3 MATERIAL E MÉTODOS

3.1 Fundamentação Teórica

Quando se quer obter informações de um grupo de variáveis ou de um conjunto total dos dados de uma região, usualmente recorre-se à análise multivariada. A análise multivariada é uma técnica estatística cujos principais objetivos são:

- reduzir a dimensão de uma matriz de dados;
- investigar o comportamento espacial e temporal das variáveis consideradas;
- obter grupos homogêneos das variáveis.

Existem várias técnicas de manejo de dados dentro da análise multivariada que são muito úteis para trabalhos em climatologia como, por exemplo, a Análise de Componentes Principais (ACP) e a Análise de Agrupamento (AA) as quais são utilizadas em processos de classificação, embora a primeira não tenha como único objetivo este fim.

Neste trabalho foi utilizada a técnica estatística multivariada da AA na determinação das regiões homogêneas das temperaturas médias trimestrais (JFM), (AMJ), (JAS) e (OND) para o período de 1913 a 2002.

Os métodos da AA partem de uma matriz de dados $\mathbf{X}_{(nxp)}$ cujas linhas correspondem “n” unidades (indivíduos) fornecendo “p” características (valores numéricos) cada. No caso de estudos climatológicos, essas unidades podem ser estações meteorológicas; as características seriam dados meteorológicos como pressão, vento, temperatura etc, ordenados ou não segundo seqüências cronológicas. No caso de uma única variável, estes dados podem constituir uma seqüência cronológica, onde 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) dada.

A matriz de dados é organizada da seguinte forma:

$$\mathbf{X}_{(nxp)} = \begin{pmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & . & x_{2p} \\ . & & & & & \\ . & & & & & \\ x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix} \qquad (1)$$

Ou:

$$\mathbf{X}_{(nxp)} = \begin{pmatrix} X_1 & X_2 & . & . & . & X_p \end{pmatrix} \qquad (2)$$

Observa-se que a matriz $\mathbf{X}_{(nxp)}$ pode ser interpretada como um ordenamento de “p” vetores colunas (cada vetor equivalente à série temporal de cada uma das “p” estações), ou de “n” vetores linha (cada vetor indicando valores de uma determinada rede de estações numa época dada). O primeiro caso descreve principalmente o comportamento temporal de uma rede, enquanto que o segundo ilustra a situação espacial da variável em cada época.

3.1.1 Análise de Agrupamento

Conforme Everitt (1974), classificação pode ser entendida como um processo para localizar entidades em classes inicialmente indefinidas, de modo que indivíduos da mesma classe sejam similares, entre si, em algum sentido definido pelas variáveis consideradas. Essas classes de indivíduos similares serão então os agrupamentos.

De acordo com Curi (1983), a análise de agrupamento não pressupõe a existência de grupos; as etapas desta metodologia visam transformar um conjunto heterogêneo de unidades, em grupos que se caracterizam pela homogeneidade interna e pela heterogeneidade externa.

Dentre os diversos autores que abordam o estudo da Análise de Agrupamento algumas definições são feitas como, por exemplo, as de Everitt (1974), descritas a seguir:

- Técnicas de Partição – Otimização;
- Técnicas de Densidade;
- Técnicas de Agrupamentos com interseção;
- Outras (que não se encaixam claramente dentro das citadas acima).

As técnicas da AA podem ser hierárquicas e não-hierárquicas, na primeira são produzidas seqüências de partições em classes cada vez mais vastas e nas não-hierárquicas, efetua-se uma partição em um número de classes fixado a “priori”.
As técnicas hierárquicas de agrupamento ainda se subdividem em:

- Aglomerativas - onde através de fusões sucessivas dos “n” elementos vão sendo obtidos n-1, n-2, etc, grupos, até reunir todos os elementos num único grupo;
- Divisivas – partem de um único grupo e por divisões sucessivas vão sendo obtidos dois, três, etc. grupos.

Nestes processos a união de dois grupos numa determinada etapa, produz um dos agrupamentos da etapa superior o que caracteriza o processo hierárquico.

3.1.2 Função de Agrupamento

Na análise de agrupamento, todos os processos de hierarquização são similares, iniciando-se pela determinação de uma função de agrupamento. Esta função é usada como critério para medir a distância entre dois objetos ou para estabelecer o quanto eles são parecidos. Alguns autores chamam-na de coeficiente de parecença, que pode se dividir em duas categorias: medida de similaridade e dissimilaridade. Na primeira, quanto maior o valor observado mais parecidos são os objetos. Já na segunda, quanto maior o valor observado menos parecidos serão os objetos. Um exemplo de medida de similaridade é o coeficiente de correlação e de dissimilaridade, é a distância euclidiana. A maioria dos algoritmos utilizam medidas de similaridade e dissimilaridade entre os elementos de um conjunto, cujos valores de similaridade (dissimilaridade) calculados com base na matriz de dados originais $\mathbf{X}_{(nxp)}$ são representados por uma matriz de distâncias ou de similaridade denominada matriz de parecença.

3.1.3 Distância euclidiana e métodos de agrupamento

De acordo com Wilks (1995), a idéia central do agrupamento de um conjunto de dados distribuídos em pontos é a idéia de distância. A mais utilizada é a distância euclidiana em dados de vetores num espaço P-dimensional. Assim, a distância entre dois pontos x_i e x_j é:

$$d_{ij} = \left\| x_i - x_j \right\| = \left[\sum_{k=1}^{P} \left(x_{i,k} - x_{j,k} \right)^2 \right]^{1/2} \qquad (3)$$

Quando se usam grandezas que não são diretamente comparáveis (com unidades e naturezas diferentes), a mudança de uma das unidades de medida pode alterar completamente o significado e o valor do coeficiente de parecença. Essa é uma das razões pelas quais a padronização ou redução das variáveis se

torna necessário. A redução é feita da seguinte forma: a matriz $\mathbf{X}_{(nxp)}$ é transformada numa outra $\mathbf{Z}_{(nxp)}$ tal que

$$Z_{i,k} = \frac{x_{ik} - \bar{x}_k}{s_k} \qquad (4.a)$$

$$Z_{j,k} = \frac{x_{jk} - \bar{x}_k}{s_k} \qquad (4.b)$$

sendo $\bar{x}_k$ e s_k a média e o desvio padrão dos valores na k-ésima coluna.
Feita a transformação a distância euclidiana passa a ser :

$$d_{ij} = \left\| z_i - z_j \right\| = \left[\sum_{k=1}^{P} \left(z_{i,k} - z_{j,k} \right)^2 \right]^{1/2} \qquad (5)$$

ou, substituindo (4.a) e (4.b) em (5), temos:

$$d_{ij} = \left\| x_i - x_j \right\| = \left[\sum_{k=1}^{P} \left(\frac{x_i - x_j}{s_k} \right)^2 \right]^{1/2} \qquad (6)$$

que é a média dos desvios quadráticos dos dados padronizados.

Tanto a distância euclidiana (medida de dissimilaridade) quanto o coeficiente de correlação (medida de similaridade) podem ser usados na determinação dos grupos. Para este fim existem dois métodos de agrupamento, os hierárquicos e os não hierárquicos. Nos métodos hierárquicos várias técnicas ou critérios de agrupamentos são possíveis. Dentre eles podemos destacar, os relacionados por Wilks (1995), que são os seguintes: o método da ligação simples ou vizinho mais próximo, o da ligação completa ou vizinho mais distante, o método da centróide e

o método de Ward. Estes métodos serão aqui descritos por terem sido utilizados neste trabalho e em vários trabalhos encontrados na bibliografia.

a) - Método da ligação simples ou vizinho mais próximo.

Este método usa como medida de parecença entre dois grupos, tanto a similaridade quanto a distância. Por este método a distância entre os grupos é definida como a distância entre seus membros mais próximos, ou seja, os grupos com menor distância se fundem primeiro. Assim, a distância entre dois grupo G_1 e G_2 será definida por:

$$d_{G_1G_2} = \min \left[d_{ij}\right] \qquad (7)$$

onde: $i \in G_1, j \in G_2$

No caso de similaridade (definida por s),

$$s_{G_1G_2} = \max \left[s_{ij}\right] \qquad (8)$$

onde: $i \in G_1, j \in G_2$

b) - Método da ligação completa ou vizinho mais distante.

É exatamente o oposto ao da ligação simples. Aqui os grupos são formados fundindo os membros mais distantes entre os grupos. No caso da parecença ser definida pela distância, esta será calculada por:

$$d_{G_1G_2} = \max \left[d_{ij}\right] \qquad (9)$$

onde: $i \in G_1, j \in G_2$

No caso de similaridade (definida por s),

$$s_{G_1G_2} = \min \left[s_{ij} \right] \qquad (10)$$

onde: $i \in G_1, j \in G_2$

c) - Método da centróide.

Este processo é o mais direto, pois substitui cada fusão de grupos por um único ponto representado pelas coordenadas de seu centro. A distância entre os grupos é definida pela distância entre os centros. Em cada etapa procura-se fundir grupos que tenham a menor distância entre si. A distância entre G_1 e G_2 será:

$$d_{G_1G_2} = \left\| \bar{x}_{G_1} - \bar{x}_{G_2} \right\| \qquad (11a)$$

onde os vetores médios são definidos conforme equação abaixo.

$$\bar{X}_g = \frac{1}{n_g} \left[\mathbf{X}_g \right]^T \cdot 1 \qquad (11b)$$

d) - Método de Ward

Ward (1963), citado por Everitt (1974), propôs que em qualquer estágio de uma análise, a perda de informações, que resulta do agrupamento de indivíduos em grupos, pode ser medida pela soma total do quadrado dos desvios de todos os pontos em torno da média do grupo para o qual estão contidos. Em cada passo dentro da análise a união de todos os pares possíveis do grupo é considerada, e os dois grupos cuja fusão resulte no mínimo incremento da inércia, ou seja, cuja fusão resulta no menor incremento do erro na soma dos desvios quadráticos são combinados. O agrupamento é feito diretamente através da equação:

$$W = \sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum x_i\right)^2 \qquad (12)$$

As etapas descritas anteriormente, embora instrutivas acerca do processo de agrupar indivíduos, não facilitam a interpretação dos resultados a respeito de como serão obtidos os grupos e o número deles num processo de agrupamento. Necessita-se de instrumentos mais apropriados; um deles é o dendrograma que é uma representação gráfica muito usada na análise de agrupamento para apresentar o resultado de um agrupamento.

Entendendo que esta análise sempre deve ser aplicada com caráter introdutório, e nesse sentido tem seu mérito, a consideração a ser adotada é que o melhor método para a obtenção dos grupos é aquele que fornece os resultados mais coerentes com a realidade climatológica da região estudada. Existem dois métodos de determinação do número de grupos a ser obtido e dos indivíduos que os constituem: um método objetivo (variação da inércia) e outro subjetivo (corte transversal no dendrograma) que determina, através da necessidade e dos dados agrupados, o número de grupos a ser obtido.

3.2 Dados

3.2.1 Área de estudo

Para este trabalho a área utilizada foi o Estado do Rio Grande do Sul o qual localiza-se entre as latitudes de 27°05’ e 33°45’ Sul e longitudes de 49°43’ e 57°39’ Oeste. Sua distribuição topográfica, constituída de vales, depressões e regiões montanhosas está representada na Figura 1, as quais contribuem na distribuição da temperatura média do ar.

Utilizando o sistema de classificação climática de Köppen, Berlato (1983) observou que o Rio Grande do Sul se enquadra na zona fundamental temperada

(C), tipo fundamental úmido (Cf) com duas variedades específicas: subtropical (Cfa) e temperada (Cfb)

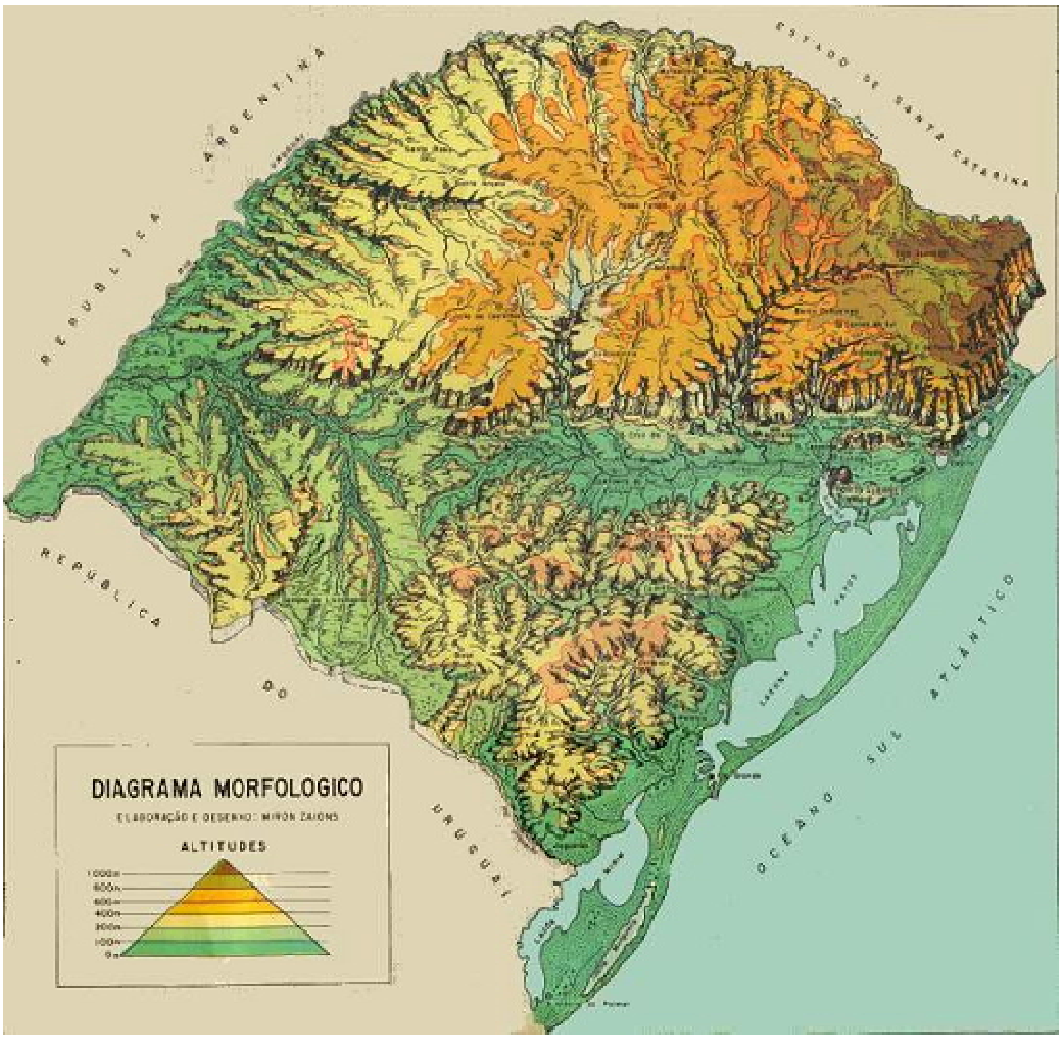

FIGURA 1- Mapa topográfico do Rio Grande do Sul. Fonte: http://www.paginadogaucho.com.br/ 23 de março de 2001

3.2.2 Período de estudo e dados utilizados

Para este trabalho foram utilizados dados mensais de temperatura média de 40 estações meteorológicas do Rio Grande do Sul (Tabela 1) para o período de 1913 a 2002, perfazendo um total de 90 anos de dados, referentes a três normais

climatológicas, o qual propiciou um bom período para o estudo climatológico proposto.

TABELA 1- Estações meteorológicas do Estado do Rio Grande do Sul, coordenadas geográficas e instituições a que pertencem

Estação	Latitude	Longitude	Altitude (m)	Fonte
1-Alegrete	29º41’	55º31’	124	8º DISME
2-Bagé	31º21’	54º06’	215	8º DISME
3-Bento Gonçalves	29º15’	51º31’	619	8º DISME
4-Bom Jesus	28º40’	50º26’	1047	8º DISME
5-Caçapava do Sul	30º30’	53º29’	450	8º DISME
6-Cachoeira do Sul	30º02’	52º53’	72	8º DISME
7-Caxias do Sul	29º10’	51º12’	785	8º DISME
8-Cruz Alta	28º38’	53º36’	472	8º DISME
9-Encruzilhada do Sul	30º32’	52º31’	427	8º DISME
10-Farroupilha	29º14’	51º26’	702	FEPAGRO
11-Guaporé	28º55’	51º54’	772	8º DISME
12-Ijuí	28º23’	53º54’	448	FEPAGRO
13-Iraí	27º11’	53º14’	222	8º DISME
14-Júlio de Castilhos	29º13’	53º40’	516	FEPAGRO
15-Lagoa Vermelha	28º25’	51º35’	836	8º DISME
16-Marcelino Ramos	27º27’	51º54’	414	8º DISME
17-Osório	29º41’	50º13’	32	FEPAGRO
18-Palmeira das Missões	27º53’	53º26’	634	8º DISME
19-Passo Fundo	28º15’	52º24’	676	8º DISME
20-Pelotas	31º52’	52º21’	13	8º DISME
21-Porto Alegre	30º01’	51º13’	46	8º DISME
22-Rio Grande	32º01’	52º05’	5	8º DISME
23-Santa Cruz do Sul	29º43’	52º25’	56	8º DISME
24-Santa Maria	29º42’	53º42’	95	8º DISME
25-Santa Rosa	27º51’	54º25’	360	8º DISME
26-Santa Vitória do Palmar	33º31’	53º21’	6	8º DISME
27-Santana do Livramento	30º53’	55º32’	210	8º DISME
28-Santiago	29º11’	54º53’	426	8 DISME
29-Santo Ângelo	28º18’	54º15’	289	8º DISME
30-São Borja	28º39’	56º00’	96	FEPAGRO
31-São Francisco de Paula	29º20'	51º30'	912	8º DISME
32-São Gabriel	30º20’	54º19’	124	8º DISME
33-São Luiz Gonzaga	28º23’	54º58’	254	8º DISME
34-Soledade	29º03’	52º26’	720	8º DISME
35-Tapes	30º52’	51º21’	5	8º DISME
36-Taquari	29º48’	51º49’	76	8º DISME
37-Torres	29º20’	49º43’	43	8º DISME
38-Uruguaiana	29º45’	57º05’	74	8º DISME
39-Vacaria	28º33’	50º42’	960	8º DISME
40-Veranópolis	28º56’	51º33’	705	FEPAGRO

Os dados das estações meteorológicas foram obtidos através do Instituto Nacional de Meteorologia - INMET (8º Distrito de Meteorologia - Porto Alegre) do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) e Fundação Estadual de Pesquisas Agropecuárias (FEPAGRO), da Secretaria de Ciência e Tecnologia (SCT) no Estado do Rio Grande do Sul, com séries já preenchidas as falhas

através do método das correlações utilizado por Diniz (2002). Observa-se que estes dados têm um período de observação muito grande e as estações meteorológicas bem distribuídas no Estado, conforme Figura 2. No Apêndice A, consta um mapa mostrando a localização das regiões ecoclimáticas.

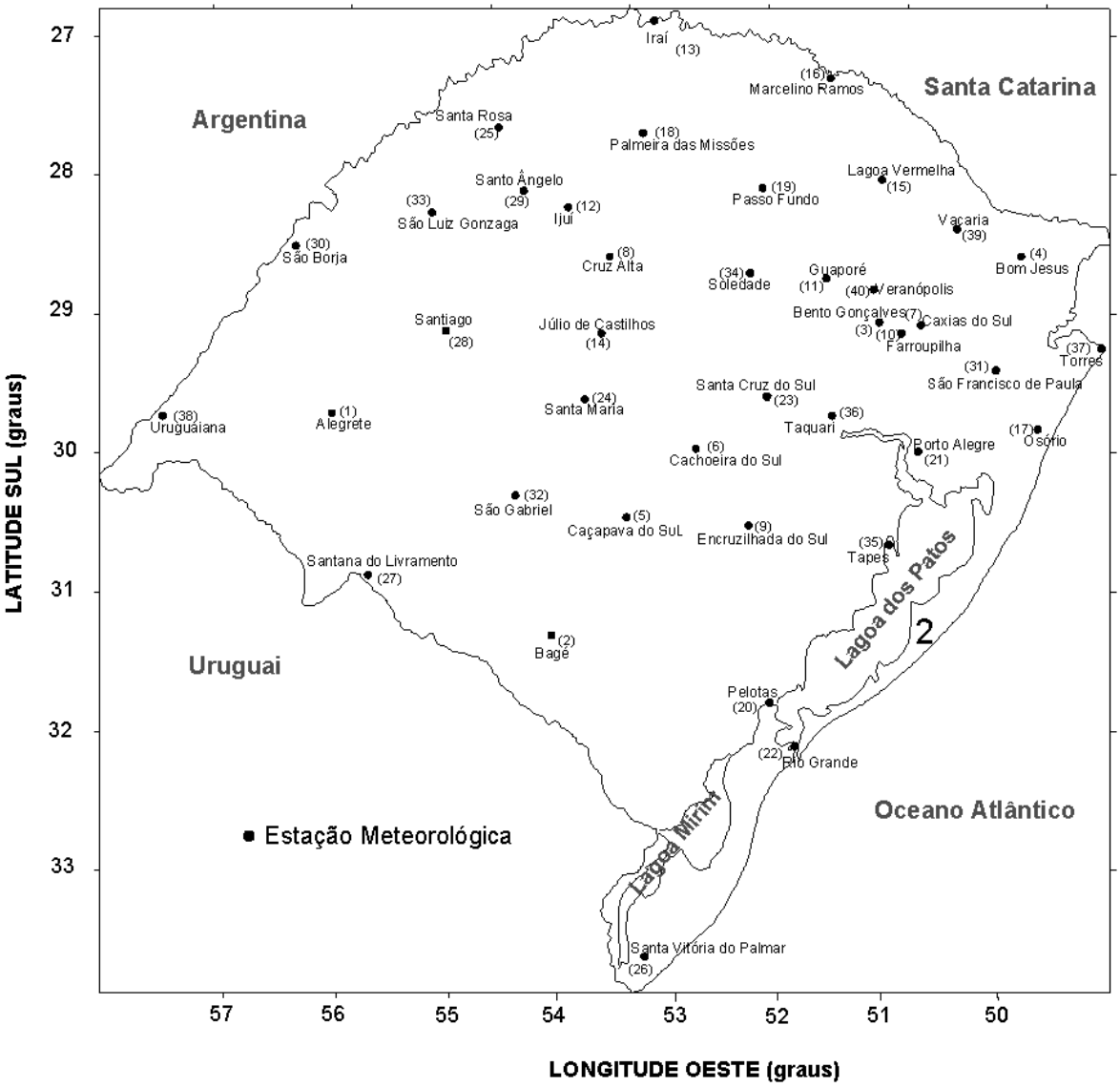

FIGURA 2- Distribuição espacial das estações meteorológicas do Estado do Rio Grande do Sul utilizadas no trabalho.

3.3 Metodologia

Quatro etapas foram realizadas neste trabalho que são descritas a seguir:

Na primeira foi feito o cálculo da temperatura média mensal, a qual foi obtida através da média aritmética simples entre as temperaturas máxima e mínima médias mensais para as 40 estações meteorológicas, após foi calculada a média trimestral, que foi usada como variável de agrupamento nos processos de classificação, dos 90 anos de dados para as 40 estações. Os trimestres foram os seguintes: (jan – fev – mar), (abr – mai – jun), (jul – ago – set) e (out – nov – dez). Não usou-se os trimestres astronômicos, que definem as estações do ano, por facilitar os cálculos com dados dentro do mesmo ano. Adotou-se, então, denominar como período quente o trimestre (JFM), temperado quente (OND), frio (JAS) e temperado frio (AMJ).

Na segunda etapa, foram determinadas as regiões homogêneas através de uma matriz de dados, a qual foi organizada da seguinte forma: os dados das temperaturas médias trimestrais foram ordenados conforme matrizes (1 e 2). A esta matriz de dados foram introduzidas também as coordenadas de posição de, cada estação meteorológica, o que exigiu a padronização destes dados conforme equação (4a).

Na terceira etapa foram aplicados quatro métodos hierárquicos aglomerativos de agrupamento à matriz de dados obtida na etapa anterior. Os métodos foram: o da ligação simples, o da ligação completa, o da centróide e o de Ward, tendo como função de agrupamento, a distância euclidiana; estes cálculos foram feitos através de “Software” estatístico. A utilização de mais de um método, e desses especificamente, deveu-se ao fato de que diferentes métodos aplicados ao mesmo conjunto de dados, geram diferentes estruturas. Assim, pôde-se verificar as composições dos grupos obtidos por cada um deles e escolher os grupos gerados pelo método que melhor refletisse as características climatológicas do Estado do Rio Grande do Sul. Tanto o número de grupos bem como as estações contidas em cada um deles, foram determinadas de forma subjetiva através de corte transversal no dendrograma.

Na quarta etapa, já definidas as regiões homogêneas (grupos) e separadas as estações constituintes de cada região, foram estabelecidas duas séries temporais de dados. A primeira série foi uma média regional constituída de valores de temperatura média mensal dos 90 anos das estações meteorológicas de cada região. A segunda série é também uma média regional, agora constituída de valores de temperatura média trimestral das estações meteorológicas de cada região homogênea para os 90 anos de dados.

Conhecendo-se as séries temporais dos dados de cada região homogênea, estabeleceu-se dois estudos climatológicos das mesmas. Na primeira série foi realizado o estudo de determinação do intervalo referente à normalidade através do método estatístico dos tercis, onde as séries dos 90 anos analisados foram ordenadas de forma crescente e subdivididas em três categorias com a mesma proporção de dados estabelecendo intervalos de: abaixo da normal, para os valores abaixo do limite inferior (1º tercil), normal para os valores entre o limite inferior e superior (2º tercil) e acima da normal para os valores acima do limite superior (3º tercil).

O segundo estudo, correspondente a segunda série de dados, foi feito através de parâmetros estatísticos como: média, desvio padrão, variância e coeficiente de variação.