2} \quad \text{Eq. (49)}$$

em que: t representa CV-L, t_3 representa assimetria-L, t_4 representa curtose-L.

Estimado os quocientes dos momentos-L foi realizada a estimativa dos parâmetros das distribuições de probabilidade.

Os parâmetros da distribuição gama foram estimados pelas equações 50 e 51.

$$\frac{l_1}{l_2} = \frac{\Gamma(\alpha + 0{,}5)}{\sqrt{\pi}\Gamma(\alpha + 1)} \quad \text{Eq. (50)}$$

$$\beta = \frac{l_1}{\alpha} \quad \text{Eq. (51)}$$

Os parâmetros da Pearson tipo III foram estimados pelas equações 52 a 54.

$$t_3 = 6I_{\frac{1}{3}}(\alpha, 2\alpha) - 3 \quad \text{Eq. (52)}$$

$$l_2 = \pi^{-0{,}5} + \beta\frac{\Gamma(\alpha + 0{,}5)}{\Gamma(\alpha)} \quad \text{Eq. (53)}$$

$$l_1 = \gamma + \alpha\beta \quad \text{Eq. (54)}$$

em que:

$$I_x(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^x t^{p-1}(1-t)^{q-1}dt \quad \text{Eq. (55)}$$

Estimados os parâmetros das distribuições de probabilidade foi necessário verificar o ajuste das frequências teóricas com as frequências empíricas. Para este propósito, foi utilizado o teste de Kolmogorov-Smirnov, que tem como base a máxima diferença entre as funções de probabilidade acumuladas (empírica e teórica), conforme Equação 56.

$$D_N = \sup_{-\infty<x<\infty} \left|F_N(x) - F_X(x)\right| \quad \text{Eq. (56)}$$

em que: F_N frequência empírica e F_X frequência teórica estimada pela distribuição gama ou Pearson tipo III.

Foi verificado o ajuste das distribuições aos dados adimensionalisados de cada estação (Equação 18) e estimados os quantis locais.

Foram estimados os parâmetros regionais de cada grupo, conforme Equação 19, com os dados de cada estação, adimensionalisados conforme Equação 18. Em seguida, a frequência teórica regional foi confrontada à frequência empírica de cada estação pertencente ao grupo pelo teste de Kolmogorov-Smirnov, conforme Equação 57.

$$D_R = \max\left[\sup_{-\infty<x<\infty} \left|F_i(x) - F_R(x)\right|\right]; \text{ i=1, 2,..., N} \quad \text{Eq. (57)}$$

em que: F_i é a frequência empírica adimensional, F_R é a frequência teórica regional adimensional e N é o número de estações do grupo.

Verificados os ajustes das distribuições de probabilidade locais e regionais foram estimados os quantis. Dessa forma, foi possível verificar o erro cometido pela regionalização dos parâmetros das distribuições de probabilidade teóricas, como indicam Ouarda *et al*. (2008).

$$RSME = \sqrt{\frac{1}{N}\sum_{1}^{N}\left(q_r - q_{li}\right)^2} \quad \text{Eq. (58)}$$

$$E(\%) = \frac{RSME}{\bar{q}_r} \times 100 \quad \text{Eq. (59)}$$

$$\bar{q}_r = \frac{1}{N}\sum_{1}^{N} q_{li} \quad \text{Eq. (60)}$$

em que: RSME é a raiz do erro quadrático médio, N é o número de estações do grupo, q_r é a estimativa do quantil regional, q_l é a estimativa do quantil local e E é o erro médio percentual.

5 RESULTADOS E DISCUSSÃO

5.1 Seleção do Número de Grupos

5.1.1 Algoritmos hierárquicos

A Tabela 2 mostra os valores do coeficiente de correlação cofenética obtidos com as metodologias utilizadas neste trabalho.

Tabela 2 - Coeficiente de correlação cofenético das metodologias de ligação hierárquica

Método	CCC
Centroide	0,65
Simples	0,71
Completa	0,75
Média	0,80
Ward	0,55

Pode-se observar na Tabela 2 que, para as metodologias de ligação simples, completa e média, os coeficientes cofenéticos ficaram acima de 0,7, isso demonstra que tais métodos tendem a distorcer menos a matriz de distâncias. Os piores desempenhos foram dos métodos *ward* e centroide, indicando que estas metodologias aumentam a distorção da matriz de distâncias em cada passo do algoritmo, principalmente a primeira.

Na Figura 5 é plotado o gráfico da função objetivo em relação ao número de grupos das metodologias de agrupamento hierárquico.

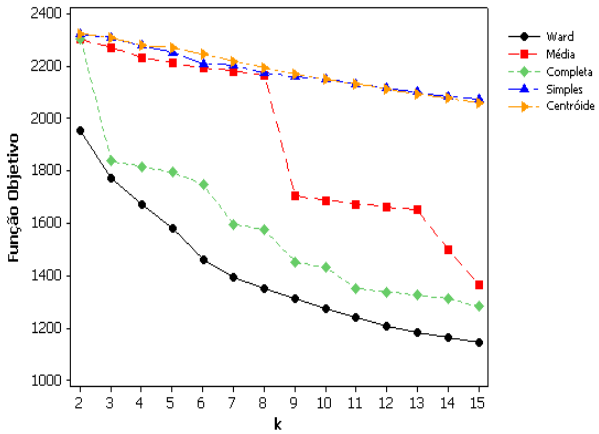

Figura 5 - Gráfico da função objetivo para as metodologias de agrupamentos hierárquicos.

Observa-se pela Figura 5 que as metodologias simples, centroide e *ward* apresentam monotocidade da função objetivo, dificultando a localização de mínimo local. Por outro lado, as metodologias completa e média apresentam oscilações facilmente identificáveis, facilitando a localização de mínimo local.

Para solucionar este problema foi utilizada uma metodologia para localizar mínimo local, conforme Equação 61.

$$\min\left(\left(F\left(k+1\right)-F\left(k\right)\right)-\left(F\left(k\right)-F\left(k-1\right)\right)\right) \qquad \text{Eq. (61)}$$

em que: F é a função objetivo e k é o número de grupos.

A Tabela 3 mostra os valores da função objetivo para os métodos de agrupamento hierárquico estudados, bem como os mínimos locais.

Tabela 3 - Função objetivo das metodologias de ligação hierárquica em função do número de grupos (k)

k	Centroide	Simples	Completa	Média	*Ward*
2	2324,5	2324,5	2302,2	2302,2	1956,2
3	2309,1	2309,1	1837,2	2272,0	1774,1
4	2278,8	2278,8	1815,9	2233,2	1672,7
5	**2271,4**	**2252,5**	1794,0	2211,2	**1582,4**
6	2245,2	2209,8	**1749,2**	2193,2	1462,3
7	2219,4	2201,9	1595,8	2180,2	1395,6
8	2193,9	2176,4	1574,8	**2166,3**	1350,1
9	2169,0	2158,5	1452,5	1705,9	1313,7
10	2149,7	2151,1	1430,0	1687,1	1272,3
11	2131,0	2132,3	1350,9	1673,2	1240,3
12	2112,9	2115,1	1337,8	1662,7	1207,9
13	2095,1	2100,3	1324,0	1649,9	1181,1
14	2077,0	2085,1	1311,7	1501,1	1161,5
15	2059,7	2074,5	1282,8	1366,7	1142,8

Pode-se observar na Tabela 3 que as metodologias centroide e simples são as que menos reduzem a função objetivo. A função objetivo é bastante reduzida pelo método de *ward*, também constatado por Rao e Srinivas (2006b).

Apesar destes resultados, as metodologias de ligação simples, centroide, média e completa não produzem grupos de tamanhos homogêneos e com qualidade satisfatória para uso posterior, pois são sensíveis a valores atípicos, ou seja, meses com valores extremos de precipitação pluviométrica.

As tabelas 4 a 8 mostram o tamanho do grupo em função do número de grupos escolhidos. Nota-se que as metodologias simples e centroide tendem a formar um grande grupo com quase todas as estações e conforme o número de grupos aumenta há uma segregação em grupos unitários. O método de ligação média forma grandes grupos que são divididos apenas quando há um aumento significativo do número de grupos (2, 9, 14), além disso, possui segregação de grupos unitários.

Tabela 4 - Tamanho do grupo em função do número de grupos para a metodologia hierárquica de *ward*

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	141	111	111	30	30	30	30	30	30	30	30	30	20	20
2	86	86	37	81	81	42	42	19	19	19	19	19	19	19
3		30	30	37	37	39	39	23	23	23	23	23	23	23
4			49	30	25	37	20	39	39	22	22	22	10	10
5				49	49	25	25	20	20	17	17	17	22	22
6					5	49	17	25	25	20	20	20	17	17
7						5	49	17	17	25	25	25	20	9
8							5	49	13	17	17	17	25	25
9								5	36	13	13	13	17	17
10									5	36	18	10	13	11
11										5	18	18	10	13
12											5	8	18	10
13												5	8	18
14													5	8
15														5

Tabela 5 - Tamanho do grupo em função do número de grupos para a metodologia hierárquica centroide

	Número de grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	224	224	223	223	222	221	220	219	218	217	216	215	214	213
2	3	2	1	1	1	1	1	1	1	1	1	1	1	1
3		1	2	1	1	1	1	1	1	1	1	1	1	1
4			1	1	1	1	1	1	1	1	1	1	1	1
5				1	1	1	1	1	1	1	1	1	1	1
6					1	1	1	1	1	1	1	1	1	1
7						1	1	1	1	1	1	1	1	1
8							1	1	1	1	1	1	1	1
9								1	1	1	1	1	1	1
10									1	1	1	1	1	1
11										1	1	1	1	1
12											1	1	1	1
13												1	1	1
14													1	1
15														1

Tabela 6 - Tamanho do grupo em função do número de grupos para a metodologia hierárquica simples

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	224	224	223	222	220	220	219	218	218	217	216	215	214	213
2	3	2	1	1	1	1	1	1	1	1	1	1	1	1
3		1	2	1	1	1	1	1	1	1	1	1	1	1
4			1	2	2	2	1	1	1	1	1	1	1	1
5				1	1	1	2	1	1	1	1	1	1	1
6					2	1	1	2	1	1	1	1	1	1
7						1	1	1	1	1	1	1	1	1
8							1	1	1	1	1	1	1	1
9								1	1	1	1	1	1	1
10									1	1	1	1	1	1
11										1	1	1	1	1
12											1	1	1	1
13												1	1	1
14													1	1
15														1

Tabela 7 - Tamanho do grupo em função do número de grupos para a metodologia hierárquica completa

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	222	161	161	161	161	68	68	68	68	68	68	68	68	68
2	5	61	60	60	23	93	93	71	71	33	33	33	33	33
3		5	1	1	37	23	23	22	22	38	38	38	38	38
4			5	2	1	37	35	23	2	22	22	22	22	22
5				3	2	1	2	35	35	2	2	2	1	1
6					3	2	1	2	21	35	35	35	1	1
7						3	2	1	2	21	21	21	35	21
8							3	2	1	2	1	1	21	14
9								3	2	1	1	1	1	21
10									3	2	1	1	1	1
11										3	2	2	1	1
12											3	1	2	1
13												2	1	2
14													2	1
15														2

Tabela 8 - Tamanho do grupo em função do número de grupos para a metodologia hierárquica média

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	222	221	219	219	218	218	218	114	114	114	114	113	113	91
2	5	1	2	2	1	1	1	104	102	101	101	1	1	1
3		5	1	1	2	1	1	1	1	1	1	101	59	59
4			5	2	1	1	1	1	2	1	1	1	1	22
5				3	2	1	1	1	1	2	1	1	1	1
6					3	2	2	1	1	1	1	1	1	1
7						3	1	2	1	1	1	1	42	1
8							2	1	2	1	1	1	1	42
9								2	1	2	2	1	1	1
10									2	1	1	2	1	1
11										2	2	1	2	1
12											1	2	1	2
13												1	2	1
14													1	2
15														1

Rao e Srinivas (2006b) também constatam que a metodologia de ligação simples tende a formar um grande grupo e diversos pequenos grupos. Além disso, também chegam à conclusão da ineficiência do coeficiente cofenético na identificação do número de grupos.

A metodologia de *ward* foi aquela que apresentou os melhores resultados, pois há uma tendência em cada passo do algoritmo em dividir os grupos com maior número de objetos, criando, dessa forma, grupos mais homogêneos e com qualidade para análises posteriores.

Das análises já feitas, conclui-se que a metodologia hierárquica de *ward* apresenta os melhores resultados no agrupamento de estações pluviométricas, baseado em séries históricas. A análise de pertinência do número de grupos será feita apenas para este método, conforme mostra a Figura 6, na qual estão dispostos os índices de Dunn e Davies-Bouldin em função do número de grupos. A partição com o maior valor do índice de Dunn e o menor valor do Davies-Boulding são consideradas ideais ou otimizadas.

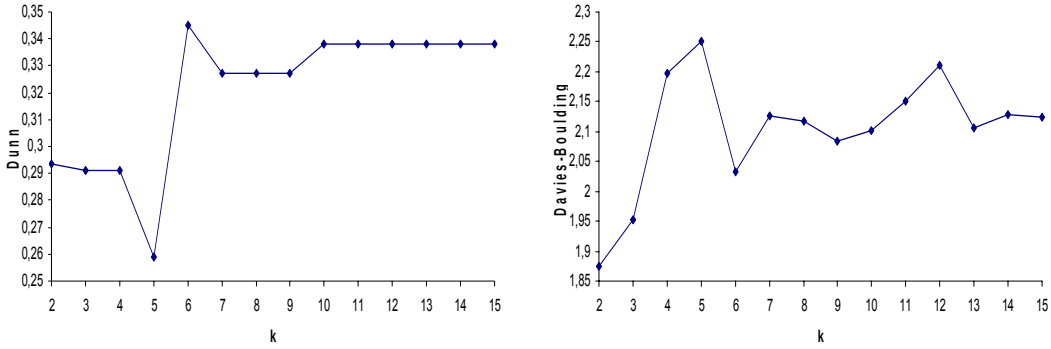

Figura 6 - Índices de Dunn e Davies-Bouldin, em função do número de grupos (k).

Na Figura 6, observa-se que o índice de Dunn apresenta uma oscilação entre um mínimo e um máximo para, em seguida, se manter praticamente constante. O índice de Davies-Boulding apresenta oscilações maiores que Dunn, além disso, apresenta um mínimo global e um mínimo local.

Para o índice de Dunn é facilmente identificável a partição ideal, sendo ela igual seis, pois neste ponto o índice apresenta seu máximo global. O índice de Davies-Boulding apresenta um mínimo global para dois agrupamentos e um mínimo local para seis, porém dois grupos não é uma escolha interessante, porque resultará em grupos com número elevado de estações, o que não é interessante para a regionalização hidrológica. Sendo assim, os índices corroboram a escolha do número de grupos igual a seis. A Figura 7 mostra a configuração final das estações pelo agrupamento hierárquico de *ward* com seis grupos.

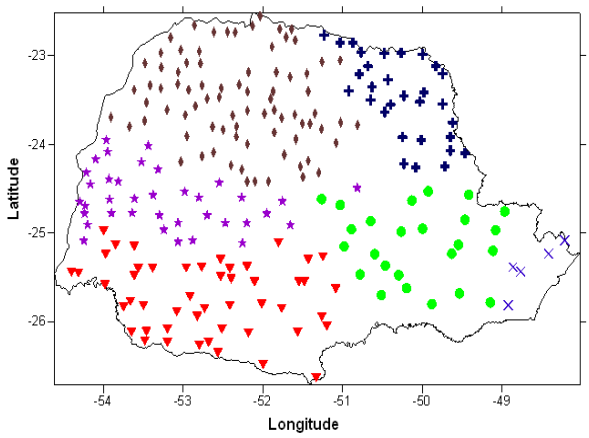

Figura 7 - Configuração final pelo método hierárquico de *ward*.

5.1.2 Algoritmo k-médias

A Figura 8 mostra a função objetivo da metodologia k-médias, assim como suas formas híbridas, ou seja, o algoritmo k-médias inicia com os centroides calculados pelas metodologias hierárquicas.

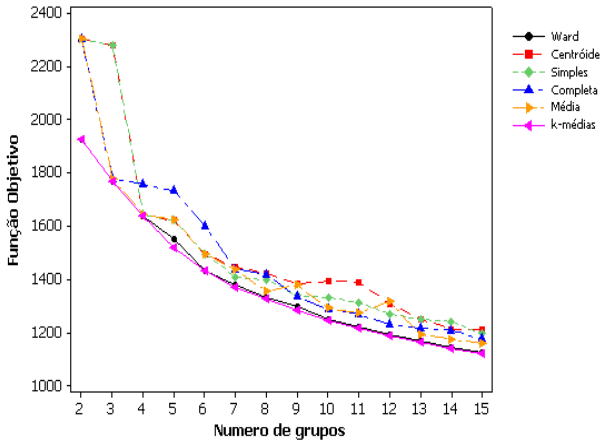

Figura 8 - Gráfico da função objetivo para a metodologia k-médias e suas formas híbridas.

Na Figura 8, pode-se observar que os métodos híbridos centroide, simples, completa e média possuem um comportamento muito semelhante na função objetivo. O método híbrido *ward* e o método k-médias possuem os menores valores da função objetivo, independentemente do número de grupos, além disso, suas curvas estão sobrepostas, apresentando visível monotocidade.

Na Tabela 9 são apresentados os valores da função objetivo em função do número de grupos e da metodologia empregada.

Tabela 9 - Função objetivo das metodologias híbridas e k-médias

k	Centroide	Simples	Completa	Média	*Ward*	k-médias
2	2302,20	2302,20	2302,20	2302,20	1925,45	1923,19
3	1780,66	1780,66	1780,66	1780,66	1768,71	1766,02
4	1758,67	1758,67	1759,38	1759,38	1637,56	1636,06
5	1744,84	1624,99	1737,39	1737,39	**1550,38**	1515,52
6	1611,17	1497,51	1606,01	**1717,98**	1432,25	1430,37
7	1483,69	**1704,16**	1534,74	1586,60	1376,62	1371,36
8	**1696,76**	1572,77	**1515,33**	**1572,77**	1334,12	**1325,30**
9	1565,38	1444,43	1381,43	1446,60	1305,70	1281,17
10	1426,02	1437,03	1317,54	1427,01	1258,73	1245,69
11	1360,21	1381,70	1273,98	1300,71	1231,70	1214,91
12	1322,32	1308,79	1234,04	1260,32	1198,21	**1188,49**
13	1251,46	1255,42	1220,22	1224,71	1172,54	1162,92
14	1216,27	1242,55	1209,15	1178,41	1146,42	**1141,57**
15	1196,01	1195,55	1148,15	1165,19	1124,44	1120,80

Comparando-se as metodologias híbridas centroide, simples e média com seus algoritmos hierárquicos (Tabela 3), observa-se a sensível redução na função objetivo. Já para as metodologias híbridas *ward* e completa não houve redução substancial da função objetivo quando comparada com seus algoritmos hierárquicos.

As tabelas 10 a 15 mostram o número de objetos em função do número de grupos para as metodologias híbridas e para k-médias. Nota-se que as metodologias híbridas centroide, simples, completa e média tendem a produzir grupos de tamanho unitário ou com poucos objetos com o aumento do número de grupos. Isso demonstra a sensibilidade destas metodologias para estações que apresentam alta variabilidade temporal ou estações que possam ser consideradas *outliers,* devido às suas características intrínsecas.

Tabela 10 - Tamanho do grupo em função do número de grupos para metodologia híbrida *ward*

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	135	112	95	44	44	40	37	30	30	29	29	29	24	24
2	92	84	54	54	54	50	35	16	16	16	16	16	14	14
3		31	29	52	52	33	29	27	27	28	26	26	26	26
4			49	29	25	34	24	29	26	26	25	25	14	14
5				48	47	25	25	24	25	20	18	18	26	26
6					5	40	26	24	24	21	21	19	18	16
7						5	46	26	22	23	23	22	21	15
8							5	46	28	17	18	17	17	17
9								5	24	20	16	14	13	13
10									5	22	17	11	14	10
11										5	13	17	10	12
12											5	8	17	10
13												5	8	17
14													5	8
15														5

Tabela 11 - Tamanho do grupo em função do número de grupos para metodologia híbrida centroide

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	222	135	135	135	110	95	135	110	95	49	49	46	44	39
2	5	5	87	87	49	48	85	50	24	66	66	50	51	28
3		87	2	2	63	54	1	1	1	44	18	22	28	31
4			3	1	2	2	1	1	45	1	43	29	21	4
5				2	1	1	1	1	1	34	1	1	14	23
6					2	25	1	1	2	1	19	24	1	24
7						2	2	60	1	1	1	1	21	1
8							1	2	51	1	1	26	1	20
9								1	5	27	1	1	19	1
10									2	2	25	1	1	27
11										1	2	23	1	1
12											1	2	22	1
13												1	2	24
14													1	2
15														1

Tabela 12 - Tamanho do grupo em função do número de grupos para metodologia híbrida simples

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	222	135	135	110	95	135	110	95	95	88	44	42	44	28
2	5	5	87	49	48	1	50	47	47	32	53	31	3	3
3		87	2	63	54	1	1	54	54	47	50	37	32	28
4			3	2	2	2	1	1	1	29	27	25	24	30
5				3	3	1	2	1	1	1	23	21	21	23
6					25	85	1	2	1	1	1	1	1	20
7						2	60	1	1	1	1	1	1	1
8							2	24	24	1	1	41	41	1
9								2	2	24	1	1	1	41
10									1	2	23	1	1	1
11										1	2	23	22	1
12											1	2	2	22
13												1	33	2
14													1	25
15														1

Tabela 13 - Tamanho do grupo em função do número de grupos para metodologia híbrida completa

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	222	135	135	135	110	70	70	49	44	40	40	40	40	33
2	5	87	85	85	50	75	75	66	51	49	48	48	48	28
3		5	2	2	60	37	37	27	25	24	22	22	22	23
4			5	2	2	38	38	34	41	25	23	23	22	22
5				3	2	2	1	44	27	26	26	26	3	27
6					3	2	1	1	32	26	26	26	24	23
7						3	2	1	1	30	18	18	26	20
8							3	2	1	1	17	17	18	17
9								3	2	1	1	1	17	11
10									3	2	1	1	1	16
11										3	2	2	1	1
12											3	1	2	1
13												2	1	2
14													2	1
15														2

Tabela 14 - Tamanho do grupo em função do número de grupos para metodologia híbrida média

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	222	135	135	135	135	110	110	87	87	46	46	41	31	33
2	5	87	85	85	85	50	50	50	48	47	18	40	39	39
3		5	2	2	1	60	60	59	5	29	29	18	21	20
4			5	2	1	1	1	24	56	24	30	22	28	21
5				3	2	1	1	1	24	50	23	26	25	3
6					3	2	2	1	1	24	21	22	14	24
7						3	1	2	1	1	1	17	19	14
8							2	1	2	1	1	1	21	19
9								2	1	2	2	1	1	21
10									2	1	1	2	1	1
11										2	2	1	2	1
12											53	2	1	2
13												34	2	1
14													22	2
15														26

Tabela 15 - Tamanho do grupo em função do número de grupos para metodologia k-médias

	Número de Grupos													
Grupo	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	135	87	67	48	5	5	5	22	24	22	5	18	21	1
2	92	5	5	5	47	24	50	25	24	24	31	29	5	5
3		135	52	95	52	48	29	5	30	35	28	27	17	1
4			103	54	25	50	25	25	28	21	1	19	1	27
5				25	54	29	29	29	26	1	21	16	20	20
6					44	25	21	26	41	26	22	18	16	16
7						46	22	35	2	20	39	26	21	24
8							46	23	25	5	19	23	15	13
9								37	22	22	1	2	1	15
10									5	37	25	3	11	18
11										14	21	23	23	28
12											14	22	27	17
13												1	29	16
14													20	15
15														11

As figuras 9 e 10 mostram o número de objetos em função do número de grupos para a metodologia híbrida de *ward* e k-médias, respectivamente.

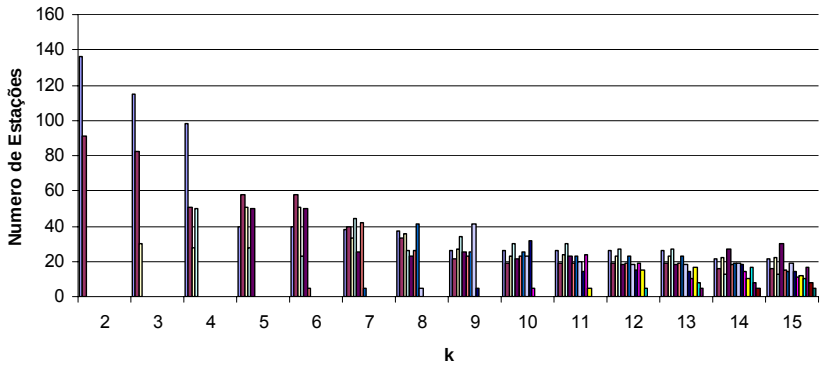

Figura 9 - Número de objetos, em função do número de grupos, para a metodologia de agrupamento híbrido *ward*.

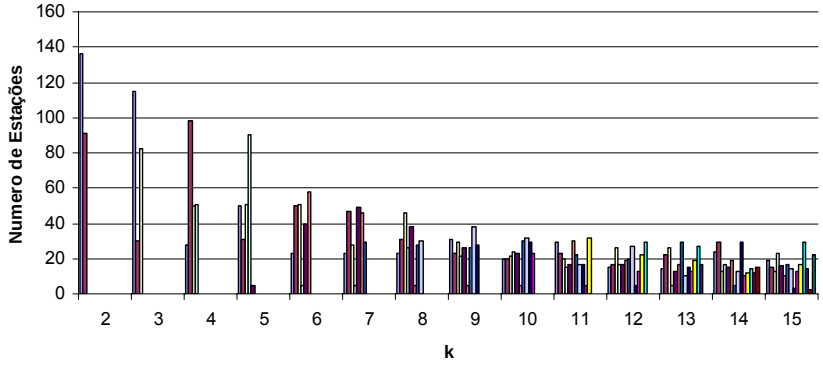

Figura 10 - Número de objetos, em função do número de grupos, para a metodologia de agrupamento k-médias.

Comparando k-médias, método híbrido *ward* e algoritmo hierárquico *ward* observa-se grande semelhança nos grupos obtidos, bem como na função objetivo. Em contraste, as demais metodologias hierárquicas são profundamente modificadas pelo método k-médias (algoritmo híbrido). Conclusão semelhante à dos autores Rao e Srinivas (2006b).

Apenas as metodologias k-médias e sua forma híbrida com *ward* foram avaliadas quanto à pertinência do número de grupos, pois não apresentam a tendência de formar grupos com menos de cinco estações. As figuras 11 e 12 apresentam os índices dos agrupamentos para as metodologias híbrida de *ward* e k-médias, respectivamente.

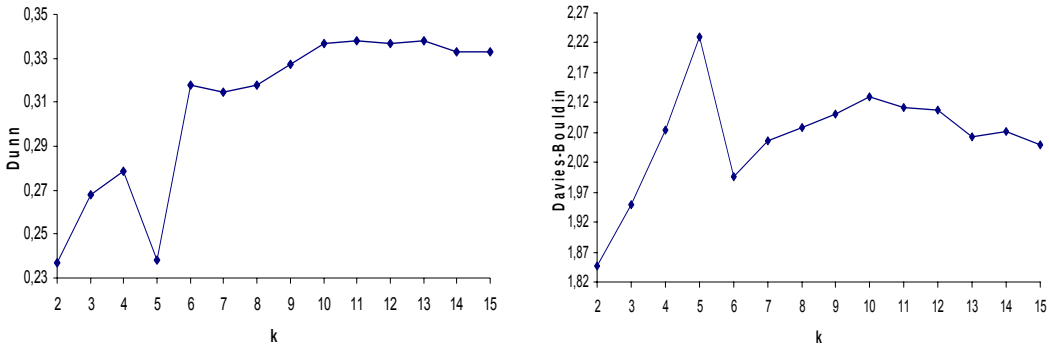

Figura 11 - Índices de Dunn e Davies-Bouldin, em função do número de grupos (k), para o método híbrido *ward*.

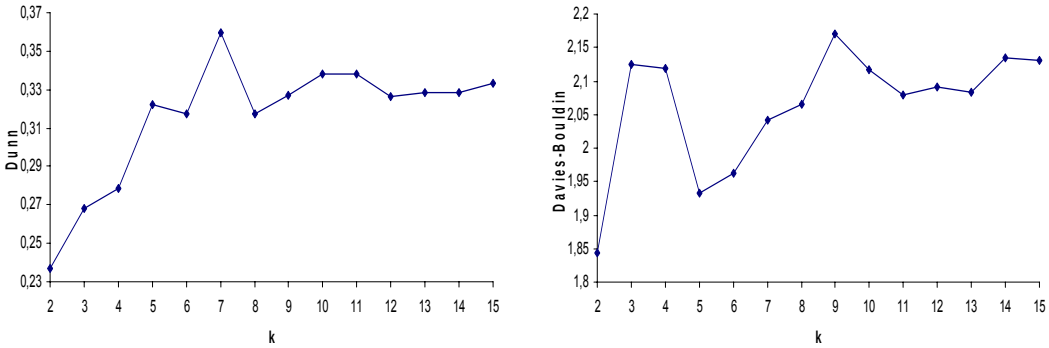

Figura 12 - Número de objetos, em função do número de grupos, para a metodologia de agrupamento k-médias.

A Figura 11 mostra que o índice de Dunn tem sua maior variação quando este sobe de 5 para 6 grupos e seu máximo global está situado entre 10 e 13 grupos. O índice de Davies-Boulding tem um mínimo global para dois grupos e um mínimo local para seis grupos e seu máximo global é alcançado em 5 grupos. Desta forma, os índices evidenciam que seis classes é a melhor partição para o método híbrido de *ward*.

Na Figura 12 o índice de Dunn tem um máximo global para 7 grupos, já o índice de Davies-Boulding tem um mínimo global para dois grupos e um mínimo local para 5 grupos, além disso, possui um máximo global para 9 grupos. Para metodologia k-médias, os índices não corroboraram o mesmo número de grupos, dessa forma, serão averiguadas as três soluções.

As Figuras 13, 14, 15 e 16 mostram as configurações finais para o método híbrido e as três soluções do método k-médias. As soluções das Figuras 12 e 13 são semelhantes, pois a configuração de três grupos é idêntica e os demais grupos são redistribuídos. As soluções apresentadas nas figuras 14 e 15 são interessantes para estudos de regionalização hidrológica, pois apresentam grupos com tamanhos uniformes e bem distribuídos.

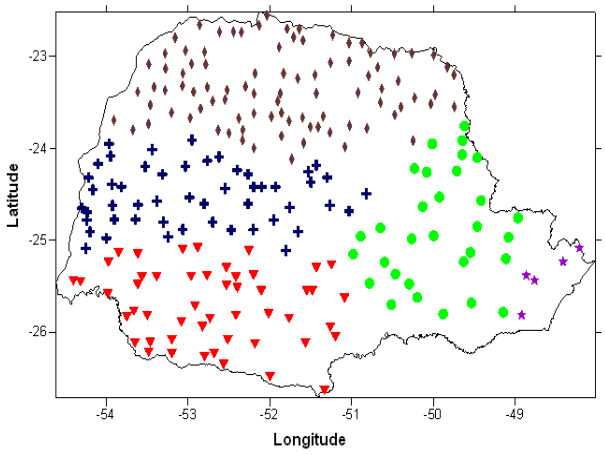

Figura 13 - Configuração final pelo método híbrido de *ward*.

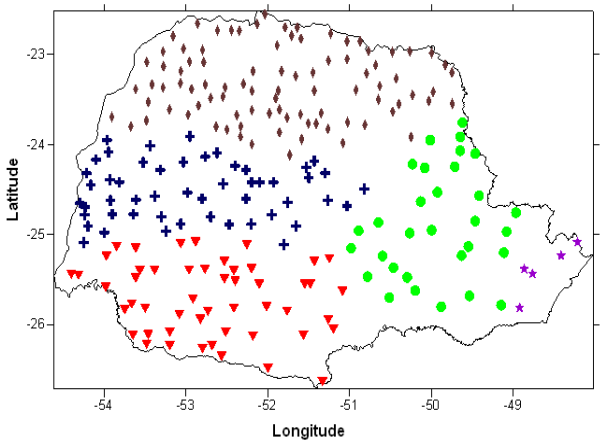

Figura 14 - Configuração final pelo método k-médias para cinco grupos.

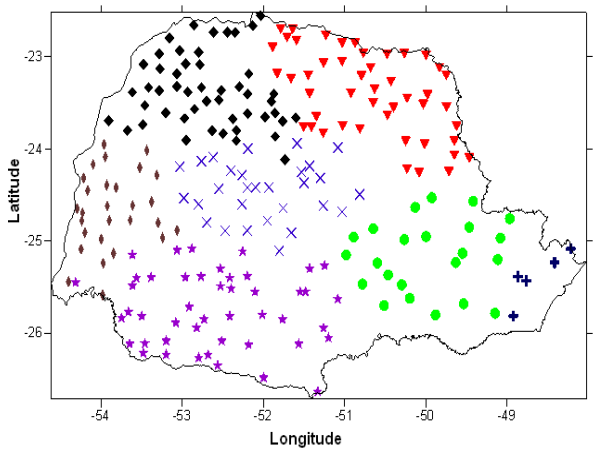

Figura 15 - Configuração final pelo método k-médias para sete grupos.

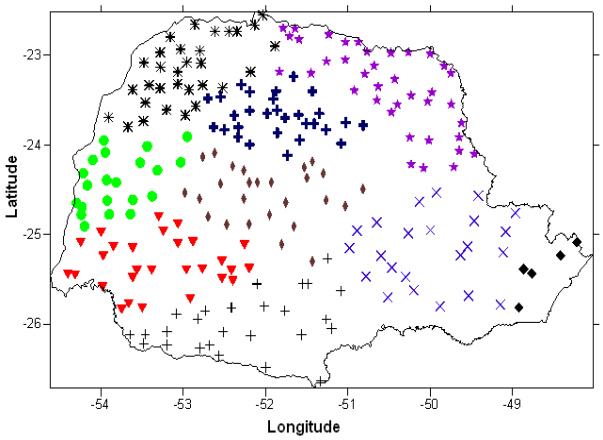

Figura 16 - Configuração final pelo método k-médias para nove grupos.

5.2 Teste de Discrepância e Heterogeneidade

Foram obtidas cinco soluções pelas metodologias de agrupamento multivariados: uma para o método de hierárquico de *ward* e sua forma híbrida, três para k-médias. Obteve-se como partição ideal cinco, seis, sete e nove grupos (Apêndices A e B). Cada grupo foi submetido ao teste de heterogeneidade e discrepância. Os testes foram efetuados sobre os 372 registros de cada estação e não foi dada preferência a nenhum período específico.

5.2.1 Algoritmos hierárquicos

Os grupos foram submetidos à medida de discordância, conforme Figura 17. Observa-se que as estações x38, x102, x109, x212, x182 e x196 são discordantes do grupo ao qual pertencem, pois os valores da medida de discordância ultrapassam a linha de corte.

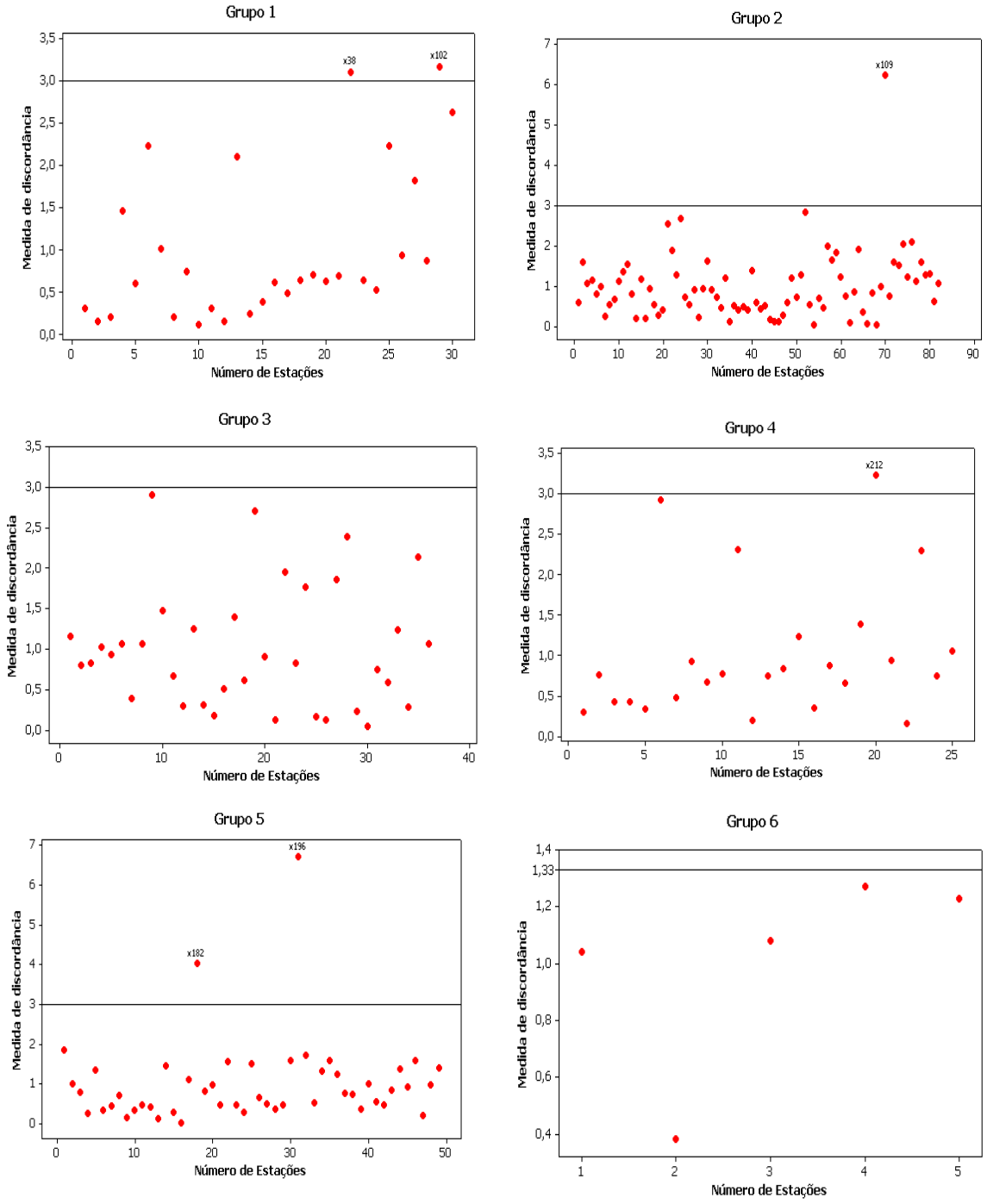

Figura 17 - Medida de discordância para a solução obtida pela metodologia hierárquica de *ward*.

Para complementar estes resultados foram confeccionados diagramas de dupla massa das estações discrepantes em relação às demais do mesmo grupo, conforme Figura 18.

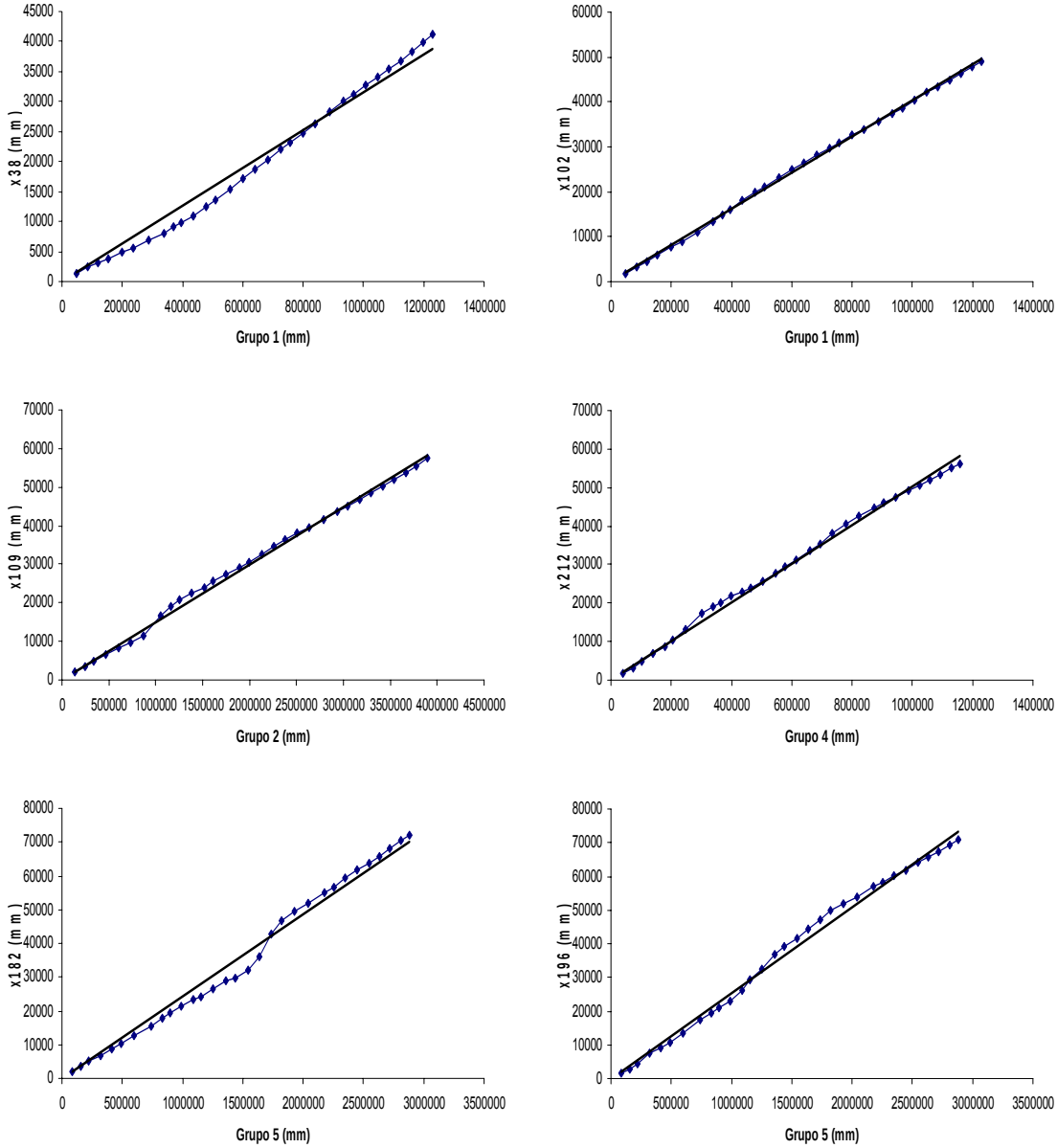

Figura 18 - Diagrama de dupla massa para totais anuais das estações discrepantes em relação ao grupo a qual pertencem.

Observa-se que o teste de discrepância foi eficiente em detectar estações que apresentam anormalidade, pois os diagramas de dupla massa mostram que há desvios entre a estação discrepante e o grupo ao qual pertence. Saf (2010) aponta tendências e mudanças das séries de dados de precipitação como importantes causas de discordância.

A estação x102 apresenta-se linearmente no gráfico de dupla massa, sem desvios consideráveis, porém, ela foi detectada pela medida de discrepância, pois, possivelmente seja um *outlier*. A Figura 19 mostra os gráficos do CV-L e assimetria-L da solução obtida pela metodologia hierárquica de *ward*.

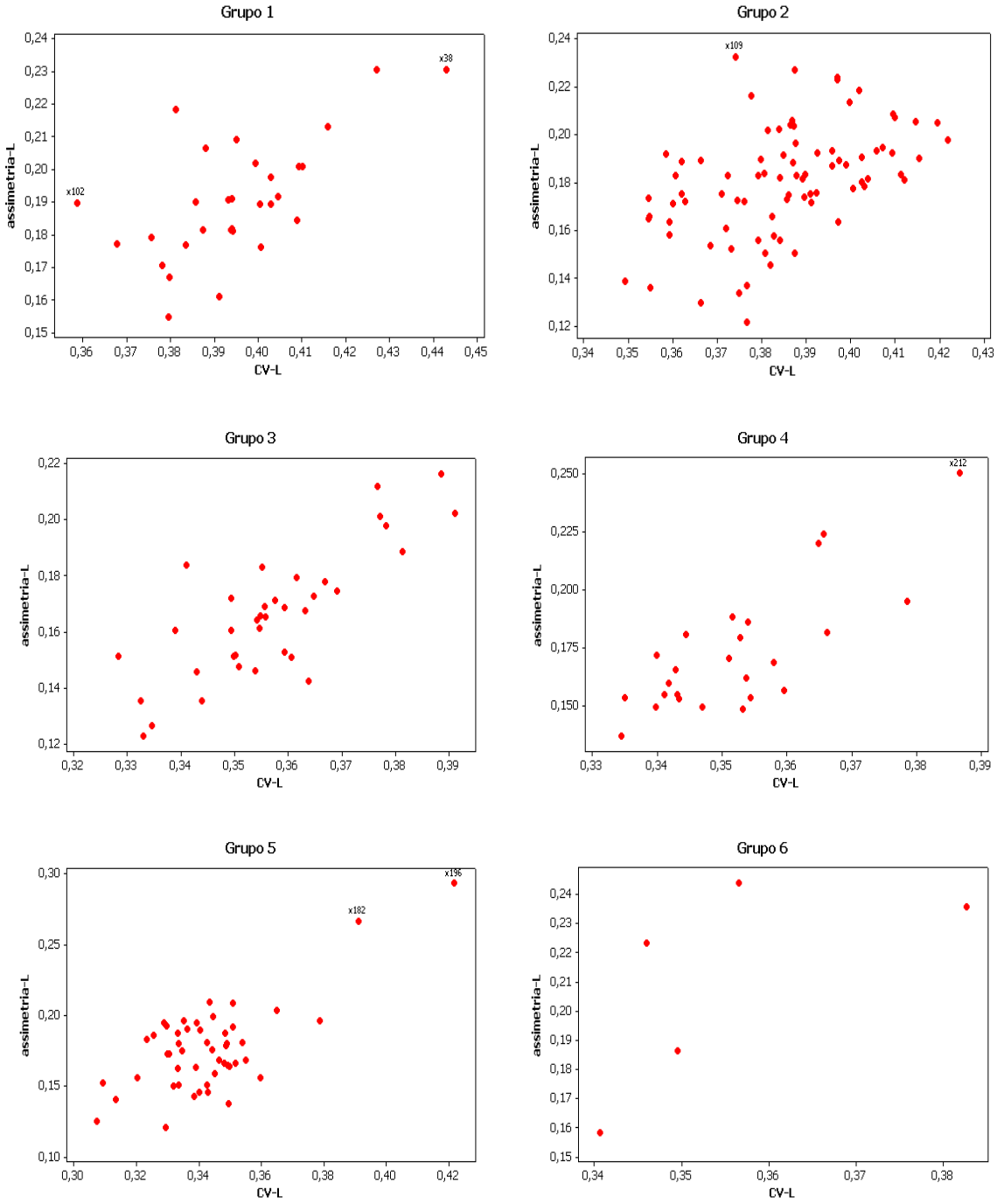

Figura 19 - Gráfico de CV-L e assimetria-L para os grupos obtidos pela metodologia hierárquica de *ward*.

A Figura 19 mostra que as estações discordantes apresentam momentos-L que diferem significativamente do grupo ao qual pertencem, provavelmente provocados pelas distorções apresentadas na Figura 18 ou por possuir variações intrínsecas do local onde está instalada. Isso corrobora Fowler e Kilsby (2003) que identificaram precipitações extremas pontuais como causa de altos valores de curtose-L.

Dessa forma, fica demonstranda a robustez da metodologia, pois conseguiu identificar estações com anormalidades dos seus dados, bem como, possíveis *outlier*s.

Na Tabela 16 estão as seis estações discrepantes de um total de 227, que representa menos de 3% das estações estudadas. Os grupos três e seis não apresentaram estações discrepantes.

Tabela 16 - Estações discrepantes da metodologia hierárquica de *ward* com seis grupos

Grupo	N	Estação (ID)	D
1	30	x38	3,11
		x102	3,18
2	82	x109	6,24
3	36	-	-
4	25	x212	3,24
5	49	x182	4,03
		x196	6,71
6	5	-	-

Nota: N = número de estações.

Após o teste de discrepância, os grupos foram submetidos ao teste de heterogeneidade, como mostra a Tabela 17.

Tabela 17 - Medidas de heterogeneidade para a solução obtida pela metodologia hierárquica de *ward*

Grupo	N	H1	H2	H3
1	30	1,89	-0,99	-2,17
2	82	**4,1**	**2,25**	0,52
3	36	1,89	-0,05	-0,52
4	25	0,39	0,52	0,34
5	49	**5,39**	1,57	1,56
6	5	1,21	1,97	1,54

Nota: N = número de estações.

Observa-se na Tabela 17 que os grupos 1, 3 e 6 podem ser considerados possivelmente homogêneos enquanto o grupo 4 pode ser considerado aceitavelmente homogêneo pela medida de heterogeneidade H1. Os grupos 1 e 3 podem se considerados aceitavelmente homogêneos pelas medidas de heterogeneidade H2 e H3. Os grupos 4 e 6 podem ser considerados possivelmente homogêneos pelas medidas de heterogeneidade H2 e H3.

O grupo 2 pelas medidas de heterogeneidade H1 e H2 deve ser considerado definitivamente heterogêneo e a medida H3 indica que a região deve ser considerada aceitavelmente homogênea. O grupo 5 pelas medidas H2 e H3 deve ser considerado possivelmente homogêneo e a medida H1 indica o contrário, que a região deve ser considerada definitivamente heterogênea.

Saf (2010) aponta que a principal fonte de erros na regionalização são as estações discordantes, dessa forma, para avaliar uma melhora no desempenho das medidas de heterogeneidade as estações discrepantes foram retiradas e foi realizado novo teste de

discrepância, obtendo-se novas estações discrepantes. O procedimento foi repetido até que não fossem obtidas estações discrepantes, conforme Tabela 18.

Tabela 18 - Estações discrepantes, medida de discrepância e medidas de heterogeneidade para metodologia híbrida de *ward* para seis classes

Grupo	Estação (ID)	D	H1	H2	H3
1	x103	3,45	-0,12	-1,73	-2,50
	-	-	-0,11	-1,88	-2,80
2	x77	3,55	4,08	1,98	0,00
	x49	3,07			
	-	-	4,10	1,53	-0,70
4	x106	3,29	-0,65	-0,63	-0,69
	-	-	-0,75	-1,32	-1,29
5	x195	3,37	1,07	-0,40	0,14
	-	-	0,14	-0,62	0,15

Pode-se observar uma melhora considerável nas medidas de heterogeneidade, exceto o grupo 2, conforme as estações discrepantes são retiradas. Dessa forma, foram retiradas 10 estações de 227, o que proporcionalmente representa, aproximadamente, 5%. Nota-se que algumas medidas de heterogeneidade foram menores que -2, denotando alta correlação cruzada entre as estações pertencentes ao grupo, como apontam Castellarin, Burn e Brath (2008).

No grupo 5 observou-se melhor o efeito de estações discrepantes, pois com a remoção delas o grupo passou de definitivamente heterogêneo para aceitavelmente homogêneo.

Para o grupo 2 não houve melhoras no desempenho da medida de heterogeneidade que pode ser atribuída ao seu tamanho. Para solucionar este problema as estações pertencentes ao grupo foram novamente classificadas pelo método de *ward*, porém, em apenas dois grupos. Após esta divisão as estações foram novamente submetidas ao teste de discrepância e heterogeneidade, conforme Tabela 19.

Tabela 19 - Medidas de heterogeneidade e discordância para as subdivisões do grupo 2

Grupo	N	Estação (ID)	D	H1	H2	H3
2a	39	-	-	0,22	-0,62	-1,29
2b	40	-	-	1,44	0,80	-0,19

Não foram encontradas estações discrepantes nas subdivisões do grupo 2. O grupo 2a pode ser classificado como aceitavelmente homogêneo, pois as três medidas de heterogeneidade são menores que 1. O grupo 2b pode ser classificado como possivelmente homogêneo pela medida H1 e aceitavelmente homogêneo pelas medidas H2 e H3. Dessa forma, a subdivisão do grupo 2 possibilitou a formação de dois grupos aptos para a regionalização hidrológica. Isso corrobora os resultados de Cannarazzo *et al*. (2009) que,

mesmo após eliminar todas as estações discordantes, não obteve homogeneidade, sendo necessária a reclassificação das estações para obter a homogeneidade.

5.2.2 Algoritmo k-médias

A Tabela 20 mostra as medidas de discrepância e heterogeneidade para a solução com seis grupos, obtida pelo algoritmo híbrido de *ward*, até não serem encontradas estações discrepantes.

Tabela 20 - Teste de discrepância e heterogeneidade para a solução com seis grupos, obtida pelo algoritmo híbrido de *ward*

Grupo	N	Estação (ID)	D	H1	H2	H3
1	40	x102	3,07	1,73	-1,57	-2,29
		x103	3,05			
		x38	3,89			
2	37	-	-	-0,09	-2,46	-3,02
	58	x77	3,12	0,50	1,42	1,35
	57	-	-	0,50	1,23	1,07
3	51	x109	6,06	0,32	-0,83	-0,64
	50	x124	3,01	0,30	-1,33	-1,31
	49	x107	3,02	0,20	-1,29	-1,37
	48	x54	3,27	-0,31	-1,47	-1,34
	47	-	-	-0,33	-1,48	-1,31
4	23	x212	3,02	0,64	0,57	0,13
	22	x106	3,01	-0,40	-0,60	-0,88
	21	-	-	-0,53	-1,26	-1,40
5	50	x182	4,16	5,35	1,68	1,79
		x196	7,02			
	48	x195	3,96	0,57	-0,34	0,33
	47	-	-	-0,47	-0,57	0,33
6	5	-	-	1,21	1,97	1,54

Observa-se na Tabela 20 que, para se obter grupos homogêneos foi necessário eliminar 13 estações, representando, aproximadamente, 6% das estações usadas neste estudo. Pelas medidas de heterogeneidade, os grupos 1, 2, 3, 4 e 5 podem ser classificados como aceitavelmente homogêneos. O grupo 6 pode ser classificado, pelas medidas de heterogeneidade, como possivelmente homogêneo.

O algoritmo híbrido de *ward* demonstrou ser mais eficiente na formação de grupos homogêneos que sua forma hierárquica, pois não houve necessidade de subdivisão de nenhum grupo para se obter homogeneidade, apenas remoção das estações discrepantes.

Isso se deve bastante ao tamanho uniforme dos grupos e à complementaridade das metodologias. Primeiramente, aplica-se o método hierárquico de *ward* que, em cada passo do algoritmo, agrupa objetos que têm o menor incremento possível na soma de quadrado dos erros, ou seja, fundamenta-se na mudança de variação entre grupos e dentro de grupos que estão sendo formados em cada passo do agrupamento. Posteriormente, os centroides obtidos do algoritmo hierárquico de *ward* são usados como início do algoritmo k-médias, que

tem por objetivo minimizar a soma dos quadrados das distâncias euclidianas de cada elemento ao centro da classe qual pertence. Assim, fica evidente que, no método híbrido, os objetos passam por uma dupla análise de variância, em que a primeira tem a função de produzir grupos heterogêneos entre si e a segunda homogeneíza o tamanho dos grupos.

A Figura 19 mostra o digrama de dupla massa das estações eliminadas, em função do grupo ao qual pertencem.

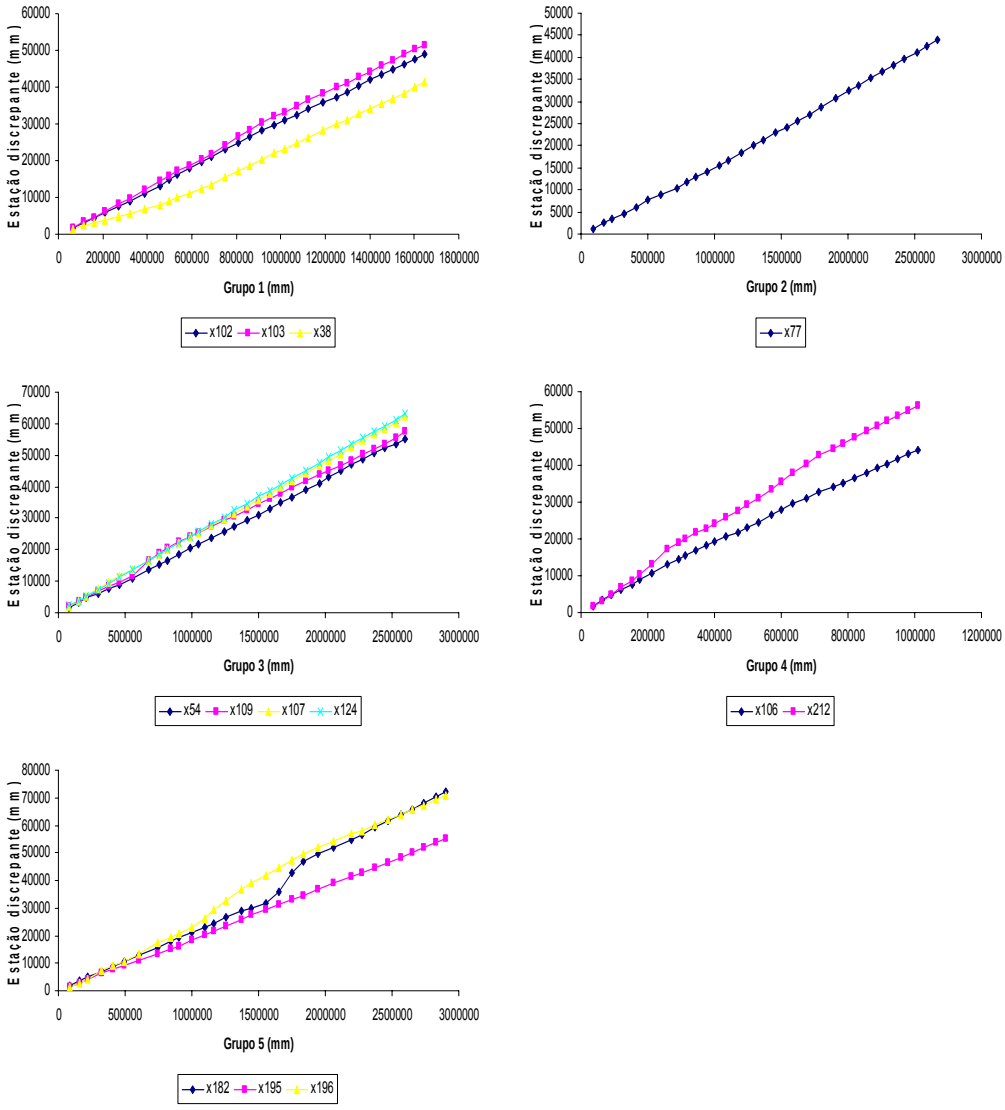

Figura 20 - Diagrama de dupla massa das estações discrepantes, obtidas pela metodologia híbrida de *ward*.

Na Figura 20, observa-se que foram identificadas estações que apresentam consideráveis desvios em relação ao grupo ao qual pertencem. Algumas estações foram identificadas pelo teste de discrepância por apresentarem variação maior que a esperada para o grupo ao qual pertencem.

Para o algoritmo k-médias foram testadas as soluções com sete e nove grupos, pois com cinco grupos não há possibilidade de formação de grupos homogêneos apenas removendo as estações discrepantes, porque um dos grupos possui 90 estações, sendo necessária a subdivisão deste grupo para se obter homogeneidade, como foi demonstrado para a metodologia hierárquica de *ward*.

A Tabela 21 mostra as medidas de discrepância e heterogeneidade para a solução com sete grupos, obtida pelo algoritmo k-médias, até não serem encontradas estações discrepantes.

Tabela 21 - Teste de discrepância e heterogeneidade para a solução com sete grupos, obtida pelo algoritmo k-médias

Grupo	N	Estação (ID)	D	H1	H2	H3
1	47	x38	3,97	2,64	-0,24	-1,35
		x46	3,15			
		x102	3,47			
		x103	3,20			
2	43	-	-	0,74	-1,48	-2,42
	18	x117	3,10	0,34	0,17	0,29
	17	-	-	-1,18	-0,43	-0,06
3	18	x109	4,58	0,17	-0,15	0,39
	17	-	-	0,06	-0,73	-0,49
4	23	x212	3,02	0,64	0,57	0,13
	22	x106	3,01	-0,40	-0,60	-0,88
	21	-	-	-0,53	-1,26	-1,40
5	28	-	-	0,85	-0,08	-0,78
6	46	x182	4,11	4,74	1,21	1,22
		x196	7,03			
7	44	-	-	-0,30	-0,76	-0,24
	5	-	-	1,21	1,97	1,54

Observa-se na Tabela 21 que foram removidas 10 estações, representando menos de 5% do total de 227 estações. Os grupos 1, 2, 3, 4, 5 e 6 são classificados como aceitavelmente homogêneos pelas medidas de heterogeneidade. O grupo 7 é classificado como possivelmente homogêneo pelas medidas de heterogeneidade.

A Tabela 22 mostra as medidas de discrepância e heterogeneidade para a solução com nove grupos obtida pelo algoritmo k-médias até não serem encontradas estações discrepantes. Observa-se que para obter homogeneidade dos grupos (Tabela 13) foi necessária a remoção de 8 estações, isto representa, aproximadamente, 3,5% do total de 227 estações. Pelas medidas de heterogeneidade são aceitavelmente homogêneos todos os grupos, quando retiradas as estações discrepantes, exceto o grupo 8 que é considerado possivelmente homogêneo. A metodologia k-médias foi eficaz na identificação do número de grupos que são homogêneos apenas pela eliminação das estações discrepantes, sem necessidade de subdivisão dos grupos para se obter homogeneidade.

Tabela 22 - Teste de discrepância e heterogeneidade para a solução com nove grupos, obtida pelo algoritmo k-médias

Grupo	N	Estação (ID)	D	H1	H2	H3
1	38	x38	3,76	1,83	-1,22	-1,98
	37	-	-	0,89	-1,69	-2,21
2	31	x76	3,08	-0,33	-0,54	-0,46
	30	-	-	-0,56	-0,76	-0,58
3	29	-	-	-0,43	1,37	1,51
4	21	-	-	-0,10	-0,49	-1,03
5	23	x212	3,02	0,64	0,57	0,13
	22	x106	3,01	-0,4	-0,60	-0,88
	21	-	-	-0,53	-1,26	-1,40
6	28	x182	3,52	4,32	1,47	1,12
		x196	4,76			
	26	x195	3,24	-0,76	-1,82	-2,19
	25	-	-	-1,95	-2,26	-2,46
7	26	x109	4,27	-0,67	-0,34	0,48
	25	-	-	-0,99	-0,93	-0,41
8	5	-	-	1,21	1,97	1,54
9	26	-	-	-0,20	-1,54	-1,95

A Tabela 23 mostra o algoritmo de classificação utilizado e quais estações foram eliminadas em função do número de grupos.

Tabela 23 - Estações discrepantes em função da metodologia de classificação e do número de grupos

K-médias		Híbrido	*Ward*
k=9	k=7	k=6	k=6
x38	x38	x38	x38
	x102	x102	x102
	x103	x103	x103
x106	x106	x106	x106
x109	x109	x109	x109
x196	x196	x196	x196
x212	x212	x212	x212
x182	x182	x182	X182
x195		x195	x195
		x77	x77
	x46	x54	x49
	x117	x107	
x76		x124	
8	10	13	11

Nota: k=número de grupos.

Observa-se na Tabela 23 que as estações x38, x106, x109, x182 x196 e x212 aparecem como discrepantes em todas as situações testadas. Isso concorre para que sejam descartadas para análises posteriores, pois sua eliminação independe do algoritmo de classificação e do número de classes adotado.

As estações x102 e x103 são consideradas discrepantes pelas metodologias de *ward*, híbrida e k-médias, com sete grupos. A estação x195 foi considerada discrepante pelas metodologias *ward*, híbrida e k-médias, com nove grupos. Estas estações foram identificadas no teste de discrepância porque apresentam pequenos desvios em relação ao grupo ao qual pertencem, dependendo da sensibilidade das metodologias de classificação.

As demais estações eliminadas dependem do algoritmo utilizado e do número de grupos adotados, pois, conforme o tamanho do grupo haverá variações nos momentos-L que fazem que determinadas estações tornem-se discrepantes.

A metodologia híbrida foi mais sensível na identificação de estações discrepantes seguida pela metodologia hierárquica de *ward*. Nota-se que a metodologia híbrida corroborou as demais, na maioria dos casos, nas estações eliminadas.

Fica claro também que o aumento no número de classes da metodologia k-médias diminui sua sensibilidade na identificação de estações discrepantes. Para sete grupos, as estações x102 e x103 são identificadas, porém para nove classes elas não são. Para nove classes identifica-se a estação x195 que não aparece para sete classes, porém há uma redução do número de estações discrepantes.

Desta forma a solução mais adequada é a solução da forma híbrida entre a metodologia hierárquica de *ward* com k-médias, pois apresenta grupos de tamanho uniformes e satisfatórios para regionalização hidrológica. Além disso, produz grupos homogêneos e possui maior sensibilidade para identificar estações discrepantes. Corroborando Rao e Srinivas (2006b) que apontam o algoritmo híbrido *ward* como melhor solução. Cheng e Liao (2009) também obtiveram bons resultados com o algoritmo híbrido de *ward* no zoneamento regional de sistemas de armazenamento da água da chuva.

5.3 Estudo de Caso

Nas tabelas 24 e 25 é confrontada a distribuição teórica regional com a distribuição empírica local, pelo teste de Kolmogorov-Smirnov, para os grupos obtidos pela metodologia híbrida de *ward* usando a distribuição gama e Pearson tipo III.

Na Tabela 24 são verificadas somente cinco situações em que a séries não passou pelo teste de aderência de Kolmogorov-Smirnov, ao nível de 10% de significância, usando a distribuição gama. Isso mostra a versatilidade da distribuição gama em representar os dados locais por meio de parâmetros regionais, portanto, ela pode ser usada na estimação de quantis regionais adimensionais.

Tabela 24 - Teste de Kolmogorov-Smirnov regional da distribuição gama

	Grupo					
Mês	1	2	3	4	5	6
jan	0,1591	0,1787	0,2013	0,2019	0,1869	0,1597
fev	0,1905	0,2014	0,1985	0,2132	**0,2249**	0,1293
mar	0,2064	0,2048	0,1842	0,1533	**0,2405**	0,1372
abr	0,2137	**0,2456**	0,1743	**0,2271**	**0,226**	0,1306
mai	0,1658	0,1689	0,141	0,138	0,1807	0,1255
jun	0,1793	0,1967	0,1611	0,1563	0,1965	0,1344
jul	0,2085	0,2002	0,159	0,1362	0,1841	0,1623
ago	0,1932	0,1887	0,2058	0,1832	0,1962	0,1848
set	0,1696	0,1701	0,2058	0,1643	0,1684	0,1632
out	0,1675	0,1902	0,1954	0,155	0,1864	0,1739
nov	0,1612	0,1721	0,1784	0,2028	0,1917	0,1597
dez	0,1797	0,1782	0,1734	0,1767	0,1902	0,182

Nota: $KS_{(10\%,31)} = 0,214$.

Observa-se na Tabela 25 que a distribuição Pearson tipo III se ajustou bem aos dados, pois apenas em três situações não houve aderência. Como a distribuição Pearson tipo III é a gama com três parâmetros, observa-se uma suavização nos valores do teste de aderência.

Tabela 25 - Teste de Kolmogorov-Smirnov regional da distribuição Pearson tipo III

	Grupo					
Mês	1	2	3	4	5	6
jan	0,1570	0,1751	0,1788	0,1875	0,1518	0,1595
fev	0,1499	0,1680	0,2033	0,1798	0,2078	0,1221
mar	0,2062	0,1985	0,1820	0,1552	0,2444	0,1386
abr	0,1617	0,2238	0,1586	0,1967	0,2207	0,1161
mai	0,1827	0,1628	0,1649	0,1382	0,1762	0,1277
jun	0,1560	0,1706	0,1456	0,1472	0,1694	0,1346
jul	0,2067	0,1755	0,1583	0,1398	0,1954	0,1408
ago	0,1434	0,1663	0,1568	0,1360	0,1628	0,1807
set	0,1489	0,1792	0,1749	0,1456	0,1684	0,1775
out	0,1639	0,1825	0,1836	0,1489	0,1786	0,1805
nov	0,1629	0,1815	0,1673	0,1920	0,1942	0,1602
dez	0,1729	0,1754	0,1731	0,1850	0,1893	0,1434

Nota: $KS_{(10\%,31)} = 0,214$.

Pode-se concluir, mesmo para um nível de 10% de significância, um nível restritivo, que as duas distribuições de probabilidade modelam a distribuição de frequência de precipitação total mensal no estado do Paraná.

Foram estimados os quantis locais (valor observado) e confrontados com a estimativa regional (valor estimado), conforme equações 58, 59 e 60. Dessa forma, pode-se avaliar qual das distribuições de probabilidade é a mais adequada na modelagem da precipitação mensal.

As tabelas 26 a 31 mostram os erros cometidos mensalmente pela regionalização, aplicando-se a distribuição gama e de Pearson tipo III, em função do período de retorno.

Tabela 26 - Erro médio percentual das distribuições de probabilidade, em função do período de retorno para o grupo 1

	Gama						Pearson tipo III					
	Período de retorno						Período de retorno					
Mês	5	10	20	50	75	100	5	10	20	50	75	100
jan	1,89	3,50	4,66	5,80	6,22	6,48	2,35	3,54	5,27	7,38	8,23	8,79
fev	1,93	3,53	4,67	5,80	6,21	6,47	2,09	3,91	5,97	8,41	9,38	10,04
mar	2,85	5,23	6,96	8,68	9,30	9,70	2,75	5,67	8,43	11,46	12,62	13,38
abr	2,07	3,88	5,16	6,40	6,85	7,13	2,51	3,98	5,74	7,90	8,78	9,38
mai	1,92	4,20	5,75	7,21	7,72	8,04	1,90	4,61	6,88	9,23	10,10	10,67
jun	1,17	3,34	4,77	6,07	6,52	6,80	1,92	3,47	5,42	7,46	8,20	8,68
jul	1,85	4,69	6,60	8,38	9,00	9,39	2,17	5,07	8,43	11,87	13,13	13,92
ago	0,88	4,95	7,61	9,92	10,68	11,15	1,25	5,61	9,44	12,96	14,17	14,92
set	1,70	3,24	4,33	5,40	5,78	6,03	1,70	3,73	5,78	8,14	9,09	9,71
out	2,98	5,45	7,24	9,03	9,69	10,11	3,20	5,56	7,73	10,15	11,10	11,71
nov	1,97	3,56	4,70	5,82	6,23	6,50	2,20	3,64	5,28	7,20	7,96	8,46
dez	2,11	3,67	4,81	5,95	6,37	6,64	1,95	4,28	6,68	9,42	10,49	11,20

Tabela 27 - Erro médio percentual das distribuições de probabilidade, em função do período de retorno para o grupo 2

	Gama						Pearson tipo III					
	Período de retorno						Período de retorno					
Mês	5	10	20	50	75	100	5	10	20	50	75	100
jan	2,46	4,44	5,87	7,28	7,79	8,12	2,43	4,95	7,42	10,19	11,27	11,98
fev	2,44	4,48	5,93	7,37	7,89	8,22	2,36	5,15	7,89	11,00	12,23	13,05
mar	3,06	5,65	7,56	9,48	10,19	10,64	3,20	5,94	8,53	11,45	12,59	13,34
abr	2,53	4,96	6,67	8,33	8,93	9,31	2,38	5,62	8,70	12,05	13,33	14,16
mai	1,28	3,09	4,31	5,43	5,82	6,07	1,57	3,49	5,69	7,93	8,75	9,27
jun	1,49	3,72	5,19	6,53	7,00	7,29	2,06	3,76	5,41	7,14	7,78	8,18
jul	1,93	4,24	5,82	7,30	7,81	8,14	1,97	4,93	7,78	10,76	11,87	12,58
ago	1,04	4,89	7,38	9,55	10,28	10,72	1,73	5,19	8,07	10,74	11,66	12,23
set	2,06	3,96	5,32	6,63	7,11	7,41	2,03	4,30	6,35	8,57	9,43	9,99
out	2,03	3,72	4,93	6,12	6,56	6,83	2,28	3,82	5,52	7,50	8,29	8,81
nov	2,15	3,87	5,11	6,32	6,76	7,04	2,41	4,04	5,90	8,09	8,95	9,52
dez	2,88	5,00	6,54	8,10	8,68	9,05	3,04	5,25	7,58	10,31	11,39	12,11

Tabela 28 - Erro médio percentual das distribuições de probabilidade, em função do período de retorno para o grupo 3

	Gama						Pearson tipo III					
	Período de retorno						Período de retorno					
Mês	5	10	20	50	75	100	5	10	20	50	75	100
jan	2,44	4,43	5,87	7,30	7,82	8,15	2,35	4,99	7,43	10,19	11,28	12,00
fev	2,57	4,78	6,35	7,89	8,45	8,80	2,70	5,16	7,61	10,40	11,51	12,24
mar	2,92	5,39	7,17	8,94	9,58	9,99	3,06	5,59	8,03	10,75	11,80	12,49
abr	2,17	4,29	5,78	7,22	7,73	8,06	2,32	4,48	6,46	8,59	9,40	9,93
mai	1,52	3,33	4,56	5,71	6,12	6,37	1,47	3,91	6,37	8,96	9,92	10,55
jun	2,29	4,53	6,07	7,55	8,07	8,40	2,14	5,07	7,63	10,34	11,36	12,02
jul	1,70	3,55	4,83	6,04	6,47	6,74	1,71	4,02	6,24	8,61	9,50	10,07
ago	1,52	3,83	5,37	6,78	7,27	7,58	1,85	4,32	6,69	9,19	10,12	10,72
set	1,75	3,29	4,39	5,46	5,85	6,09	1,63	4,15	7,03	10,30	11,59	12,43
out	2,13	3,71	4,86	6,00	6,42	6,68	2,24	3,98	5,94	8,24	9,17	9,78
nov	1,99	3,54	4,65	5,75	6,15	6,41	1,92	4,00	6,09	8,44	9,35	9,95
dez	2,73	4,88	6,44	8,00	8,57	8,93	2,67	5,36	8,03	11,02	12,17	12,93

Tabela 29 - Erro médio percentual das distribuições de probabilidade, em função do período de retorno para o grupo 4

	Gama						Pearson tipo III					
	Período de retorno						Período de retorno					
Mês	5	10	20	50	75	100	5	10	20	50	75	100
jan	2,10	3,73	4,93	6,13	6,57	6,85	2,69	3,78	5,81	8,48	9,59	10,33
fev	2,30	3,94	5,14	6,32	6,76	7,04	2,41	4,21	5,98	8,05	8,87	9,42
mar	2,49	4,54	6,00	7,44	7,97	8,30	1,87	5,50	8,99	12,77	14,21	15,15
abr	2,80	5,26	6,99	8,66	9,26	9,64	2,68	5,98	8,89	12,03	13,24	14,02
mai	1,12	2,87	4,02	5,09	5,46	5,68	1,38	2,97	4,54	6,12	6,69	7,04
jun	2,27	4,45	5,98	7,46	7,98	8,31	1,75	5,26	8,16	11,20	12,35	13,09
jul	1,88	3,55	4,71	5,82	6,22	6,47	1,68	4,25	6,62	9,18	10,17	10,81
ago	1,37	3,05	4,19	5,24	5,61	5,84	1,68	3,51	5,62	7,95	8,85	9,43
set	2,04	3,87	5,18	6,46	6,92	7,21	2,02	4,23	6,23	8,44	9,30	9,86
out	1,69	2,97	3,91	4,86	5,21	5,44	1,94	3,04	4,40	6,09	6,79	7,25
nov	1,92	3,47	4,57	5,66	6,05	6,30	1,86	3,97	6,26	8,86	9,87	10,53
dez	2,51	4,34	5,68	7,03	7,53	7,85	2,44	4,80	7,33	10,36	11,59	12,41

Tabela 30 - Erro médio percentual das distribuições de probabilidade, em função do período de retorno para o grupo 5

	Gama						Pearson tipo III					
	Período de retorno						Período de retorno					
Mês	5	10	20	50	75	100	5	10	20	50	75	100
jan	2,39	4,23	5,56	6,88	7,37	7,68	2,44	4,57	6,57	8,86	9,77	10,37
fev	2,50	4,50	5,92	7,31	7,81	8,13	2,99	4,34	5,84	7,69	8,45	8,96
mar	3,17	5,73	7,57	9,38	10,03	10,45	3,35	5,83	8,08	10,54	11,48	12,08
abr	1,92	3,64	4,86	6,04	6,46	6,73	2,09	3,78	5,39	7,16	7,83	8,27
mai	1,23	2,91	4,05	5,12	5,50	5,73	1,22	3,46	5,79	8,23	9,14	9,73
jun	2,08	3,81	5,04	6,24	6,67	6,95	2,10	4,37	6,79	9,59	10,70	11,44
jul	1,22	2,42	3,26	4,05	4,33	4,51	2,28	2,69	4,83	7,29	8,22	8,82
ago	1,74	3,85	5,29	6,65	7,12	7,42	1,93	4,08	5,97	7,94	8,69	9,17
set	2,35	4,20	5,52	6,82	7,29	7,59	2,34	4,64	6,94	9,56	10,59	11,27
out	1,94	3,34	4,38	5,42	5,81	6,06	2,23	3,46	5,27	7,53	8,44	9,05
nov	2,32	4,21	5,57	6,92	7,41	7,72	2,10	4,74	7,45	10,45	11,61	12,36
dez	2,90	5,07	6,65	8,24	8,82	9,20	2,95	5,37	7,76	10,50	11,59	12,30

Tabela 31 - Erro médio percentual das distribuições de probabilidade, em função do período de retorno para o grupo 6

	Gama						Pearson tipo III					
	Período de retorno						Período de retorno					
Mês	5	10	20	50	75	100	5	10	20	50	75	100
jan	2,07	3,50	4,56	5,62	6,02	6,27	2,18	3,64	5,22	7,08	7,83	8,33
fev	1,87	3,15	4,10	5,08	5,44	5,67	1,92	3,82	6,29	9,34	10,60	11,45
mar	2,55	4,27	5,57	6,90	7,41	7,73	3,14	3,63	4,06	4,67	4,95	5,16
abr	1,70	3,04	4,01	4,96	5,31	5,53	1,69	3,31	4,87	6,65	7,36	7,84
mai	1,20	2,50	3,39	4,23	4,52	4,71	1,86	2,13	2,40	2,74	2,88	2,98
jun	1,98	3,73	4,97	6,17	6,61	6,88	2,15	4,14	6,74	9,68	10,81	11,55
jul	2,84	5,66	7,64	9,55	10,24	10,67	1,61	7,43	12,88	18,74	20,97	22,42
ago	2,90	5,67	7,62	9,50	10,17	10,59	2,81	6,05	8,69	11,42	12,42	13,07
set	3,34	5,91	7,78	9,64	10,33	10,76	3,29	6,01	8,17	10,41	11,25	11,78
out	3,96	6,62	8,58	10,57	11,31	11,78	3,28	7,51	11,71	16,51	18,40	19,65
nov	3,88	7,22	9,69	12,18	13,09	13,68	3,10	8,39	13,64	19,39	21,59	23,02
dez	2,25	4,01	5,31	6,62	7,10	7,41	1,97	4,61	7,62	10,92	12,16	12,97

Nas tabelas 26 a 31 observa-se que há monotocidade nos erros com aumento do período de retorno, como também concluem Ouarda *et al*. (2008), estudando regionalização

de vazões. Dessa forma, os maiores erros ocorrem quando se aproxima dos caudais das distribuições. Nota-se que os erros são menores para a distribuição gama, pois seus erros estão próximos a 10%, enquanto na distribuição Pearson tipo III os erros podem chegar a 23%.

As figuras 21 a 26 mostram os erros percentuais médios para cada período de retorno estudado, em função da época ano.

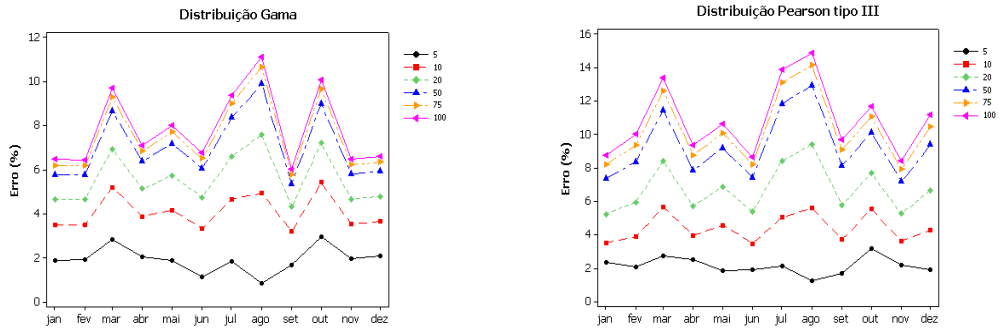

Figura 21 - Erro médio percentual das distribuições de probabilidade na estimativa mensal dos quantis regionais para o grupo 1.

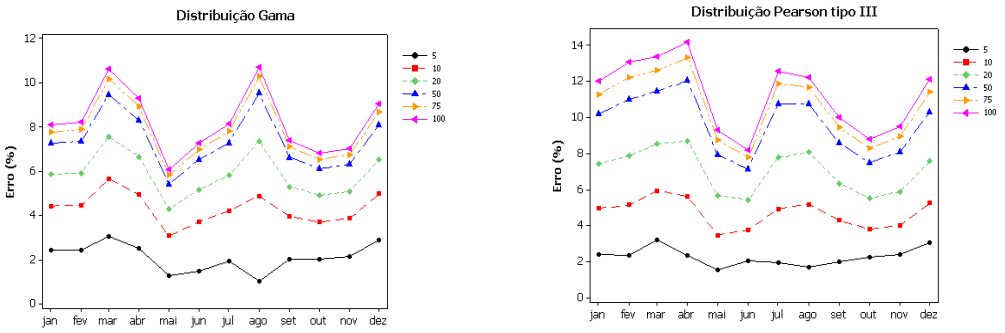

Figura 22 - Erro médio percentual das distribuições de probabilidade na estimativa mensal dos quantis regionais para o grupo 2.

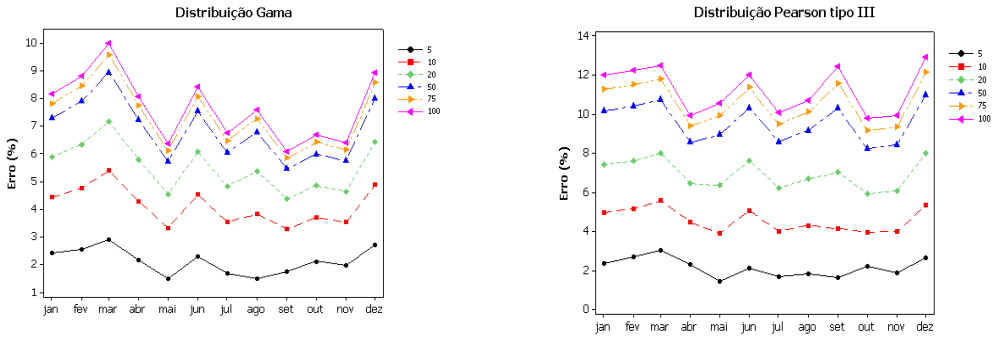

Figura 23 - Erro médio percentual das distribuições de probabilidade na estimativa mensal dos quantis regionais para o grupo 3.

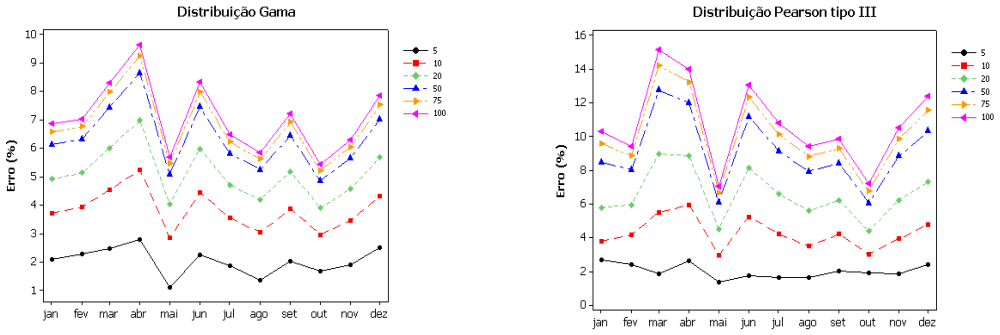

Figura 24 - Erro médio percentual das distribuições de probabilidade na estimativa mensal dos quantis regionais para o grupo 4.

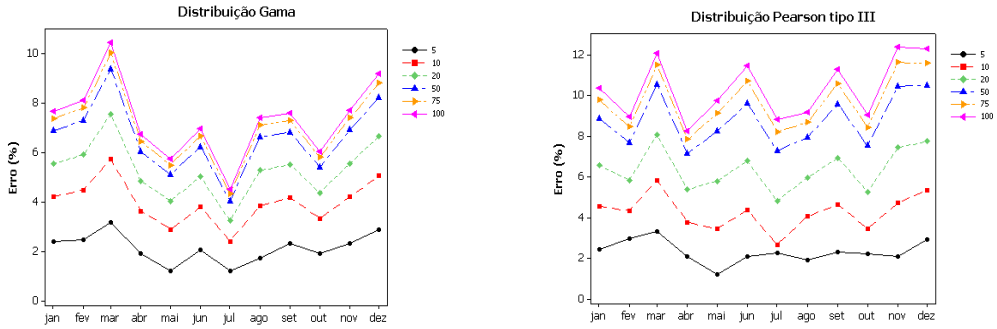

Figura 25 - Erro médio percentual das distribuições de probabilidade na estimativa mensal dos quantis regionais para o grupo 5.

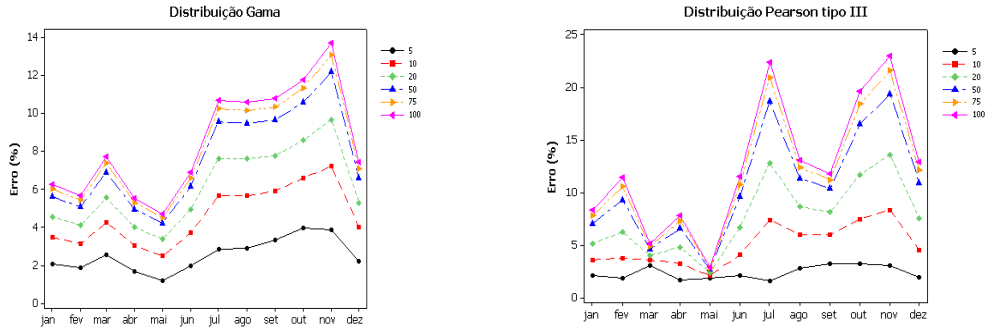

Figura 26 - Erro médio percentual das distribuições de probabilidade na estimativa mensal dos quantis regionais para o grupo 6.

Observa-se que para os períodos de retorno de 50, 75 e 100 anos não há um incremento significativo nos erros, sendo verificados pela proximidade das curvas, já para os períodos de recorrência 5, 10 e 20 os incrementos nos erros são significativos.

Observam-se sazonalidades nos erros cometidos na estimativa dos quantis regionais, porém não foi encontrado um padrão. Como a distribuição gama apresenta as estimativas regionais mais próximas das estimativas locais esta será usada na estimação de quantis locais.

As tabelas 32 a 37 mostram os parâmetros, médias e quantis regionais.

Tabela 32 - Parâmetros da distribuição gama regional, média regional e quantis adimensionais para o grupo 1

	Parâmetros			Período de Retorno					
Mês	alfa	beta	Média	5	10	20	50	75	100
jan	2,9320	0,3411	202,9	1,4303	1,7830	2,1125	2,5261	2,7046	2,8276
fev	3,0526	0,3276	160,3	1,4234	1,7674	2,0882	2,4901	2,6634	2,7828
mar	3,2343	0,3092	128,5	1,4136	1,7456	2,0542	2,4401	2,6061	2,7205
abr	2,5679	0,3894	93,2	1,4532	1,8361	2,1964	2,6511	2,8479	2,9839
mai	1,5665	0,6384	111,0	1,5398	2,0619	2,5675	3,2203	3,5069	3,7060
jun	1,0802	0,9258	80,6	1,5989	2,2592	2,9155	3,7793	4,1630	4,4309
jul	1,2793	0,7817	56,8	1,5734	2,1667	2,7492	3,5090	3,8447	4,0785
ago	0,6922	1,4448	49,6	1,6444	2,5168	3,4176	4,6351	5,1841	5,5699
set	2,5844	0,3869	114,1	1,4521	1,8335	2,1922	2,6448	2,8407	2,9761
out	3,4025	0,2939	124,6	1,4051	1,7269	2,0253	2,3977	2,5577	2,6679
nov	3,2215	0,3104	132,0	1,4143	1,7471	2,0565	2,4435	2,6100	2,7247
dez	4,1978	0,2382	180,9	1,3711	1,6541	1,9140	2,2357	2,3732	2,4677

Tabela 33 - Parâmetros da distribuição gama regional, média regional e quantis adimensionais para o grupo 2

	Parâmetros			Período de retorno					
Mês	alfa	beta	Média	5	10	20	50	75	100
jan	3,2108	0,3114	184,2	1,4148	1,7483	2,0584	2,4463	2,6132	2,7283
fev	3,0009	0,3332	154,1	1,4263	1,7740	2,0984	2,5053	2,6807	2,8016
mar	3,4632	0,2888	124,1	1,4022	1,7205	2,0155	2,3832	2,5412	2,6500
abr	2,3327	0,4287	110,8	1,4700	1,8766	2,2611	2,7484	2,9599	3,1062
mai	1,3125	0,7619	131,5	1,5693	2,1531	2,7253	3,4706	3,7995	4,0287
jun	1,2071	0,8284	93,3	1,5825	2,1980	2,8048	3,5987	3,9501	4,1951
jul	1,5661	0,6385	58,4	1,5399	2,0620	2,5677	3,2206	3,5073	3,7065
ago	0,7360	1,3586	59,0	1,6407	2,4806	3,3427	4,5034	5,0256	5,3922
set	2,4326	0,4111	131,1	1,4626	1,8587	2,2324	2,7052	2,9101	3,0518
out	3,0072	0,3325	147,4	1,4259	1,7732	2,0972	2,5034	2,6786	2,7993
nov	3,0643	0,3263	136,0	1,4227	1,7660	2,0859	2,4868	2,6595	2,7786
dez	4,4285	0,2258	180,2	1,3627	1,6367	1,8876	2,1977	2,3301	2,4210

Tabela 34 - Parâmetros da distribuição gama regional, média regional e quantis adimensionais para o grupo 3

	Parâmetros			Período de retorno					
Mês	alfa	beta	Média	5	10	20	50	75	100
jan	3,2228	0,3103	186,8	1,4142	1,7469	2,0563	2,4431	2,6096	2,7243
fev	2,7834	0,3593	160,2	1,4392	1,8035	2,1447	2,5739	2,7593	2,8873
mar	3,0849	0,3242	129,3	1,4216	1,7634	2,0819	2,4809	2,6528	2,7713
abr	2,1820	0,4583	135,7	1,4818	1,9057	2,3081	2,8196	3,0420	3,1960
mai	1,5661	0,6385	174,2	1,5399	2,0620	2,5677	3,2206	3,5073	3,7065
jun	1,9917	0,5021	121,6	1,4979	1,9468	2,3751	2,9218	3,1601	3,3253
jul	1,8054	0,5539	94,9	1,5152	1,9926	2,4509	3,0385	3,2954	3,4736
ago	1,2236	0,8173	82,7	1,5804	2,1907	2,7917	3,5775	3,9251	4,1674
set	2,5833	0,3871	156,1	1,4521	1,8337	2,1925	2,6453	2,8412	2,9766
out	3,8548	0,2594	192,1	1,3847	1,6828	1,9576	2,2988	2,4451	2,5456
nov	3,4066	0,2935	166,8	1,4049	1,7265	2,0247	2,3967	2,5566	2,6667
dez	3,6018	0,2776	184,5	1,3957	1,7065	1,9939	2,3517	2,5052	2,6109

Tabela 35 - Parâmetros da distribuição gama regional, média regional e quantis adimensionais para o grupo 4

	Parâmetros			Período de retorno					
Mês	alfa	beta	Média	5	10	20	50	75	100
jan	4,0062	0,2496	191,4	1,3785	1,6697	1,9376	2,2699	2,4121	2,5099
fev	4,1168	0,2429	154,6	1,3742	1,6605	1,9238	2,2498	2,3893	2,4851
mar	3,0543	0,3274	126,9	1,4233	1,7672	2,0879	2,4896	2,6628	2,7822
abr	2,4526	0,4077	88,0	1,4612	1,8553	2,2269	2,6969	2,9006	3,0414
mai	1,2053	0,8297	125,3	1,5827	2,1988	2,8063	3,6011	3,9528	4,1981
jun	2,1740	0,4600	102,3	1,4824	1,9073	2,3107	2,8236	3,0466	3,2010
jul	2,1992	0,4547	102,2	1,4804	1,9023	2,3025	2,8110	3,0322	3,1852
ago	1,4431	0,6930	73,8	1,5537	2,1036	2,6390	3,3330	3,6384	3,8509
set	2,6160	0,3823	143,2	1,4499	1,8285	2,1843	2,6330	2,8271	2,9612
out	4,4397	0,2252	153,6	1,3623	1,6358	1,8864	2,1959	2,3281	2,4188
nov	3,0614	0,3267	125,6	1,4229	1,7663	2,0865	2,4876	2,6605	2,7796
dez	4,5668	0,2190	154,7	1,3580	1,6268	1,8728	2,1764	2,3060	2,3949

Tabela 36 - Parâmetros da distribuição gama regional, média regional e quantis adimensionais para o grupo 5

	Parâmetros			Período de retorno					
Mês	alfa	beta	Média	5	10	20	50	75	100
jan	3,6590	0,2733	190,2	1,3932	1,7009	1,9853	2,3392	2,4910	2,5955
fev	2,9469	0,3393	179,0	1,4294	1,7810	2,1094	2,5215	2,6993	2,8219
mar	3,1086	0,3217	137,0	1,4203	1,7605	2,0774	2,4742	2,6451	2,7629
abr	2,3998	0,4167	149,8	1,4650	1,8645	2,2416	2,7191	2,9261	3,0693
mai	1,4315	0,6986	183,7	1,5551	2,1077	2,6461	3,3444	3,6517	3,8656
jun	2,8296	0,3534	153,4	1,4364	1,7969	2,1344	2,5586	2,7418	2,8682
jul	1,9887	0,5028	132,1	1,4982	1,9475	2,3762	2,9235	3,1621	3,3275
ago	1,5701	0,6369	106,8	1,5394	2,0608	2,5656	3,2172	3,5033	3,7021
set	3,1850	0,3140	168,7	1,4162	1,7514	2,0631	2,4532	2,6211	2,7369
out	4,6553	0,2148	231,9	1,3550	1,6208	1,8637	2,1634	2,2911	2,3789
nov	3,1833	0,3141	177,5	1,4163	1,7516	2,0635	2,4537	2,6217	2,7374
dez	4,1023	0,2438	180,3	1,3747	1,6617	1,9255	2,2524	2,3922	2,4883

Tabela 37 - Parâmetros da distribuição gama regional, média regional e quantis adimensionais para o grupo 6

	Parâmetros			Período de retorno					
Mês	alfa	beta	Média	5	10	20	50	75	100
jan	4,9388	0,2025	366,3	1,3460	1,6025	1,8363	2,1241	2,2467	2,3307
fev	5,8112	0,1721	323,2	1,3222	1,5547	1,7654	2,0232	2,1326	2,2075
mar	6,9208	0,1445	280,7	1,2980	1,5075	1,6961	1,9255	2,0225	2,0888
abr	3,3450	0,2990	169,4	1,4080	1,7332	2,0350	2,4118	2,5738	2,6854
mai	1,7284	0,5786	142,6	1,5228	2,0135	2,4858	3,0926	3,3583	3,5427
jun	2,5481	0,3925	109,7	1,4545	1,8394	2,2015	2,6587	2,8567	2,9935
jul	2,1459	0,4660	128,7	1,4847	1,9131	2,3201	2,8378	3,0630	3,2190
ago	2,2058	0,4533	91,4	1,4799	1,9009	2,3003	2,8078	3,0284	3,1810
set	3,7368	0,2676	184,5	1,3897	1,6936	1,9740	2,3228	2,4723	2,5751
out	5,8150	0,1720	195,3	1,3221	1,5545	1,7651	2,0228	2,1321	2,2070
nov	3,7320	0,2680	194,3	1,3899	1,6940	1,9747	2,3238	2,4734	2,5764
dez	4,0282	0,2483	255,1	1,3776	1,6678	1,9348	2,2658	2,4075	2,5049

6 CONCLUSÕES

Os resultados obtidos neste trabalho permitem concluir que:

- a metodologia híbrida entre os algoritmos k-médias e *ward* foi capaz de agrupar estações em regiões homogêneas, conforme demonstrado pelas medidas de discordância e heterogeneidade;
- os momentos-L são eficazes em identificar estações que apresentam anormalidades nos seus registros, sendo identificadas pela medida de discordância;
- a medida de heterogeneidade é uma medida eficaz para validar a existência de homogeneidade nos registros das estações de uma região;
- Para se obter uma região homogênea deve-se retirar as estações discrepantes e se persistir a heterogeneidade deve-se dividir o grupo, na região de estudo;
- A distribuição gama pode ser usada na regionalização hidrológica produzindo erros de até 10%, no caso deste trabalho.

REFERÊNCIAS

AMIN-NASERI, M. R.; SOROUSH, A. R.; Combined use of unsupervised and supervised learning for daily peak load forecasting. **Energy Conversion and Management**, Kidlington, v. 49, n. 6, p. 1302-1308, 2008.

BARBOSA, J. P. S.; VALERIANO, M. M.; SCOFIELD, G. B. Cálculo do excedente hídrico no alto curso do Rio Paraíba do Sul, SP. SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12., Goiânia, 2005. **Resumos...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2005. p. 2463-2470.

BARRETO, S.; FERREIRA, C.; PAIXÃO, J.; SANTOS, B. S. Using clustering analysis in a capacitated location-routing problem. **European Journal of Operational Research**, Amsterdam, v. 179, n. 3, p. 968-977, 2007.

BEAVER, S.; PALAZOĞLU, A. A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. **Atmospheric Environment**, Amsterdam, v. 40, n. 4, p. 713-725, 2006.

BOCCHIOLA, D.; MEDAGLIANI, M.; ROSSO, R. Regional snow depth frequency curves for avalanche hazard mapping in central Italian Alps. **Cold regions science and technology**, Lebanon, v. 46, n. 3, p. 204-221, 2006.

BURN, D. H. Catchment similarity for regional flood frequency analysis using seasonality measures. **Journal of Hydrology**, Amsterdam, v. 202, n. 1-4, p. 212-230, 1997.

CANNAROZZO, M.; NOTO, L. V.; VIOLA, F.; LA LOGGIA, G. Annual runoff regional frequency analysis in Sicily. **Physics and Chemistry of the Earth**, Amsterdam, v. 34, n. 10-12, p. 679-687, 2009.

CASTELLARIN, A.; BURN, D.H.; BRATH, A. Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem?. **Journal of Hydrology**, Amsterdam, v. 360, n. 1-4, p. 67-76, 2008.

CHENG, C. L.; LIAO, M. C. Regional rainfall level zoning for rainwater harvesting systems in northern Taiwan. **Resources, Conservation and Recycling**, Amsterdam, v. 53, n. 8, p. 421-428, 2009.

CORAL, G.; PINTO, H. S.; ASSAD, E. D.; LAFFE, A. Utilização de um modelo agrometeorológico na estimativa de produtividade da cultura da soja no estado de São Paulo. CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA, 14., Campinas, 2005. **Resumos...** Santa Maria: Sociedade Brasileira de Agrometeorologia, 2005.

CORTÉS, J. A.; PALMA, J. L.; WILSON, M. Deciphering magma mixing: The application of cluster analysis to the mineral chemistry of crystal populations. **Journal of Volcanology and Geothermal Research**, Amsterdam, v. 165, n. 3-4, p. 163-188, 2007.

COSTA, J. C. G.; HERNANDEZ, F. B. T.; VANZELA, L. S.; BISPO, E. M. Agroclimatologia da região da Nova Alta Paulista, Irapuru e Junqueirópolis, SP. In: CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNESP, 18., 2006, Jaboticabal. **Resumos...** Jaboticabal: Universidade Estadual Paulista, 2006.

FERREIRA, D. F. **Estatística multivariada**. 1. ed., Lavras: UFLA, 2008. 662 p.

FIGUEIREDO, L. C. C.; RUBERT, O. A. V. A integração entre o sistema nacional de informações sobre saneamento e o sistema nacional de informações sobre recursos hídricos. CONGRESSO BRASILEIRO DE ENGENHARIA SANITÁRIA E AMBIENTAL, 21., João Pessoa, 2001. **Resumos...** Rio de Janeiro: Associação Brasileira de Engenharia Sanitária e Ambiental, 2001.

FISTAROL, O.; FRANK, B.; REFOSCO, J. C. Sistema de Informações de Recursos Hídricos da Bacia do Itajaí. CONGRESSO BRASILEIRO DE CADASTRO TÉCNICO MULTIFINALITÁRIO, 6., Florianópolis, 2004. **Resumos...** Rio de Janeiro: Sociedade Brasiliera de Cartografia, 2004.

FOWLER, H. J.; KILSBY, C. G. A regional frequency analysis of united kingdom extreme rainfall from 1961 to 2000. **International journal of climatology**, Malden, n. 23, n. 11, p. 1313-1334, 2003.

GARCÍA, H. L.; GONZÁLEZ, I. M.; Self-organizing map and clustering for wastewater treatment monitoring. **Engineering Applications of Artificial Intelligence**, Amsterdam, v. 17, n. 3, p. 215-225, 2004.

GELLENS, D. Combining regional approach and data extension procedure for assessing GEV distribution of extreme precipitation in Belgium. **Journal of Hydrology**, Amsterdam, v. 268, n. 1-4, p. 113-126, 2002.

GOEL, N. K.; BURN, D. H.; PANDEY, M. D.; YING, A. Wind quantile estimation using a pooled frequency analysis approach. **Journal of Wind Engineering and Industrial Aerodynamics**, Amsterdam, v. 92, n. 6, p. 509-528, 2004.

GULDEMIR, H.; SENGUR, A. Comparison of clustering algorithms for analog modulation classification. **Expert Systems with Applications**, Amsterdam, v. 30, n. 4, p. 642-649, 2006.

HAIR JR., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise multivariada de dados**. 5. ed., Porto Alegre: Bookman, 2005. 600 p.

HOSKING, J. R. M.; WALLIS, J. R. **Regional frequency analysis**: an approach based on L-moments. New York: Cambridge University Press, 1997. 224 p.

KELLER FILHO, T.; ASSAD, E. D.; LIMA, P. R. S. R. Regiões pluviometricamente homogêneas no Brasil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 40, n. 4, p. 311-322, abr. 2005.

KJELDSEN, T. R.; SMITHERS, J. C.; SCHULZE, R. E. Regional flood frequency analysis in the KwaZulu-Natal province, South Africa, using the index flood method. **Journal of Hydrology**, Amsterdam, v. 255, n. 1-4, p. 194-211, 2002.

LOPES, F. Z. **Relação entre o MEI (Multivariate enso index) e a precipitação pluvial no estado do Rio Grande do Sul**. 2006. 159 f. Dissertação (Mestrado em Meteorologia) – Universidade Federal de Pelotas, Faculdade de Meteorologia, Pelotas, 2006.

LUCHINI, A. M.; SOUZA, M. D.; PINTO, A. L. Aportes e limites da perspectiva de redes de políticas públicas: O caso da gestão da água. **Caderno de Pesquisas em Administração**, São Paulo, v. 10, n. 2, p. 87-94, abr./jun. 2003.

MANLY, B. F. J. **Métodos estatísticos multivariados**: uma introdução. 3. ed. Porto Alegre: Bookman, 2008. 229 p.

MARTINS, R. M.; PINHO, S. Z.; GONÇALVES, S. L. Utilização da técnica hierárquica aglomerativa pelo método do vizinho mais próximo do risco de geada no estado do Paraná para a cultura do milho safrinha. **Revista Energia na Agricultura**, Botucatu, v. 23, n. 3, p. 95-107, 2008.

MELO JÚNIOR, J. C. F.; SEDIYAMA, G. C.; FERREIRA, P. A.; LEAL, B. G. 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**, Campina Grande, v. 10, n. 2, p. 408-416, 2006.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 297 p.

MODARRES, R. Regional maximum wind speed frequency analysis for the arid and semi-arid regions of Iran. **Journal of Arid Environments**, Amsterdam, v. 72, n. 7, p. 1329-1342, 2008.

MOITA NETO, J. M.; MOITA, G. C. Uma introdução à análise exploratória de dados multivariados. **Química Nova**, São Paulo, v. 21, p. 467-469, 1998.

NAGHETTINI, M.; PINTO, E. J. A. **Hidrologia estatística**. Belo Horizonte: CPRM, 2007. 552 p.

OUARDA, T. B. M. J.; BA, K. M.; DIAZ-DELGADO, C.; CARSTEANU, A.; CHOKMANI, K.; GINGRAS, H.; QUENTIN, E.; TRUJILLO, E.; BOBEE, B. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. **Journal of Hydrology**, Amsterdam, v. 348, n. 1-2, p. 40-58, 2008.

PAKHIRA, M. K., BANDYOPADHYAY, S., MAULIK, U. Validity index for crisp and fuzzy clusters. **Pattern Recognition**, Amsterdam, v. 37, n. 3, p. 487–501, 2004.

PEREIRA, D. S. P.; JOHNSSON, R. M. F. Descentralização da gestão dos recursos hídricos em bacias nacionais no Brasil. **REGA**, Porto Alegre, v. 2, n. 1, p. 53-72, jan./jun. 2005.

RAHNAMA, M. B.; ROSTAMI, R. Halil-River basin regional flood frequency analysis based on L-moments approach. **International Journal of Agricultural Research**, Amsterdam, v. 2, n. 3, p. 261-267, 2007.

RAO, A. R.; SRINIVAS, V. V. Regionalization of watersheds by fuzzy cluster analysis. **Journal of Hydrology**, Amsterdam, v. 318, n. 1-4, p. 57-79, 2006a.

RAO, A. R.; SRINIVAS, V. V. Regionalization of watersheds by hybrid-cluster analysis. **Journal of Hydrology**, Amsterdam, v. 318, n.1-4, p.37-56, 2006b.

SAF, B. Assessment of the effects of discordant sites on regional flood frequency analysis. **Journal of Hydrology**, Amsterdam, v. 380, n. 3-4, p. 362-375, 2010.

SILVA, J. F.; D'ANGIOLELLA, G. A climatologia aplicada na manutenção de sistemas hidrológicos. Geo-Simpósio. Novembro 2001. Asunción, PY.

WU, J.; XIONG, H.; CHEN, J. To*Ward*s understanding hierarchical clustering: a data distribution perspective. **Neurocomputing**, Amsterdan, v. 72, n. 10-12, p. 2319-2330, 2009.

YANG, T.; SHAO, Q.; HAO, Z. C.; CHEN, X.; ZHANG, Z.; XU, C. Y.; SUN, L. Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. **Journal of Hydrology**, Amsterdam, v. 380, n. 3-4, p. 386-405, 2010.

APÊNDICES

APÊNDICE A - COORDENADAS GEOGRÁFICAS DAS ESTAÇÕES

Tabela 1A - Coordenadas geográficas das estações utilizadas neste trabalho

Código	Estação (ID)	Município	Latitude	Longitude	Altitude
02250028	X1	SERTANEJA	-22:51:40	-50:52:28	365
02250030	X2	LEÓPOLIS	-22:57:35	-50:46:16	344
02250032	X3	ITAMBARACA	-22:58:20	-50:28:44	402
02250033	X4	ANDIRÁ	-22:57:55	-50:15:58	423
02250035	X5	CAMBARÁ	-22:58:59	-50:00:00	528
02251033	X6	COLORADO	-22:53:53	-51:53:17	487
02251037	X7	CAFEARA	-22:47:16	-51:42:41	377
02251038	X8	ALVORADA DO SUL	-22:46:00	-51:13:59	373
02251039	X9	PRIMEIRO DE MAIO	-22:51:60	-51:01:55	370
02251041	X10	SANTO INÁCIO	-22:41:47	-51:47:23	373
02251042	X11	LUPIONÓPOLIS	-22:41:59	-51:38:32	377
02251069	X12	CENTENÁRIO DO SUL	-22:49:22	-51:35:44	500
02252013	X13	JARDIM OLINDA	-22:33:30	-52:02:11	318
02252015	X14	DIAMANTE DO NORTE	-22:39:15	-52:51:38	329
02252019	X15	PARANAPOEMA	-22:39:39	-52:07:59	299
02252022	X16	TERRA RICA	-22:43:50	-52:36:59	437
02252023	X17	PARANAVAI	-22:43:52	-52:26:47	400
02252024	X18	SANTO ANTÔNIO DO CAIUÁ	-22:43:59	-52:21:00	420
02252025	X19	GUAIRAÇA	-22:57:00	-52:48:00	460
02253008	X20	SÃO PEDRO DO PARANÁ	-22:47:42	-53:09:33	419
02253011	X21	SANTA CRUZ DE MONTE CASTELO	-22:58:00	-53:16:59	482
02253013	X22	LOANDA	-22:55:59	-53:1:59	488
02349036	X23	RIBEIRÃO CLARO	-23:12:00	-49:45:00	782
02349038	X24	JACAREZINHO	-23:07:00	-49:49:59	580
02349059	X25	SANTO ANTÔNIO DA PLATINA	-23:24:50	-49:58:51	603
02349060	X26	CARLÓPOLIS	-23:32:44	-49:44:49	563
02349061	X27	SANTANA DO ITARARÉ	-23:45:16	-49:37:21	543
02349064	X28	SÃO JOSÉ DA BOA VISTA	-23:54:52	-49:39:00	550
02350006	X29	SANTO ANTÔNIO DO PARAÍSO	-23:30:00	-50:39:00	670
02350021	X30	NOVA FÁTIMA	-23:21:00	-50:37:59	570
02350023	X31	URAÍ	-23:12:26	-50:47:43	458
02350027	X32	CORNÉLIO PROCÓPIO	-23:07:00	-50:40:59	580
02350029	X33	SANTA AMÉLIA	-23:15:59	-50:25:55	471
02350032	X34	ASSAÍ	-23:23:38	-50:55:26	533
02350037	X35	SÃO JERÔNIMO DA SERRA	-23:46:50	-50:48:51	989
02350041	X36	IBAITI	-23:55:00	-50:15:00	600
02350043	X37	PINHALÃO	-23:57:00	-50:01:00	750
02350048	X38	CONGONHINHAS	-23:38:02	-50:28:28	531
02350052	X39	RIBEIRÃO DO PINHAL	-23:33:00	-50:24:00	800
02350053	X40	JUNDIAÍ DO SUL	-23:27:00	-50:13:59	500
02350054	X41	GUAPIRAMA	-23:31:00	-50:01:59	600
02351004	X42	BOM SUCESSO	-23:42:00	-51:46:00	560

Código	Estação (ID)	Município	Latitude	Longitude	Altitude
02351020	X43	BORRAZÓPOLIS	-23:56:27	-51:35:16	640
02351023	X44	SÃO PEDRO DO IVAÍ	-23:51:51	-51:51:30	404
02351025	X45	NOVO ITACOLOMI	-23:45:50	-51:30:25	606
02351026	X46	RIO BOM	-23:45:50	-51:24:39	648
02351027	X47	MARILÂNDIA DO SUL	-23:49:37	-51:15:59	860
02351028	X48	ITAMBÉ	-23:39:00	-51:58:59	420
02351029	X49	MARIALVA	-23:36:35	-51:51:36	372
02351031	X50	CAMBÉ	-23:03:58	-51:15:40	438
02351032	X51	SERTANÓPOLIS	-23:03:00	-51:01:59	380
02351037	X52	CALIFORNIA	-23:39:00	-51:21:00	790
02351040	X53	LONDRINA	-23:45:00	-51:01:36	673
02351041	X54	ORTIGUEIRA	-23:58:59	-51:04:59	1011
02351043	X55	CAMBIRA	-23:39:46	-51:36:09	601
02351044	X56	SARANDI	-23:29:02	-51:54:20	504
02351045	X57	MARINGÁ	-23:24:00	-51:52:26	584
02351048	X58	ARAPONGAS	-23:24:00	-51:26:00	793
02351050	X59	IGUARAÇU	-23:10:59	-51:49:59	581
02351051	X60	ASTORGA	-23:14:14	-51:39:41	572
02351053	X61	ROLÂNDIA	-23:12:00	-51:27:00	653
02352026	X62	TUNEIRAS DO OESTE	-23:54:24	-52:57:17	459
02352029	X63	PEABIRU	-23:54:39	-52:20:10	527
02352030	X64	PEABIRU	-23:59:40	-52:11:46	425
02352031	X65	CIANORTE	-23:48:00	-52:37:59	600
02352032	X66	ARARUNA	-23:49:59	-52:30:00	600
02352033	X67	ENGENHEIRO BELTRÃO	-23:48:00	-52:19:59	550
02352035	X68	TERRA BOA	-23:40:23	-52:22:50	474
02352036	X69	SÃO CARLOS DO IVAÍ	-23:21:49	-52:31:26	293
02352037	X70	FLORAÍ	-23:19:26	-52:17:58	521
02352038	X71	OURIZONA	-23:24:15	-52:11:45	561
02352042	X72	TAPEJARA	-23:40:05	-52:58:34	447
02352043	X73	RONDON	-23:34:00	-52:51:00	500
02352044	X74	INDIANÓPOLIS	-23:28:58	-52:42:05	501
02352045	X75	JAPURÁ	-23:28:00	-52:33:00	500
02352046	X76	CIDADE GAUCHA	-23:22:59	-52:55:59	400
02352047	X77	GUAPOREMA	-23:19:59	-52:46:00	400
02352051	X78	AMAPORÃ	-23:05:07	-52:47:05	396
02352060	X79	PLANALTINA DO PARANÁ	-23:04:38	-52:57:33	362
02352061	X80	IVATUBA	-23:37:01	-52:11:47	339
02352062	X81	NOVA ESPERANÇA	-23:10:59	-52:10:59	582
02353005	X82	XAMBRE	-23:44:03	-53:29:11	412
02353010	X83	QUERÊNCIA DO NORTE	-23:04:54	-53:28:52	349
02353016	X84	PÉROLA	-23:47:49	-53:40:32	438
02353019	X85	SÃO JORGE DO PATROCÍNIO	-23:41:35	-53:54:32	365
02353023	X86	MARIA HELENA	-23:36:28	-53:12:15	370
02353027	X87	UMUARAMA	-23:31:39	-53:27:46	441
02353031	X88	ICARAIMA	-23:22:59	-53:37:00	450
02353032	X89	IVATÉ	-23:19:59	-53:25:00	500
02353033	X90	DOURADINA	-23:22:00	-53:16:59	450

Código	Estação (ID)	Município	Latitude	Longitude	Altitude
02353034	X91	TAPIRA	-23:19:11	-53:04:12	401
02353038	X92	SANTA ISABEL DO IVAÍ	-23:07:52	-53:16:38	339
02353041	X93	SANTA MÔNICA	-23:10:59	-53:04:00	300
02353047	X94	ALTÔNIA	-23:57:00	-53:58:00	400
02449011	X95	PIRAÍ DO SUL	-24:31:45	-49:55:44	1068
02449040	X96	JAGUARIAIVA	-24:14:45	-49:43:00	890
02449044	X97	SENGES	-24:06:00	-49:28:00	650
02449045	X98	SÃO JOSÉ DA BOA VISTA	-24:04:00	-49:39:00	850
02450021	X99	PONTA GROSSA	-24:58:59	-50:16:00	950
02450024	X100	CARAMBEÍ	-24:57:00	-50:00:00	1000
02450026	X101	CASTRO	-24:37:59	-50:07:59	1050
02450034	X102	TIBAGI	-24:13:00	-50:13:59	1050
02450036	X103	ARAPOTI	-24:15:31	-50:05:02	957
02450040	X104	RESERVA	-24:29:31	-50:49:07	919
02450049	X105	IVAÍ	-24:57:27	-50:53:30	743
02450054	X106	IPIRANGA	-24:52:00	-50:39:00	950
02451021	X107	SANTA MARIA DO OESTE	-24:46:45	-51:57:16	929
02451022	X108	IVAIPORÃ	-24:15:00	-51:32:00	720
02451023	X109	GRANDES RIOS	-24:11:00	-51:26:00	700
02451027	X110	BOA VENTURA DE SÃO ROQUE	-24:54:30	-51:39:28	906
02451036	X111	PITANGA	-24:38:28	-51:45:29	911
02451038	X112	CÂNDIDO DE ABREU	-24:37:00	-51:16:00	900
02451039	X113	RESERVA	-24:40:59	-51:1:59	1200
02451044	X114	NOVA TEBAS	-24:25:00	-51:56:00	700
02451046	X115	ARIRANHA DO IVAÍ	-24:22:00	-51:30:00	900
02451047	X116	RIO BRANCO DO IVAÍ	-24:19:00	-51:18:00	675
02451049	X117	JARDIM ALEGRE	-24:06:52	-51:44:03	618
02452008	X118	IRETAMA	-24:25:00	-52:06:00	584
02452009	X119	UBIRATÃ	-24:32:00	-52:59:00	509
02452010	X120	JANIÓPOLIS	-24:08:00	-52:46:00	350
02452011	X121	CAMPINA DA LAGOA	-24:35:59	-52:48:15	618
02452012	X122	ALTAMIRA DO PARANÁ	-24:48:00	-52:42:00	650
02452015	X123	RONCADOR	-24:36:00	-52:160:0	730
02452016	X124	PALMITAL	-24:53:04	-52:12:10	890
02452019	X125	LARANJAL	-24:53:09	-52:28:26	741
02452029	X126	FAROL	-24:05:26	-52:37:17	582
02452035	X127	MAMBORÊ	-24:25:59	-52:33:00	650
02452044	X128	IRETAMA	-24:25:00	-52:12:00	700
02452045	X129	CAMPO MOURÃO	-24:14:05	-52:24:09	668
02452046	X130	LUIZIANA	-24:16:59	-52:16:00	800
02453008	X131	ALTO PIQUIRI	-24:00:53	-53:26:23	427
02453010	X132	FORMOSA DO OESTE	-24:16:59	-53:19:00	370
02453012	X133	CORBÉLIA	-24:48:00	-53:18:00	682
02453014	X134	CAMPO BONITO	-24:52:59	-53:04:00	700
02453016	X135	GOIOERÊ	-24:11:36	-53:01:55	497
02453026	X136	OURO VERDE DO OESTE	-24:46:31	-53:54:06	554
02453027	X137	TOLEDO	-24:46:20	-53:38:33	635
02453028	X138	TOLEDO	-24:37:12	-53:55:34	539

Código	Estação (ID)	Município	Latitude	Longitude	Altitude
02453030	X139	ASSIS CHATEAUBRIAND	-24:36:40	-53:36:51	517
02453037	X140	NOVA AURORA	-24:34:23	-53:22:48	544
02453047	X141	MARIPA	-24:25:00	-53:49:00	400
02453048	X142	NOVA SANTA ROSA	-24:23:31	-53:55:57	392
02453050	X143	BRASILÂNDIA DO SUL	-24:11:54	-53:31:32	396
02453052	X144	FRANCISCO ALVES	-24:04:59	-53:57:00	350
02453056	X145	CASCAVEL	-24:57:46	-53:14:38	697
02454003	X146	ENTRE RIOS DO OESTE	-24:41:33	-54:13:57	245
02454004	X147	PATO BRAGADO	-24:38:53	-54:17:54	337
02454006	X148	TERRA ROXA	-24:10:00	-54:06:00	400
02454011	X149	VERA CRUZ DO OESTE	-24:58:38	-54:00:00	570
02454012	X150	SANTA HELENA	-24:46:44	-54:14:22	281
02454015	X151	MERCEDES	-24:27:00	-54:10:00	364
02454016	X152	GUAÍRA	-24:19:00	-54:13:00	249
02454018	X153	DIAMANTE D'OESTE	-24:54:22	-54:12:05	243
02549040	X154	CONTENDA	-25:40:48	-49:32:11	882
02549059	X155	LAPA	-25:48:00	-49:52:58	903
02549063	X156	TIJUCAS DO SUL	-25:47:00	-49:09:00	913
02550029	X157	IRATI	-25:28:00	-50:47:00	797
02550035	X158	REBOUÇAS	-25:420:0	-50:31:00	790
02550037	X159	SÃO JOÃO DO TRIUNFO	-25:37:19	-50:12:01	856
02550041	X160	PALMEIRA	-25:28:30	-50:17:53	892
02550045	X161	TEIXEIRA SOARES	-25:22:00	-50:28:00	950
02550048	X162	IMBITUVA	-25:14:15	-50:36:02	869
02550055	X163	PRUDENTÓPOLIS	-25:09:00	-50:59:00	750
02551009	X164	GUARAPUAVA	-25:06:33	-51:48:23	1056
02551011	X165	INÁCIO MARTINS	-25:37:47	-51:05:16	1150
02551017	X166	CRUZ MACHADO	-25:56:38	-51:15:44	880
02551019	X167	PINHÃO	-25:51:00	-51:46:00	1245
02552006	X168	GUARANIAÇU	-25:05:00	-52:53:00	920
02552007	X169	LARANJEIRAS DO SUL	-25:24:00	-52:25:00	850
02552008	X170	MARQUINHO	-25:06:44	-52:15:30	872
02552010	X171	NOVA LARANJEIRAS	-25:18:00	-52:32:00	728
02552022	X172	RESERVA DO IGUAÇU	-25:48:00	-52:01:00	1000
02552025	X173	CANDÓI	-25:33:04	-52:06:25	782
02552029	X174	SÃO JOÃO	-25:51:00	-52:44:00	680
02552031	X175	CHOPINZINHO	-25:49:00	-52:25:00	650
02552036	X176	RIO BONITO DO IGUAÇU	-25:29:23	-52:31:56	704
02552037	X177	PORTO BARREIRO	-25:31:00	-52:24:00	750
02552039	X178	ESPIGÃO ALTO DO IGUAÇU	-25:23:08	-52:46:12	621
02552040	X179	VIRMOND	-25:22:50	-52:12:02	758
02552042	X180	ITAPEJARA D'OESTE	-25:56:26	-52:49:30	587
02552045	X181	SÃO JORGE D'OESTE	-25:43:00	-52:55:00	550
02552046	X182	QUEDAS DO IGUAÇU	-25:23:34	-52:57:52	666
02553005	X183	PÉROLA D'OESTE	-25:50:00	-53:45:00	400
02553009	X184	CÉU AZUL	-25:07:59	-53:51:00	610
02553010	X185	AMPÉRE	-25:49:00	-53:30:00	400
02553014	X186	SERRANÓPOLIS DO IGUAÇU	-25:34:59	-53:58:59	350

Código	Estação (ID)	Município	Latitude	Longitude	Altitude
02553019	X187	IBEMA	-25:06:00	-53:04:00	750
02553024	X188	CAPITÃO LEÔNIDAS MARQUES	-25:29:00	-53:37:00	264
02553026	X189	BOA VISTA DA APARECIDA	-25:23:40	-53:23:05	478
02553028	X190	SANTA LÚCIA	-25:24:00	-53:34:00	380
02553033	X191	SANTA TEREZA DO OESTE	-25:09:00	-53:37:00	668
02553038	X192	PLANALTO	-25:46:11	-53:39:52	400
02553046	X193	ENÉAS MARQUES	-25:53:15	-53:4:57	560
02554005	X194	MATELÂNDIA	-25:14:24	-53:58:31	581
02554012	X195	SANTA TEREZINHA DE ITAIPU	-25:26:27	-54:24:11	285
02554013	X196	SÃO MIGUEL DO IGUAÇU	-25:27:00	-54:19:00	250
02554020	X197	MISSAL	-25:5:14	-54:14:53	321
02651010	X198	GENERAL CARNEIRO	-26:37:59	-51:19:59	950
02651016	X199	UNIÃO DA VITÓRIA	-26:03:00	-51:12:00	800
02651023	X200	BITURUNA	-26:07:00	-51:34:00	1030
02652010	X201	PALMAS	-26:29:00	-52:00:00	1060
02652011	X202	MARIÓPOLIS	-26:21:00	-52:34:00	850
02652012	X203	VITORINO	-26:16:00	-52:48:00	710
02652013	X204	PATO BRANCO	-26:14:00	-52:41:00	800
02652032	X205	CORONEL VIVIDA	-26:05:00	-52:31:00	700
02653014	X206	SANTO ANTÔNIO DO SUDOESTE	-26:07:00	-53:39:00	447
02653015	X207	FRANCISCO BELTRÃO	-26:05:00	-53:12:00	757
02653017	X208	SALGADO FILHO	-26:06:44	-53:27:48	500
02653019	X209	FLOR DA SERRA DO SUL	-26:14:00	-53:12:00	700
02653021	X210	BARRACÃO	-26:13:00	-53:29:00	550
02448037	X211	ADRIANÓPOLIS	-24:45:23	-48:58:05	227
02449021	X212	DOUTOR ULYSSES	-24:34:01	-49:25:10	818
02449023	X213	CERRO AZUL	-24:51:00	-49:28:00	480
02449024	X214	TUNAS DO PARANÁ	-24:58:00	-49:04:59	880
02548043	X215	GUARAQUEÇABA	-25:13:59	-48:25:00	64
02548047	X216	MORRETES	-25:22:59	-48:52:00	159
02548052	X217	SÃO JOSÉ DOS PINHAIS	-25:48:46	-48:55:25	237
02548068	X218	ANTONINA	-25:26:00	-48:46:00	74
02549047	X219	CAMPO LARGO	-25:14:00	-49:38:00	800
02549051	X220	BOCAIÚVA DO SUL	-25:12:00	-49:07:00	980
02549053	X221	ITAPERUÇU	-25:07:59	-49:33:00	750
02551008	X222	GUARAPUAVA	-25:33:00	-51:29:00	1000
02551027	X223	GUARAPUAVA	-25:33:00	-51:33:00	1000
02551034	X224	GUARAPUAVA	-25:18:00	-51:26:00	1050
02551035	X225	GUARAPUAVA	-25:16:00	-51:15:00	1202
02652027	X226	MANGUEIRINHA	-26:07:38	-52:11:002	1009
02548042	X227	GUARAQUEÇABA	-25:04:59	-48:13:00	9

APÊNDICE B - CONFIGURAÇÃO DAS METODOLOGIAS DE AGRUPAMENTO

Tabela 1B - Configuração das metodologias de agrupamento k-médias, *ward* e sua forma híbrida

Código	Estação (ID)	*Ward*	Híbrido	k-médias (5)	k-médias (7)	k-médias (9)
02250028	X1	1	1	2	5	3
02250030	X2	1	1	2	5	3
02250032	X3	1	1	2	5	3
02250033	X4	1	1	2	5	3
02250035	X5	1	1	2	5	3
02251033	X6	2	2	2	5	8
02251037	X7	2	1	2	5	3
02251038	X8	1	1	2	5	3
02251039	X9	1	1	2	5	3
02251041	X10	2	1	2	5	3
02251042	X11	2	1	2	5	3
02251069	X12	2	1	2	5	3
02252013	X13	2	2	2	7	8
02252015	X14	2	2	2	7	8
02252019	X15	2	2	2	7	8
02252022	X16	2	2	2	7	8
02252023	X17	2	2	2	7	8
02252024	X18	2	2	2	7	8
02252025	X19	2	2	2	7	8
02253008	X20	2	2	2	7	8
02253011	X21	2	2	2	7	8
02253013	X22	2	2	2	7	8
02349036	X23	1	1	2	5	3
02349038	X24	1	1	2	5	3
02349059	X25	1	1	2	5	3
02349060	X26	1	1	2	5	3
02349061	X27	1	1	4	5	3
02349064	X28	1	1	4	5	3
02350006	X29	1	1	2	5	3
02350021	X30	1	1	2	5	3
02350023	X31	1	1	2	5	3
02350027	X32	1	1	2	5	3
02350029	X33	1	1	2	5	3
02350032	X34	1	1	2	5	3
02350037	X35	2	1	2	5	1
02350041	X36	1	1	2	5	3
02350043	X37	1	1	4	5	3
02350048	X38	1	1	2	5	3
02350052	X39	1	1	2	5	3
02350053	X40	1	1	2	5	3
02350054	X41	1	1	2	5	3
02351004	X42	2	2	2	7	1
02351020	X43	2	2	2	6	1
02351023	X44	2	2	2	7	1
02351025	X45	2	2	2	5	1
02351026	X46	2	2	2	5	1
02351027	X47	2	2	2	5	1

Código	Estação (ID)	*Ward*	Híbrido	k-médias (5)	k-médias (7)	k-médias (9)
02351028	X48	2	2	2	7	1
02351029	X49	2	2	2	7	1
02351031	X50	2	1	2	5	3
02351032	X51	2	1	2	5	3
02351037	X52	2	2	2	5	1
02351040	X53	2	1	2	5	1
02351041	X54	2	3	2	6	1
02351043	X55	2	2	2	7	1
02351044	X56	2	2	2	7	1
02351045	X57	2	2	2	7	1
02351048	X58	2	1	2	5	1
02351050	X59	2	2	2	5	3
02351051	X60	2	2	2	5	1
02351053	X61	2	1	2	5	3
02352026	X62	2	2	1	7	4
02352029	X63	2	2	2	7	1
02352030	X64	2	2	2	6	1
02352031	X65	2	2	2	7	1
02352032	X66	2	2	2	7	1
02352033	X67	2	2	2	7	1
02352035	X68	2	2	2	7	1
02352036	X69	2	2	2	7	8
02352037	X70	2	2	2	7	1
02352038	X71	2	2	2	7	1
02352042	X72	2	2	2	7	8
02352043	X73	2	2	2	7	8
02352044	X74	2	2	2	7	1
02352045	X75	2	2	2	7	1
02352046	X76	2	2	2	7	8
02352047	X77	2	2	2	7	8
02352051	X78	2	2	2	7	8
02352060	X79	2	2	2	7	8
02352061	X80	2	2	2	7	1
02352062	X81	2	2	2	7	8
02353005	X82	2	2	2	7	8
02353010	X83	2	2	2	7	8
02353016	X84	2	2	2	7	8
02353019	X85	2	2	2	7	8
02353023	X86	2	2	2	7	8
02353027	X87	2	2	2	7	8
02353031	X88	2	2	2	7	8
02353032	X89	2	2	2	7	8
02353033	X90	2	2	2	7	8
02353034	X91	2	2	2	7	8
02353038	X92	2	2	2	7	8
02353041	X93	2	2	2	7	8
02353047	X94	3	3	1	2	4
02449011	X95	4	4	4	4	6
02449040	X96	1	1	4	5	3
02449044	X97	1	1	4	5	3
02449045	X98	1	1	4	5	3
02450021	X99	4	4	4	4	6
02450024	X100	4	4	4	4	6
02450026	X101	4	4	4	4	6

Código	Estação (ID)	*Ward*	Híbrido	k-médias (5)	k-médias (7)	k-médias (9)
02450034	X102	1	1	4	5	3
02450036	X103	1	1	4	5	3
02450040	X104	3	3	1	6	2
02450049	X105	4	4	4	4	6
02450054	X106	4	4	4	4	6
02451021	X107	3	3	1	6	2
02451022	X108	2	3	1	6	2
02451023	X109	2	3	1	6	2
02451027	X110	3	3	1	6	2
02451036	X111	3	3	1	6	2
02451038	X112	4	3	1	6	2
02451039	X113	4	3	1	6	2
02451044	X114	2	3	1	6	2
02451046	X115	2	3	1	6	2
02451047	X116	2	3	1	6	2
02451049	X117	2	2	2	7	1
02452008	X118	2	3	1	6	2
02452009	X119	3	3	1	6	2
02452010	X120	2	3	1	6	2
02452011	X121	3	3	1	6	2
02452012	X122	3	3	1	6	2
02452015	X123	3	3	1	6	2
02452016	X124	3	3	1	6	2
02452019	X125	3	3	1	6	2
02452029	X126	2	3	1	6	2
02452035	X127	3	3	1	6	2
02452044	X128	2	3	1	6	2
02452045	X129	2	3	1	6	2
02452046	X130	2	3	1	6	2
02453008	X131	3	3	1	2	4
02453010	X132	3	3	1	2	4
02453012	X133	3	3	1	2	5
02453014	X134	3	3	1	2	5
02453016	X135	2	3	1	6	4
02453026	X136	3	3	1	2	4
02453027	X137	3	3	1	2	4
02453028	X138	3	3	1	2	4
02453030	X139	3	3	1	2	4
02453037	X140	3	3	1	2	4
02453047	X141	3	3	1	2	4
02453048	X142	3	3	1	2	4
02453050	X143	3	3	1	2	4
02453052	X144	3	3	1	2	4
02453056	X145	3	3	1	2	5
02454003	X146	3	3	1	2	4
02454004	X147	3	3	1	2	4
02454006	X148	3	3	1	2	4
02454011	X149	5	3	1	2	5
02454012	X150	3	3	1	2	4
02454015	X151	3	3	1	2	4
02454016	X152	3	3	1	2	4
02454018	X153	3	3	1	2	4
02549040	X154	4	4	4	4	6
02549059	X155	4	4	4	4	6

Código	Estação (ID)	*Ward*	Híbrido	k-médias (5)	k-médias (7)	k-médias (9)
02549063	X156	4	4	4	4	6
02550029	X157	4	4	4	4	6
02550035	X158	4	4	4	4	6
02550037	X159	4	4	4	4	6
02550041	X160	4	4	4	4	6
02550045	X161	4	4	4	4	6
02550048	X162	4	4	4	4	6
02550055	X163	4	4	4	4	6
02551009	X164	5	3	1	6	2
02551011	X165	5	5	5	3	9
02551017	X166	5	5	5	3	9
02551019	X167	5	5	5	3	9
02552006	X168	3	5	5	3	5
02552007	X169	5	5	5	3	5
02552008	X170	3	5	5	3	5
02552010	X171	5	5	5	3	5
02552022	X172	5	5	5	3	9
02552025	X173	5	5	5	3	9
02552029	X174	5	5	5	3	9
02552031	X175	5	5	5	3	9
02552036	X176	5	5	5	3	5
02552037	X177	5	5	5	3	5
02552039	X178	5	5	5	3	5
02552040	X179	5	5	5	3	5
02552042	X180	5	5	5	3	9
02552045	X181	5	5	5	3	5
02552046	X182	5	5	5	3	5
02553005	X183	5	5	5	3	5
02553009	X184	5	5	5	2	5
02553010	X185	5	5	5	3	5
02553014	X186	5	5	5	2	5
02553019	X187	3	5	5	3	5
02553024	X188	5	5	5	3	5
02553026	X189	5	5	5	3	5
02553028	X190	5	5	5	3	5
02553033	X191	5	5	5	3	5
02553038	X192	5	5	5	3	5
02553046	X193	5	5	5	3	9
02554005	X194	5	5	5	2	5
02554012	X195	5	5	5	2	5
02554013	X196	5	5	5	3	5
02554020	X197	3	3	1	2	5
02651010	X198	5	5	5	3	9
02651016	X199	5	5	5	3	9
02651023	X200	5	5	5	3	9
02652010	X201	5	5	5	3	9
02652011	X202	5	5	5	3	9
02652012	X203	5	5	5	3	9
02652013	X204	5	5	5	3	9
02652032	X205	5	5	5	3	9
02653014	X206	5	5	5	3	9
02653015	X207	5	5	5	3	9
02653017	X208	5	5	5	3	9
02653019	X209	5	5	5	3	9

Código	Estação (ID)	*Ward*	Híbrido	k-médias (5)	k-médias (7)	k-médias (9)
02653021	X210	5	5	5	3	9
02448037	X211	4	4	4	4	6
02449021	X212	4	4	4	4	6
02449023	X213	4	4	4	4	6
02449024	X214	4	4	4	4	6
02548043	X215	6	6	3	1	7
02548047	X216	6	6	3	1	7
02548052	X217	6	6	3	1	7
02548068	X218	6	6	3	1	7
02549047	X219	4	4	4	4	6
02549051	X220	4	4	4	4	6
02549053	X221	4	4	4	4	6
02551008	X222	5	5	5	3	9
02551027	X223	5	5	5	3	9
02551034	X224	5	5	5	3	2
02551035	X225	5	5	5	3	9
02652027	X226	5	5	5	3	9
02548042	X227	6	6	3	1	7