UNIVERSIDADE ESTADUAL DE LONDRINA

ADENILSON APARECIDO DE OLIVEIRA

AVALIAÇÃO DA SITUAÇÃO EDUCACIONAL NO NORTE DO PARANÁ ATRAVÉS DA ANÁLISE DE AGRUPAMENTO

LONDRINA - PR

2011

ADENILSON APARECIDO DE OLIVEIRA

AVALIAÇÃO DA SITUAÇÃO EDUCACIONAL NO NORTE DO PARANÁ ATRAVÉS DA ANÁLISE DE AGRUPAMENTO

Monografia apresentada ao Curso de Especialização em Estatística com Ênfase em Educação, da Universidade Estadual de Londrina, como requisito parcial para a conclusão do curso. Orientadora:
Profª. Drª. Ana Vergínia Libos Messetti.

LONDRINA

2011

ADENILSON APARECIDO DE OLIVEIRA

AVALIAÇÃO DA SITUAÇÃO EDUCACIONAL NO NORTE DO PARANÁ ATRAVÉS DA ANÁLISE DE AGRUPAMENTO

Monografia apresentada ao Curso de Especialização em Estatística com Ênfase em Educação, da Universidade Estadual de Londrina, como requisito parcial para a conclusão do curso.

COMISSÃO EXAMINADORA

_____

Profa. Dra. Ana Vergínia Libos Messetti

Universidade Estadual de Londrina

_____

Profa. Dra. Jacinta Ludovico Zamboti

Universidade Estadual de Londrina

_____

Prof. Dr. José Carlos Dalmas

Universidade Estadual de Londrina

Londrina, 31 de março de 2011.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado vida e condições para a realização deste curso de especialização.

A minha família pelo apoio e estímulo.

Aos professores do curso de pós-graduação, que de alguma maneira contribuíram para a realização deste curso e da monografia.

E principalmente a minha orientadora, que foi mais que uma orientadora, foi uma amiga em todo o momento.

OLIVEIRA, Adenilson Aparecido. **Avaliação da situação educacional no Norte do Paraná através da análise de agrupamento.** 2011. Monografia (Especialização em Estatística com ênfase em Educação) – Universidade Estadual de Londrina.

RESUMO

O presente trabalho teve o objetivo de avaliar o ensino no norte do Paraná, e também analisar as cidades que apresentam melhores, ou piores notas, e se estas são obtidas devido a influência da região que estão localizadas. As cidades do norte do Paraná foram agrupadas utilizando-se as técnicas aglomerativas, Complete Linkage (vizinho mais longe) com a aplicação da distância Euclidiana quadrática e o Método de Ward com a aplicação da distância Euclidiana. O número de grupo final foi visualizado graficamente pelo Dendrograma, que mostrou claramente os agrupamentos, possibilitando detectar os grupos formados por cada método. Para melhor visualização dos agrupamentos, os grupos foram numerados, destacando-se nos mapas, onde foi possível fazer as comparações dos resultados. Para validar o resultado, realizou-se a comparação entre os dois métodos, verificando que as cidades continuaram se agrupando de formas similares.

Palavras Chaves: Avaliação educacional no Paraná, Análise de agrupamento, Método de Ward, Método do Vizinho mais longe.

OLIVEIRA, Adenilson Aparecido. **Evaluation of the educational situation in the North of Paraná through analysis of grouping**. 2011. Monograph (Specialization in statistics with emphasis in education) – Estate University of Londrina

ABSTRACT

This study aimed to evaluate teaching in northern Paraná, and also consider whether the cities that have better or worse grades, and these are obtained due to influences that are located in the region. The cities of northern Paraná were classified using the agglomerative techniques, Complete Linkage (neighbor along) with the application of the quadratic Euclidean distance and Ward method with the application of Euclidean distance. The final group number was displayed graphically by the dendrogram clearly showed that the grouping and the level of similarities between them, allowing to detect the groups formed by each method. For best viewing of the groups were the groups were numbered, especially on maps where it was possible to make comparisons of results. To validate the results was carried out to compare the two methods, noting that cities continued to grouping of similar forms

Key words: educational assessment in Paraná, collation, Analysis Method, Method of Ward's neighbor farther away.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Qualidade em educação é um dos temas mais abordados por mídias e governos. Como fazer para detectar pontos negativos na educação a serem trabalhados para que haja um melhor aprendizado, com tantas cidades em nosso estado.

Inicialmente, nota-se a importância de avaliar o ensino da região em que vivemos, para analisar se todas as cidades possuem o mesmo padrão de qualidade no ensino ou se há uma disparidade muito grande em relação ao nível de educação. Cidades que apresentam melhores notas, que estão se empenhando mais na educação do município, pode ser um atrativo para pessoas que desejam mudar para nossa região, procurando qualidade de vida para a família, incluindo boa qualidade de ensino para os filhos.

A técnica estatística denominada análise de agrupamento, vem como uma ferramenta que possibilita auxiliar os gestores governantes, pesquisadores e todos os interessados na educação, pois através de vários métodos de agrupamentos é possível reunir municípios com determinadas condições de igualdades em relação às notas dos alunos avaliados. Assim, grupos de cidades apresentam baixos ou altos valores nas médias de proficiência em língua portuguesa e matemática, auxiliando a visualização de possíveis influências da região ou detectam focos isolados de cidades com determinados problemas de ensino prejudicando a qualidade de ensino.

Frei (2006), em seu livro traz que reunir objetos (entende-se por objetos, seres humanos, animais, plantas, municípios, etc.) similares em determinados grupos é uma atividade humana importante e necessária, uma vez que essa atividade nos possibilita a organização dos grupos para um melhor estudo.

O presente trabalho teve o objetivo de agrupar e analisar as 92 cidades do norte do Paraná (norte velho como é conhecido) [1], através das suas notas de proficiências médias em língua portuguesa e matemática, padronizadas, obtidas na

[1] Dados retirados de ttp://www.setu.pr.gov.br/arquivos/Image/mapas/mapa_pr_regioes_turisticas_jpg.jpg

prova Brasil dos anos de 2005, 2007 e 2009. Notas estas que são utilizadas para a obtenção da nota do Ideb (Índice de Desenvolvimento da Educação Básica). As cidades serão agrupadas para que haja uma comparação e visualização do nível de ensino de nossa região.

O trabalho teve a seguinte composição: primeiramente foi realizado um levantamento histórico dos métodos de avaliação educacional voltados para o ensino médio e fundamental no Paraná e Brasil. (No segundo momento) foi realizado um levantamento de dados (especificamente no norte do Paraná) e por fim a aplicação dos métodos de Ward e do Vizinho mais Longe, com seus respectivos dendrograma para compor a discussão e conclusão final dos métodos hierárquicos adotados no estudo.

2 DESENVOLVIMENTO

2.1 Históricos da Avaliação Educacional no Ensino Fundamental e Médio no Brasil

Desde tempos primórdios os dirigentes brasileiros se preocupam com a educação no Brasil, mas é por volta do ano de 1937 que é criado o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), chamado inicialmente de Instituto Nacional de Pedagogia, recebendo o nome atual em 1972. Quando se transformou em um órgão autônomo em 1997 o Inep transformou em autarquia federal. Nas décadas anteriores à sua criação, algumas tentativas de sistematizar os conhecimentos educacionais e propor melhorias ao ensino já haviam sido articuladas, sem conseguir, no entanto, ter a continuidade desejada.

O Inep configurou-se, então, no primeiro órgão nacional a se estabelecer de forma duradoura como fonte primária de documentação e investigação, com atividades de intercâmbio e assistência técnica, como prescrevia a lei. Nos anos seguintes à sua criação, o Inep tornou-se uma referência para a questão educacional no País.

A avaliação educacional de âmbito federal implantada nos últimos tempos constituiu-se, em políticas de avaliação sistêmica a partir do final dos anos de 1980. Desse procedimento, a educação básica no Brasil passou a ser avaliada por um sistema nacional de avaliação em larga escala, com a finalidade de monitorar a qualidade do ensino por meio do *Sistema Nacional de Avaliação da Educação Básica* (SAEB), e foi aplicado pela primeira vez em 1990. Em 1995, o SAEB passou por uma reestruturação metodológica que possibilitou a comparação dos desempenhos ao longo dos anos. Desde a sua primeira avaliação, forneceu dados sobre a qualidade dos sistemas educacionais do Brasil como um todo, das regiões geográficas e das unidades federadas (estados e Distrito Federal).

O SAEB é realizado a cada dois anos e avalia uma amostra representativa dos alunos regularmente matriculados nas 4ª e 8ª séries do ensino fundamental e 3º

ano do ensino médio, de escolas públicas e privadas, localizadas em área urbana ou rural.

Em 2005 foi criada a prova Brasil que é uma prova censitária onde avalia todos os alunos de 4ª, 8ª série do ensino fundamental, oferecendo dados não apenas para o Brasil e unidades da Federação, mas também para cada município e escola participante. Como a metodologia dos dois sistemas de avaliação é igual, a partir de 2007 passaram a ser aplicadas em conjunto.

Em 2007 foi criado o Ideb (Índice de Desenvolvimento da Educação Básica), representa a iniciativa pioneira de reunir num só indicador dois conceitos igualmente importantes para a qualidade da educação: fluxo escolar e médio de desempenho nas avaliações.

2.2 Histórico da Avaliação Educacional no Ensino Fundamental e Médio no Paraná

Em 1988, a Secretaria de Estado da Educação do Paraná, realizou uma avaliação dos alunos de segunda e quarta séries, onde foram aplicadas provas específicas de Língua Portuguesa, Matemática, Ciências, e Estudos Sociais. As questões foram elaboradas a partir de itens produzidos por professores locais, com base nos guias curriculares vigentes.

Mas foi em 1995 que realizou sua primeira avaliação em larga escala, como expansão do SAEB, oferecendo as escolas e municípios resultados particularizados. Este processo teve seqüência nos anos subseqüentes. No ano 2000, introduziu-se nos anos finais do ensino fundamental esta mesma metodologia que até então era aplicada somente nos anos iniciais do ensino fundamental, tomando como referência o conteúdo do Currículo Básico para as Escolas Publicas do Paraná.

2.3 Análise Multivariada

A estatística multivariada segundo Reis (1997) teve seu início, como corpo teórico diferenciado no século 20, a partir de trabalhos de Pearson (1901) e Fisher (1928). Mas de certa forma até algumas décadas, a sua aplicação era bem reduzida, devido à dificuldade dos cálculos que envolvem os métodos multivariados. Mas com a acessibilidade dos computadores pessoais os métodos de estatística se popularizaram. Atualmente esta vem sendo utilizada com maior freqüência no nosso dia-a-dia, devido à evolução da tecnologia da computação, cada dia surge novos softwares computacionais adequado para os métodos utilizados.

De modo geral a estatística multivariada se divide principalmente formando dois grupos: para Mingoti (2005), um consistindo em técnicas exploratórias de sintetização da estrutura de variabilidade dos dados, e um segundo, consistidas em técnicas de inferência estatística. Dentre estes dois grupos podem-se listar algumas técnicas para o primeiro grupo: análise de componentes principais, análise de agrupamento, análise discriminante e análise de correspondência; para o segundo grupo se pode listar: métodos de estimação de parâmetros, testes de hipóteses, análise de variância, covariância e de regressão multivariada.

Para Hair (2005) análise multivariada refere-se “a todos os métodos estatísticos que simultaneamente analisam múltiplas medidas sobre cada indivíduo ou objeto de investigação e que qualquer análise simultânea de mais de duas variáveis é considerada análise multivariada”.

A estatística multivariada, segundo Mingoti (2005): “consiste em um conjunto de métodos estatísticos utilizados em situações nas quais várias variáveis são medidas simultaneamente”.

Em Reis (1997), para alguns autores, multivariado significa apenas examinar as relações entre duas ou mais variáveis, enquanto que para outros, o tema só se aplica quando se é possível pressupor que as variáveis seguem uma distribuição normal multivariada.

2.4 Análise de Agrupamento

A análise de agrupamento, também conhecida como análise de conglomerados, classificação ou cluster, tem como objetivo dividir os elementos da amostra, ou população, em grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si, (MINGOTI, 2005). Para a autora, a análise de agrupamento pode ser utilizada em várias situações, entre outras como: pesquisas de mercado, onde o determina saber o perfil de consumo. Na educação onde professores de uma determinada instituição de ensino podem ser avaliados pelos estudantes e agrupados de acordo com determinadas características.

Análise de agrupamentos é o nome dado a um conjunto de técnicas utilizadas na identificação de grupos homogêneos de casos. Artes; Barroso (2003), descrevem as seguintes etapas da aplicação desta técnica:

- Escolha do critério de parecença;
- Definição do número de grupos (a priori ou a posteriori);
- Formação dos grupos;
- Validação do agrupamento;
- Interpretação dos dados.

Para Reis (1997), os métodos de análise de clusters (ou de agrupamentos) são procedimentos de estatística multivariada que tentam organizar um conjunto de indivíduos, para os quais é conhecida informação detalhada, em grupos relativamente homogêneos.

> De modo sintético o método de agrupamento pode ser descrito como se segue: dado um conjunto de *n* indivíduos para os quais existe informação sobre a forma de *p* variáveis, o método de análise de cluster procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos. (REIS 1997).

2.4.1 Técnicas Hierárquicas

A técnica hierárquica subdividide-se em agrupamentos divisivos e aglomerativos. Nos hierárquicos aglomerativos, o processo se inicia com a matriz de similaridade, a qual é utilizada para identificar o par de indivíduos mais semelhantes entre si. Os dois indivíduos se agrupam e é considerado um único indivíduo. Em seguida, identifica-se o novo par mais semelhante, que formará outro grupo, e assim, novos grupos serão formados de acordo com suas similaridades até que todos estejam reunidos num único grupo.

Os algoritmos mais empregados na hierárquica aglomerativa, e apresentados em trabalhos são: método do vizinho mais próximo, método do vizinho mais distante, método das médias dos grupos, método dos centróides. Os hierárquicos divisívos, de maneira inversa, partem de um único grupo e finaliza com todos os indivíduos separadamente. (MESSETTI 2007)

Reis (1997) diz que este tipo de técnica baseia-se na construção de uma matriz de semelhanças ou diferenças em que cada elemento da matriz descreve o grau de semelhança ou diferença entre cada dois casos com base nas variáveis escolhidas.

> Métodos hierárquicos começam com uma matriz de distâncias entre objetos. Todos os objetos começam sozinhos em grupos de tamanho um, e os grupos que estão próximos se unem. Há várias maneiras de definir próximo. A mais simples é em termos de vizinhos mais próximos. Grupos são fundidos a um dado nível de distância se um dos objetos em um grupo está àquela distância ou mais próximo de pelo menos um objeto do segundo grupo. (MANLY, 2008)

As técnicas hierárquicas no início do processo de agrupamentos têm-se *n* grupos, onde cada elemento do conjunto de dados observado é considerado como sendo um grupo ou conglomerado isolado.

Na técnica hierárquica cada elemento constitui um agrupamento de tamanho um, logo vem à teoria de que se têm *n* grupos.

Em cada estágio do algoritmo, cada novo conglomerado formado é um agrupamento de conglomerados formados nos estágios anteriores. Se dois elementos amostrais aparecem juntos num mesmo cluster em algum estágio do

processo de agrupamento, eles aparecerão juntos em todos os estágios subseqüentes, ou seja, uma vez unidos estes elementos não poderão ser separados. (MINGOTI, 2005)

Devido esta propriedade de hierárquica é possível criar um Dendrograma, (gráfico vertical em que indica o nível de similaridade, ou dissimilaridade entre os grupos), que mostra o histórico de agrupamentos. A escolha da quantidade de grupo em que o conjunto de dados irá ser dividido é subjetiva, o ideal seria encontrar o número de partições que esteja associado à partição natural dos elementos agrupados. (MINGOTI, 2005)

2.4.1.1 Medidas de Similaridades e Dissimilaridades

O conceito fundamental na análise é escolher o critério que meça a distância entre os objetos em estudo, ou quantifique o quanto esses objetos são semelhantes ou dessemelhantes.

A medida de similaridade é definida como, quanto maior o valor observado, mais parecido são os objetos. A medida de dissimilaridade, quanto maior o valor observado menos parecido os objetos. Bussab (1990), o coeficiente de correlação é exemplo de medida de similaridade, e a distância Euclidiana é exemplo de medida de dissimilaridade.

Para Mingoti (2005) uma questão importante é quanto ao critério a ser utilizado para decidir até que ponto os elementos do conjunto de dados podem ser considerados como semelhantes entre si ou não, para isto é necessário considerar as medidas que descrevam a similaridade entre os elementos. Muitos algoritmos têm sido propostos para análise de agrupamentos, dentre este podemos citar: técnicas hierárquicas que produzem um dendograma que começa com o calculo da distancia de cada objeto a todos os outros objetos, e a técnicas de medidas de Similaridade.

Para que se possa proceder ao agrupamento de elementos é necessário que se decida a priori a medida de similaridade a ser utilizada, quanto menor o valor obtido mais similaridade haverá entre os elementos que estão sendo comparados.

Os métodos estatísticos procuram organizar os objetos em grupos homogêneos, aplicando para esta organização o conceito de similaridade. Para Frei (2006), a similaridade é obtida por meio de coeficientes, e a escolha deste coeficiente depende da escala de mensuração da variável.

Segundo Reis (1997) a relação de semelhança tem sido dominada pelos modelos geométricos, e estes modelos representam os objetos como ponto em um determinado espaço de coordenadas de forma que as dessemelhanças entre os objetos correspondam a distâncias métricas entre os respectivos pontos. Os métodos de classificação dos índices de semelhança exigem que se respeitem as propriedades métricas que são: simetria; desigualdade triangular; diferenciabilidade de não idênticos; indiferenciabilidade de idênticos.

Existem várias medidas apropriadas e cada uma com um jeito de formar um determinado tipo de agrupamento. As medidas apropriadas para variáveis quantitativas também são ditas de dissimilaridade, quanto menor for o valor obtido, mais similar vão ser os objetos que estão sendo estudados. (MINGOTI 2005)

> É necessário avaliar a vantagens e desvantagens de cada método medida e quais critérios e condições satisfazem. Existem muitas maneiras de definir o conceito de similaridade entre pares de objetos, cada uma enfatizando um aspecto diferente do conjunto de dados representativos desses objetos. O que constitui a similaridade total, ou a dissimilaridade total, de dois objetos depende do coeficiente adotado. (NETO 2007)

2.4.2 Algoritmos Hierárquicos

Num primeiro momento é determinada uma matriz de similaridade ou dissimilaridade. Esta matriz é definida pelo cálculo da distância e pelo algoritmo estabelecidos. Barroso e Artes (2003) descrevem alguns algoritmos a serem utilizados na análise, tais como: Método do Vizinho mais próximo (Single Linkage), Método do Vizinho mais longe (Complete Linkage), e Método de Ward (Ward´s Method).

2.4.2.1 Vizinho mais Próximo

Este método denominado como método de ligação simples, consiste no procedimento de procura de dois objetos mais similares entre si na matriz de distância. Depois disto, é analisado cada conjunto desses objetos formados, procurando novamente os dois conjuntos mais próximos, ou seja, que tenham distâncias menores.

De acordo com Mingoti (2005), em cada estágio do processo de agrupamentos os dois conglomerados que são mais similares em relação à distância, são combinados em um único cluster.

2.4.2.2 Vizinho mais Distante

Este método denominado como método de ligação completa, pois após agrupar os dois vizinhos de menor distância, verifica-se a distância máxima deste primeiro grupo para os demais objetos restantes, procurando garantir com que os objetos de um grupo guardem a máxima distância de outros grupos. (FREI, 2006)

Reis (1997) define que o procedimento é inverso ao anterior (vizinho mais próximo), uma vez que a distância entre dois grupos agora é definida como sendo a distância entre seus elementos mais afastados ou menos semelhantes. Este método tem tendências para encontrar grupos compactos compostos de indivíduos muito semelhantes entre si.

2.4.2.3 Método de Ward

Em 1963, Ward propôs um método de agrupamento que é fundamentado na mudança de variação entre os grupos e dentro dos grupos que estão sendo formados em cada passo do agrupamento. Ele segue o princípio que inicialmente

cada elemento é considerado como um único conglomerado e que em cada passo do algoritmo de agrupamento calcula-se a soma de quadrados dentro de cada grupo. (MINGOTI, 2005).

Para Reis (1997) este método pode ser resumido nas seguintes etapas: primeiro são calculadas as médias das variáveis para cada grupo; em seguida é calculado o quadrado da distância Euclidiana entre essas médias; somam-se as distâncias para todos os indivíduos; e por último procura-se otimizar a variância mínima dentro dos grupos.

Segundo Messetti (2007) a distância entre dois agrupamentos é a soma dos quadrados entre dos dois agrupamentos feita sobre todas variáveis. Em cada estágio do procedimento de agrupamento, a soma interna de quadrados é minimizada sobre todas as partições, que podem ser obtidas pela combinação de dois agregados do estágio anterior. Este procedimento tende a combinar agrupamentos com um pequeno número de observações e tende a produzir agregados com aproximadamente o mesmo número de observações.

2.5 Material

Para a realização desse trabalho foi feito um levantamento de todas as cidades do Paraná e também do indicador de proficiências na prova Brasil aplicada na oitava série (ou nono ano como é conhecida no novo sistema de ensino). Informações estas que tornam possíveis aos gestores governamentais uma visão de conjunto das unidades de ensino, e os eventuais problemas de aprendizagem, também podem servir de parâmetros para a escolha de melhores estratégias de qualificação.

Serão analisados os indicadores desde o ano de 2005, ano em que começou a aplicação da prova Brasil, e como ela é aplicada nos anos ímpares tem-se os anos de 2007 e 2009, onde foi possível analisar através da analise de agrupamento e pelo método hierárquico, quais cidades se destacaram na nota de proficiência em língua portuguesa e matemática, tanto em notas altas como em notas mais baixas, e

também a possível interferência da cidade de Londrina no grupo de cidades com notas mais altas.

2.6 Cidades do Norte do Paraná

No Paraná hoje existe 399 municípios, divididas em várias mesorregiões, adotou-se trabalhar com o norte velho, que é composto por 92 cidades denominadas:

Abatia, Alvorada do Sul, Andirá, Apucarana, Arapongas, Arapuã, Ariranha do Ivaí, Assaí, Bandeirantes, Barra do Jacaré, Bela Vista do Paraíso, Bom Sucesso Borrazópolis, Cafeara, Califórnia, Cambará, Cambé, Cambira, Carlópolis, Centenário do Sul, Congonhinhas, Conselheiro Mairinck, Cornélio Procópio, Cruzmaltina, Faxinal, Figueira, Florestópolis, Godoy Moreira, Grandes Rios, Guapirama, Guaraci, Ibaiti, Ibiporã, Itambaracá, Ivaiporã, Jaboti, Jacarezinho, Jaguapitã, Jandaia do Sul, Japira, Jardim Alegre, Jataizinho, Joaquim Távora, Jundiaí do Sul, Kaloré, Leópolis, Lidianópolis, Londrina, Lunardelli, Lupionópolis, Marilandia do Sul, Marumbi, Miraselva, Nova América da Colina, Nova Fátima, Nova Santa Bárbara, Novo Itacolomi, Pinhalão, Pintangueiras, Porecatú, Prado Ferreira, Primeiro de Maio, Quatiguá, Rancho Alegre, Ribeirão Claro, Ribeirão do Pinhal, Rio Bom, Rio Branco do Ivaí, Rolandia, Rosario do Ivaí, Sabaudia, Salto do Itararé, Santa Amélia, Santa Cecília do Pavão, Santa Mariana, Santana do Itararé, Santo Antonio da Platina, Santo Antonio do Paraíso, São Jerônimo da Serra, São João do Ivaí, São José da Boa Vista, São Pedro do Ivaí, São Sebastião da Amoreira, Sapopema, Sertaneja, Sertanópolis, Siqueira Campos, Tamarana, Tomazina, Uraí, Wenceslau Braz. (anexo 1)

2.7 Cálculo da Média de Proficiência em Língua Portuguesa e Matemática[2]

A nota do Ideb (Índice de Desenvolvimento da Educação Básica) é obtida por duas notas, ou seja, a média de proficiência em língua portuguesa e matemática padronizada e o indicador de rendimento escolar. A nota que se utilizou para a realização dos agrupamentos é a média de proficiências em língua portuguesa e matemática. O processo para a obtenção desses valores está a seguir: (Notas obtidas estão na Tabela do Anexo 1).

A média de Proficiência (N_{ji}) é um indicador padronizado para notas de zero a dez, dos alunos da unidade j, obtida em determinada edição do exame realizado ao final da etapa de ensino. É obtida a partir das médias de português e matemática dos estudantes submetidos à determinada edição do exame realizado ao final da etapa educacional considerada (Prova Brasil ou SAEB).

A N_{ij} é obtida de acordo com:

$N_{ji} = \frac{n_{ji}^{lp} + n_{ji}^{mat}}{2}$ e $n_{ji}^{\alpha} = \frac{S_{ji}^{\alpha} - S_{\inf}^{\alpha}}{S_{\sup}^{\alpha} - S_{\inf}^{\alpha}} * 10$, em que:

n_{ji}^{α} = Proficiência na disciplina, obtida pela unidade j, no ano i, padronizada para valores entre zero e 10;

α = disciplina (matemática ou língua portuguesa);

S_{ji}^{α} = Proficiência média (em língua portuguesa ou matemática), não padronizada, dos alunos da unidade j obtida no exame do ano i;

$S_{\inf}^{\alpha}$ = Limite inferior da média de proficiência (língua portuguesa e matemática) do SAEB 1997;

$S_{\sup}^{\alpha}$ = Limite superior da média de proficiência (língua portuguesa e matemática) do SAEB 1997.

---

[2] Dados retirado de http://www.inep.gov.br/download/Ideb/Nota_Tecnica_n1_concepcaoIDEB.pdf- 04/01/11

Para as unidades escolares (ou redes) que obteve $S_{ji}^{\alpha} < S_{\inf}^{\alpha}$, a proficiência média é fixada em $S_{\inf}^{\alpha}$. Por sua vez, aquelas unidades que obtiveram $S_{ji}^{\alpha} > S_{\sup}^{\alpha}$ tem-se o desempenho fixado em $S_{\sup}^{\alpha}$.

A Tabela 1 apresenta a média e o desvio padrão das proficiências dos alunos da 4ª e 8ª série do ensino fundamental e da 3ª série do ensino médio no SAEB de 1997. A Tabela 2 traz os valores dos limites inferiores e superiores utilizados na padronização das proficiências médias em língua portuguesa e matemática dos alunos da 4ª e 8ª séries do ensino fundamental e 3ª série do ensino médio.

Tabela 1 – Saeb 1997: Proficiências médias e desvio padrão

Série	Matemática		Língua Portuguesa	
	Média	Desvio Padrão	Média	Desvio Padrão
4ª do EF	190.8	44	186.5	46
8ª do EF	250.0	50	250.0	50
3ª do EM	288.7	59	283.9	56

Fonte: Saeb 1997 – Inep/MEC

A partir da média e desvio padrão das proficiências no SAEB 1997 (ano em que a escala do Saeb foi definida), calcularam-se, para cada etapa de ensino, considerando as diferentes disciplinas avaliadas no exame, os limites, inferior e superior, de acordo com:

$S_{\inf}^{\alpha} = média_{\alpha} - (3 * DP)$ e $S_{\sup}^{\alpha} = média_{\alpha} + (3 * DP)$.

Tabela 2 – Limite superior e inferior das proficiências

Série	Matemática		Língua Portuguesa	
	S_{inf}	S_{sup}	S_{inf}	S_{sup}
4ª do EF	60	322	49	324
8ª do EF	100	400	100	400
3ª do EM	111	467	117	451

Fonte: Saeb 1997 – Inep/MEC

Esses limites, inferiores e superiores, apresentados na Tabela 2, são usados para calcular todos os Ideb's, ou seja, desde 1997, a partir do SAEB, para o Brasil

(rede privada e pública; urbanas e rurais), e para os dados agregados por unidade da federação e, a partir da Prova Brasil de 2005, para municípios (rede municipal e estadual) e para as escolas.

2.8 Metodologia

As técnicas hierárquicas parte do princípio que no início de um agrupamento tem-se *n* conglomerados, onde cada elemento do conjunto analisado é considerado como um conglomerado único, já no último estágio tem-se um único conglomerado constituído de todos os elementos do conjunto.

2.8.1 Metodologia de Aplicação das Técnicas Hierárquicas Aglomerativa

Segundo Mingoti (2005) os principais passos para a aplicação dessa técnica podem ser resumidos da seguinte forma:

- Cada elemento constitui um cluster de tamanho um.
- Em cada estágio de agrupamento, os pares similares que são combinados passam a formar um único conglomerado dessa forma em cada estágio do processo o número de conglomerado vai diminuindo.

2.8.1.1 Propriedade de hierarquia

Cada novo conglomerado formado é um agrupamento de conglomerado formado nos estágios anteriores, portanto dois elementos aparecem juntos e num mesmo cluster terão que aparecerem em todos os demais subseqüentes, pois uma vez unidos esses elementos não poderão ser separados.

Devido à propriedade de hierarquia é possível construir um gráfico chamado dendograma que represente a historia de agrupamento. O dendograma é um gráfico em forma de “arvore” onde a escala vertical inicial o nível de similaridade ou dissimilaridade, e na horizontal os elementos amostrais.

Os métodos de hierarquia aqui utilizados foram o método de ligação completa, conhecido como vizinho mais distante e o método de Ward.

2.8.1.2 Aplicação das medidas de similaridades

A análise de agrupamento também chamada de análise de cluster tem como objetivo dividir os elementos de um determinado grupo, em outros novos grupos, fazendo com que os elementos desse novo grupo sejam o mais similar entre si possível.

Para este processo é importante já estar definido até que ponto dois elementos ou mais do conjunto de dados são considerados semelhantes ou não, para isto será utilizada medidas de similaridades ou dissimilaridade.

Existem várias opções de medidas para demonstrar similaridade. Nesta pesquisa foram utilizadas a distância euclidiana e a distância euclidiana quadrática.

2.8.1.2.1 Distância Euclidiana

A distância euclidiana, que define a distancia entre dois elementos χ_l e χ_k, com $l \neq k$.

É definida por: $d(\chi_l, \chi_k) = [(\chi_l - \chi_k)*(\chi_l - \chi_k)]^{1/2} = [\sum_{i=l}^{p} (\chi_{il} - \chi_{ik})^2]^{1/2}$.

2.8.1.2.2 Distância Euclidiana Quadrática

A distância euclidiana quadrática, que é a distancia entre dois casos (*i* e *j*), e é definida como o somatório dos quadrados das diferenças entre os valores de *i* e *j* para todas as variáveis (k= 1, 2,..., p). $d_{ij}^2 = \sum_{k=1}^{p} (X_{il} - X_{ik})^2$, Onde dois elementos são comparados em cada variável pertencentes ao vetor de observações.

2.8.2 Aplicação do Algoritmo

Depois de aplicado a medida de similaridade da distância euclidiana, e distância euclidiana quadrática esses dados das distâncias entre os elementos amostrais são armazenadas numa matriz de dimensão nxn , chamada de matriz de distância, na qual d_{ij} representa a distância do elemento amostral (i) ao elemento amostral (j).

$$D_{4x4} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{21} & 0 & d_{23} & d_{24} \\ d_{31} & d_{32} & 0 & d_{34} \\ d_{41} & d_{42} & d_{43} & 0 \end{bmatrix}$$

Onde o zero representa a distância entre o próprio elemento.

2.8.2.1 Método de Ward

O Método de Ward é calculado utilizando as seguintes formulas: a primeira é que calcula a soma de quadrados dentro de cada conglomerado.

$$SS_i = \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})'(X_{ij} - \overline{X}_{i.})$$

Onde, n_i é o número de elementos no conglomerado C_i quando se está no passo k do processo de agrupamento, X_{ij} é o vetor de observações do j-ésimo elemento amostral que pertence ao i-ésimo conglomerado, $\overline{X}_{i.}$ é o centróide do conglomerado C_i, e SS_i representa a soma de quadrados correspondente ao conglomerado Ci.

No passo k, a soma de quadrados total dentro dos grupos é definida como:

$$SSR = \sum_{i=1}^{g_k} SS_i$$

Onde g_k é o número de grupos existentes quando se está no passo k.

A distância entre os conglomerados entre os clusters C_l e C_i é então definida como:

$d(C_l, C_i) = [\frac{n_l n_i}{n_l + n_i}](\overline{X}_l - \overline{X}_{i.})'(\overline{X}_l - \overline{X}_{i.})$, que é a soma de quadrados entre os clusters C_l e C_i. No método de Ward as comparações de conglomerados que têm tamanhos diferentes sofrem uma penalização representada pelo fator de ponderação $\frac{n_i n_l}{n_i + n_l}$. Quanto maior forem os valores de n_i e n_l e a discrepância entre eles, maior será o valor do fator de penalização, aumentando, assim, a distância entre os centróides dos conglomerados.

O método de Ward tende a produzir grupos com aproximadamente o mesmo número de elementos e tem como base principal os princípios de análise de variância.

2.8.2.2 Método de Ligação Completa

Para Reis (1997), no método de ligação completa a distância entre dois grupos é definida como sendo a distância entre seus elementos mais afastados ou menos semelhantes.

Dados dois grupos (i, j) e (k) a distância entre eles, e a maior das distâncias entre os seus elementos: $d_{(i,j)k} = \max\{d_{ik}; d_{jk}\}$.

De acordo com esta estratégia cada grupo passa a ser definido como um conjunto de elementos em que cada um é mais semelhante a todos os restantes elementos do grupo do que a qualquer dos elementos dos restantes grupos.

Este método tem tendência para encontrar clusters compactados compostos de indivíduos muito semelhantes entre si.

2.8.3 Definição do número de grupos

O número de grupos pode ser definido “a priori” quando se tem algum conhecimento a respeito dos dados, ou pode ser definido “a posteriori” com base nos resultados da análise. Como critério para definir o número de grupos foi utilizado o dendrograma.

Os resultados dos algoritmos apresentados da técnica hierárquica aglomerativa se combinam até que seja estabelecido um diagrama de árvore denominado dendrograma, no qual no eixo das abscissas se posicionam os indivíduos e no eixo das ordenadas, as medidas obtidas após aplicação da metodologia, sendo possível desta forma visualizar a forma como será feita a divisão dos grupos. (MINGOTI 2005)

2.9 Aplicações dos Algoritmos

A aplicação dos dados no Software R, foram adotados dois métodos: Método de Ward com Distância Euclidiana e o Método do Vizinho mais Longe com Distância Euclidiana Quadrática. A seguir os resultados das duas técnicas utilizadas no estudo.

.

2.9.1 Método de Ward Distância Euclidiana

A aplicação do método de Ward com a distância Euclidiana, nos dados (Anexo um) utilizando o programa R, o qual resultou no dendrograma, que pode ser visualizado na Figura 1. Por meio do dendrograma foi feito o corte de distribuição dos grupos e estes representados nas tabelas abaixo.

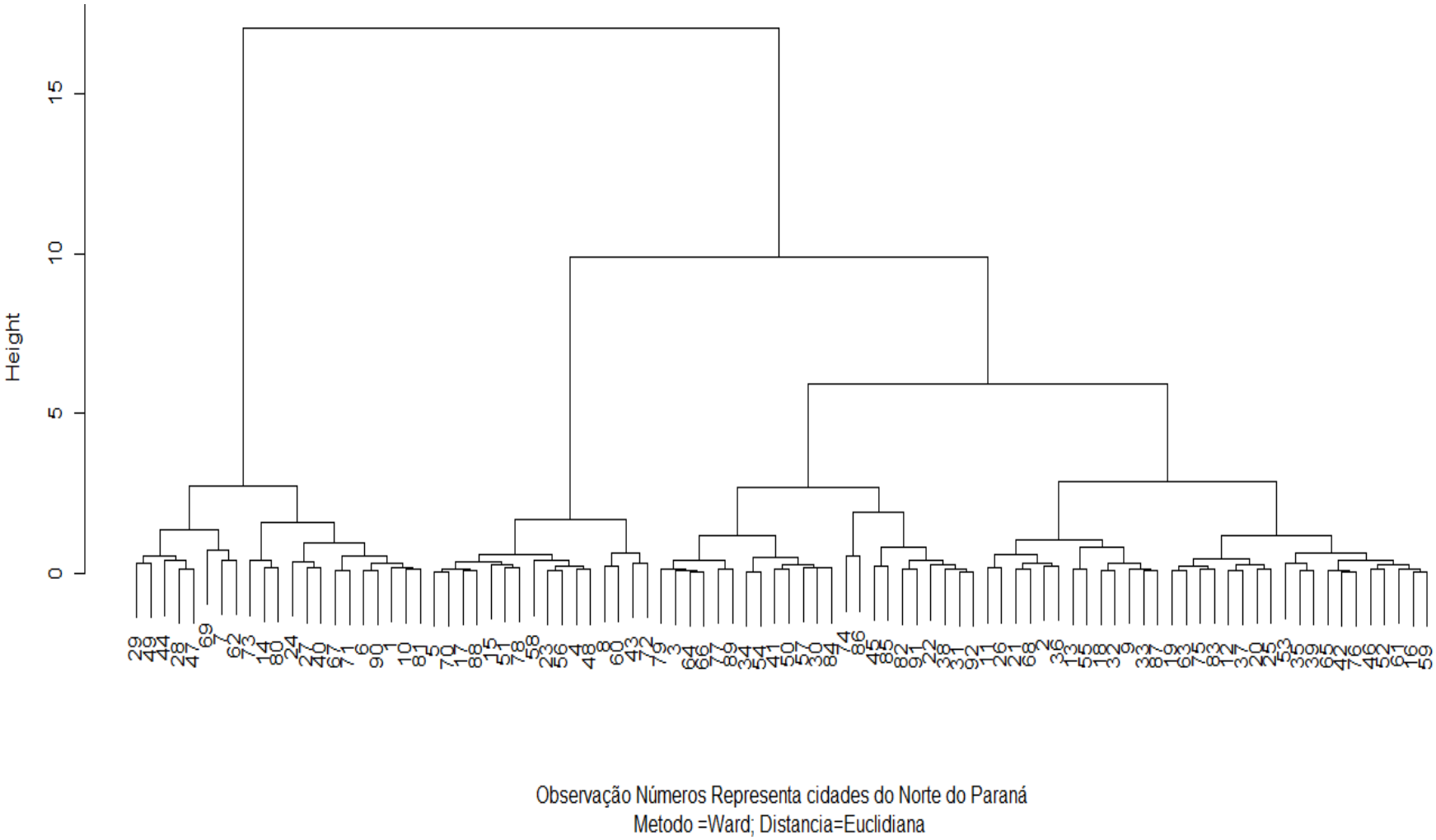


Figura 1 – Dendrograma das 92 cidades do Norte do Paraná, pelo método de Ward e Distância Euclidiana

Tabelas das 92 cidades pertencentes ao norte do Paraná, agrupadas pelo método de Ward com Distância Euclidiana

Tabela 3 – Código e Nome das Cidades pertencentes ao GRUPO 1

CÓDIGO	NOME DA CIDADE	CÓDIGO	NOME DA CIDADE
29	GRANDE RIOS	24	CRUZ MALTINA
49	LUNARDELLI	27	FLORESTÓPOLIS
44	JUNDIAÍ DO SUL	40	JAPIRA
28	GODOY MOREIRA	67	RIBEIRÃO DO PINHAL
47	LIDIANÓPOLIS	71	ROSÁRIO DO IVAÍ
69	RIO BRANCO DO IVAÍ	06	ARAPUÃ
07	ARIRANHA DO IVAÍ	90	TOMAZINA
62	PRADO FERREIRA	01	ABATIÁ
73	SALTO DO ITARARÉ	10	BARRA DO JACARÉ
14	CAFEARA	81	SÃO JOÃO DO IVAÍ
80	SÃO JERONIMO DA SERRA		

Tabela 4 - Código e Nome das Cidades pertencentes ao GRUPO 2

CÓDIGO	NOME DA CIDADE	CÓDIGO	NOME DA CIDADE
05	ARAPONGAS	23	CORNÉLIO PROCÓPIO
70	ROLANDIA	56	NOVA FATIMA
17	CAMBÉ	04	APUCARANA
88	SIQUEIRA CAMPOS	48	LONDRINA
15	CALIFORNIA	08	ASSAÍ
51	MARILANDIA DO SUL	60	PITANGUEIRAS
78	SANTO ANTONIO DA PLATINA	43	JOAQUIM TÁVORA
58	NOVO ITACOLOMI	72	SABAUDIA

Tabela 5 – Código e Nome das Cidades pertencentes ao GRUPO 3

CÓDIGO	CIDADE	CÓDIGO	CIDADE
79	SANTO ANTONIO DO PARAISO	57	NOVA SANTA BARBARA
03	ANDIRÁ	30	GUAPIRAMA
64	QUATIGUÁ	84	SÃO SEBASTIÃO DA AMOREIRA

CÓDIGO	CIDADE	CÓDIGO	CIDADE
66	RIBEIRÃO CLARO	74	SANTA AMÉLIA
77	SANTANA DO ITARARÉ	86	SERTANEJA
89	TAMARANA	45	KALORE
34	ITAMBARACÁ	85	SAPOPEMA
54	MIRASELVA	82	SÃO JOSÉ DA BOA VISTA
41	JARDIM ALEGRE	91	URAÍ
50	LUPIONÓPOLIS	22	CONSELHEIRO MAIRINCK
92	WENCESLAU BRAZ	38	JAGUAPITÃ
31	GUARACI		

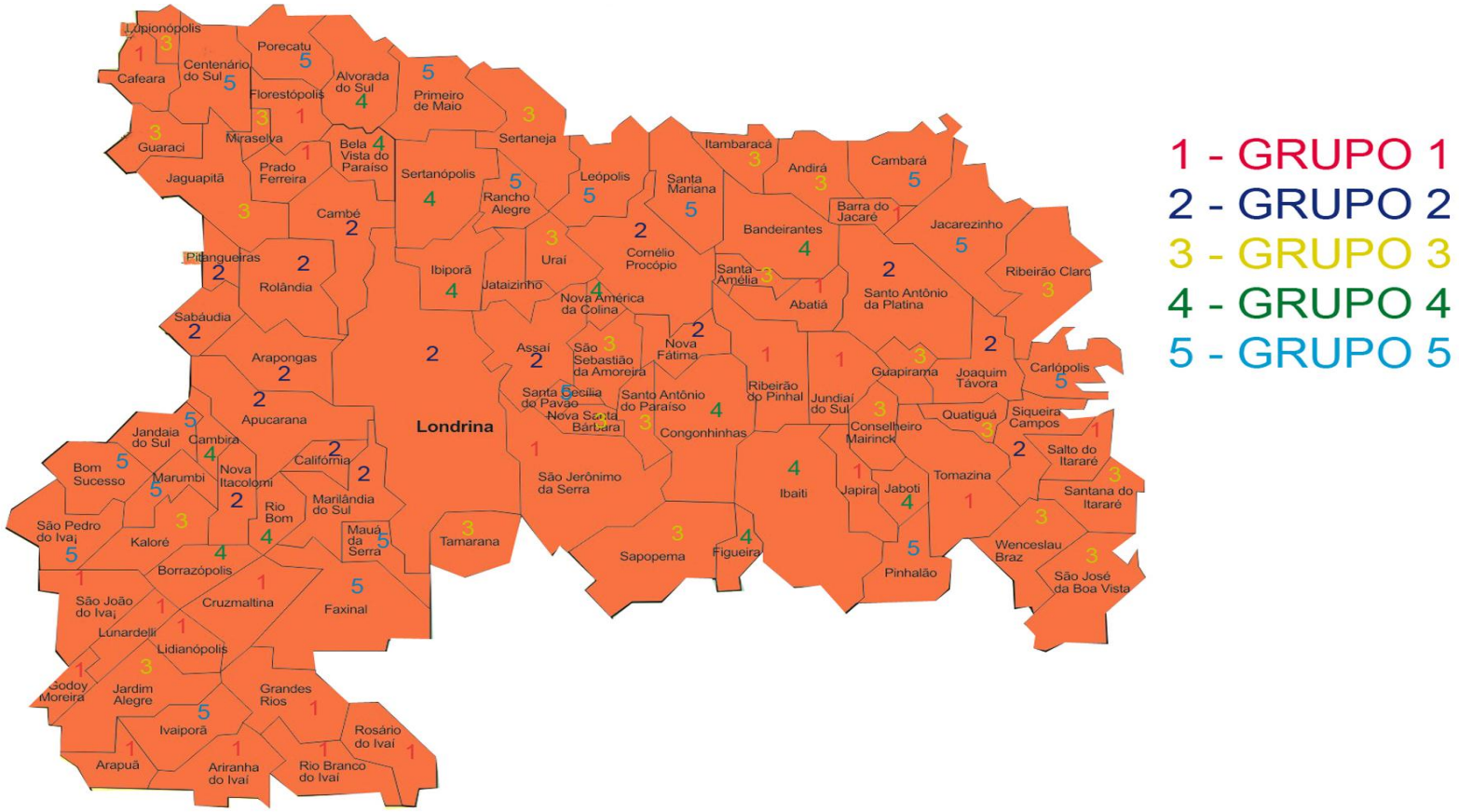
Tabela 6 – Código e Nomes das Cidades pertencentes ao GRUPO 4

CÓDIGO	CIDADE	CÓDIGO	CIDADE
11	BELA VISTA DO PARAISO	55	NOVA AMÉRICA DA COLINA
26	FIGUEIRA	18	CAMBIRA
21	CONGONHINHAS	32	IBAITI
68	RIO BOM	09	BANDEIRANTES
02	ALVORADA DO SUL	33	IBIPORÃ
36	JABOTI	87	SERTANÓPOLIS
13	BORRAZÓPOLIS		

Tabela 7 - Código e Nomes das Cidades pertencentes ao GRUPO 5

CÓDIGO	CIDADE	CÓDIGO	CIDADE
19	CARLÓPOLIS	53	MAUÁ DA SERRA
63	PRIMEIRO DE MAIO	35	IVAIPORÃ
75	SANTA CECÍLIA DO PAVÃO	39	JANDAIA DO SUL
83	SÃO PEDRO DO IVAÍ	65	RANCHO ALEGRE
12	BOM SUCESSO	42	JATAIZINHO
37	JACAREZINHO	76	SANTA MARIANA
20	CENTENÁRIO DO SUL	46	LEÓPOLIS
25	FAXINAL	52	MARUMBI
59	PINHALÃO	61	PORECATU
16	CAMBARÁ		

A aplicação do Método de Ward com Distância Euclidiana possibilitou subdividir o norte do Paraná em cinco grupos, com semelhanças aproximadas em relação às médias das três notas do IDEB (2005, 2007, 2009). Dois grupos merecem maiores destaques: o grupo um e grupo dois. O grupo um, foram agrupadas as cidades que apresentaram as menores médias nos três anos consecutivos, resultando em médias entre 3,79 a 4,39. As cidades que compõem o grupo dois, são cidades que apresentaram as maiores médias nos três anos, resultando em médias entre 4,90 a 5,40. As demais cidades apresentaram as médias intermediárias, resultando médias entre 4,39 e 4,90, portanto os grupos foram caracterizados pelas baixas, médias e altas médias.



Fonte: http://www.setu.pr.gov.br/arquivos/Image/mapas/mapa_pr_regioes_turisticas_jpg.jpg

Figura 2 – Mapa representando os grupos pela distribuição do Método de Ward e Distância Euclidiana

2.9.2 Método do Vizinho mais longe pela Distância Euclidiana Quadrática

Este método analisa o agrupamento levando em consideração a maior distância entre cada grupo, assim será possível fazer a comparação entre os dois métodos e verificar se as notas altas e baixas estão posicionadas em determinada região ou não há interferência em determinada região do Paraná. O gráfico dendrograma auxiliou para uma melhor visualização dos resultados.

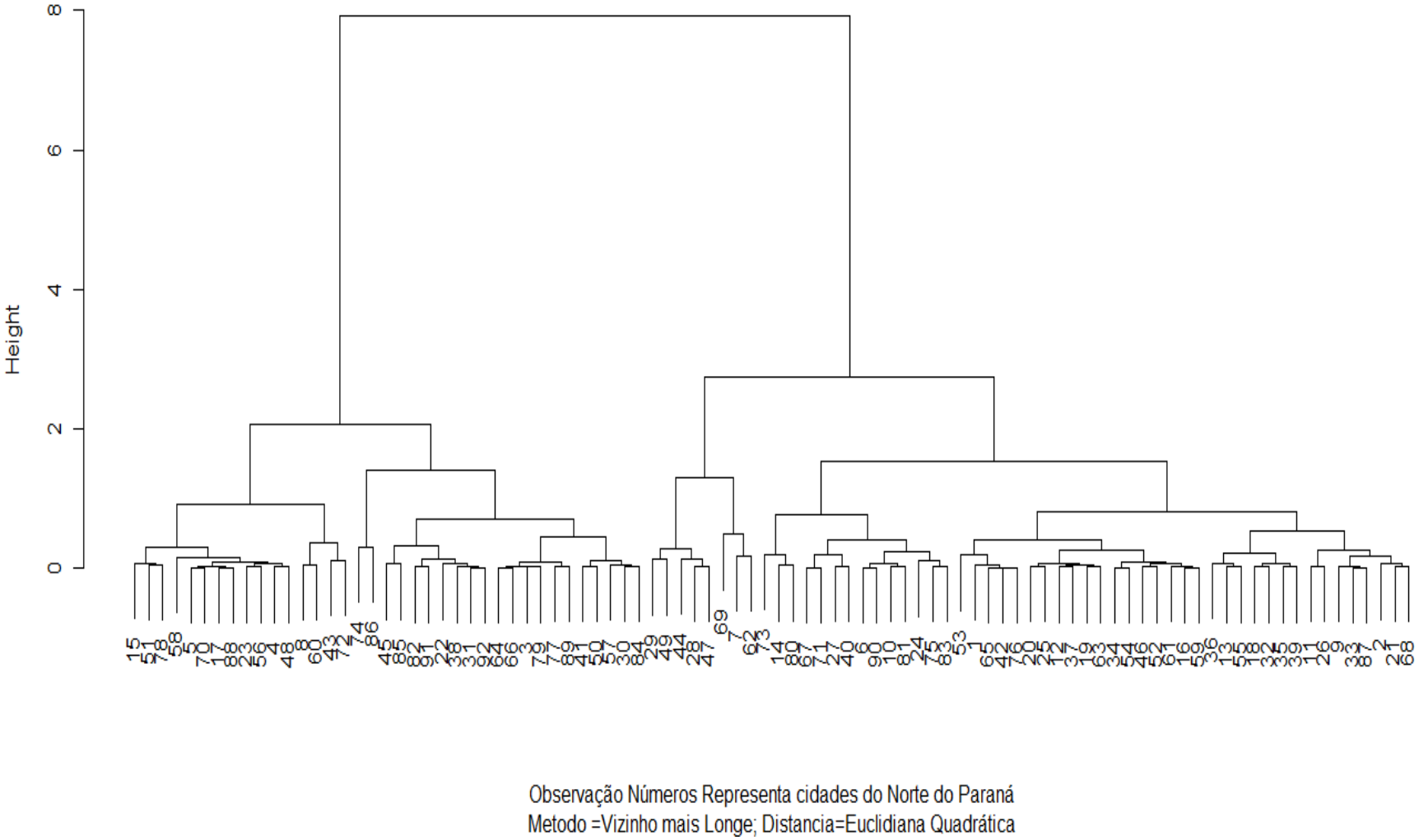


Figura 3 – Dendrograma das 92 cidades do norte do Paraná, pelo método do Vizinho mais Longe e Distância Euclidiana Quadrática

Tabelas das 92 cidades pertencentes ao norte do Paraná, agrupadas pelo método do Vizinho mais Longe utilizando a Distância Euclidiana Quadrática.

Tabela 8 - Código e Nomes das Cidades pertencentes ao GRUPO 1

CÓDIGO	CIDADES	CÓDIGO	CIDADES
15	CALIFORNIA	23	CORNÉLIO PROCÓPIO
51	MARILANDIA DO SUL	56	NOVA FATIMA
78	SANTO ANTONIO DA PLATINA	04	APUCARANA
58	NOVO ITACOLOMI	48	LONDRINA
05	ARAPONGAS	08	ASSAÍ
70	ROLANDIA	60	PITANGUEIRAS
17	CAMBÉ	43	JOAQUIM TÁVORA
88	SIQUEIRA CAMPOS	72	SABAUDIA

Tabela 9 – Código e Nomes das Cidades pertencentes ao GRUPO 2

CÓDIGO	CIDADES	CÓDIGO	CIDADES
74	SANTA AMÉLIA	66	RIBEIRÃO CLARO
86	SERTANEJA	03	ANDIRÁ
45	KALORE	79	SANTO ANTONIO DO PARAISO
85	SAPOPEMA	77	SANTANA DO ITARARÉ
82	SÃO JOSÉ DA BOA VISTA	89	TAMARANA
91	URAÍ	41	JARDIM ALEGRE
22	CONSELHEIRO MAIRINCK	50	LUPIONÓPOLIS
38	JAGUAPITÃ	57	NOVA SANTA BARBARA
31	GUARACI	30	GUAPIRAMA
92	WENCESLAU BRAZ	84	SÃO SEBASTIÃO DA AMOREIRA
64	QUATIGUA		

Tabela 10 - Códigos e Nomes das Cidades pertencente ao GRUPO 3

CÓDIGO	CIDADE	CÓDIGO	CIDADE
29	GRADES RIOS	47	LIDIANÓPOLIS
49	LUNARDELLI	69	RIO BRANCO DO IVAÍ
44	JUNDIAÍ DO SUL	07	ARIRANHA DO IVAÍ
28	GODOY MOREIRA	62	PRADO FERREIRA

Tabela 11 – Códigos e Nomes das Cidades pertencentes ao GRUPO 4

CÓDIGO	CIDADE	CÓDIGO	CIDADE
73	SALTO DO ITARARÉ	06	ARAPUÃ
14	CAFEARA	90	TOMAZINA
80	SÃO JERONIMO DA SERRA	10	BARRADO JACARÉ
67	RIBEIRÃO DO PINHAL	81	SÃO JOÃO DO IVAÍ
71	ROSÁRIO DO IVAÍ	24	CRUZ MALTINA
27	FLORESTÓPOLIS	75	SANTA CECÍLIA DO PAVÃO
40	JAPIRA	83	SÃO PEDRO DO IVAÍ

Tabela 12 – Código e Nomes das Cidades pertencentes ao GRUPO 5

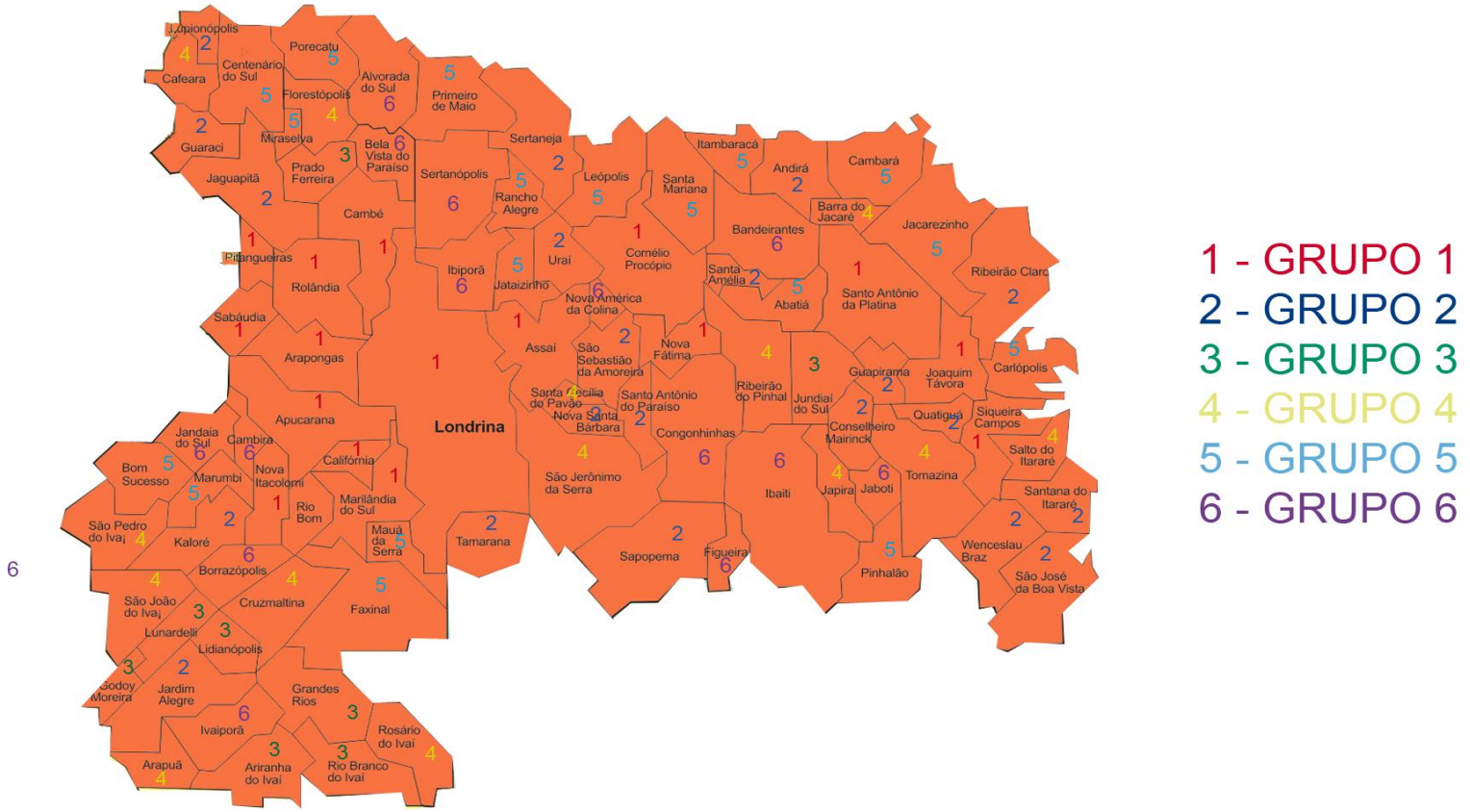
CÓDIGO	CIDADE	CÓDIGO	CIDADE
53	MAUÁ DA SERRA	63	PRIMEIRO DE MAIO
01	ABATIÁ	34	ITAMBARACÁ
65	RANCHO ALEGRE	54	MIRASELVA
42	JATAIZINHO	46	LEÓPOLIS
76	SANTA MARIANA	52	MARUMBI
20	CENTENÁRIO DO SUL	61	PORECATU
25	FAXINAL	16	CAMBARÁ
12	BOM SUCESSO	59	PINHALÃO
37	JACAREZINHO	19	CARLÓPOLIS

Tabela 13 – Código e nomes das Cidades pertencentes ao GRUPO 6

CÓDIGO	CIDADE	CÓDIGO	CIDADE
36	JABOTI	26	FIGUEIRA
13	BORRAZÓPOLIS	09	BANDEIRANTES
55	NOVA AMÉRICA DA COLINA	33	IBIPORÃ
18	CAMBIRA	87	SERTANÓPOLIS
32	IBAITI	02	ALVORADA DO SUL
35	IVAIPORÃ	21	CONGONHINHAS
39	JANDAI DO SUL	68	RIO BOM
11	BELA VISTA DO PARAIZO		

Através do método do Vizinho Mais Longe com Distância Euclidiana Quadrática, foi possível dividir as cidades que compõe o norte do Paraná em seis

grandes grupos com médias dos três anos do IDEB (2005, 2007, 2009). Com a aplicação deste método os grupos que mereceram maiores destaques, foram os grupos um e três. As cidades que compõem o grupo um foram as cidades que apresentaram maiores médias dos três anos, com valores médios entre 4,90 e 5,42. As cidades do grupo três foram agrupadas devido aos mais baixos valores médios dos três anos, permanecendo entre 3,79 a 4,16. Os demais grupos, considerados os intermediários, foram agrupados com notas próximas entre si com valores médios entre 4,16 e 4,90.



Fonte: http://www.setu.pr.gov.br/arquivos/Image/mapas/mapa_pr_regioes_turisticas_jpg.jpg

Figura 4 – Mapa representando os grupos pela distribuição do método do Vizinho mais Longe com Distância Euclidiana Quadrática

2.10 Discussão final

Analisando os dois métodos aplicados, tanto o Método do Vizinho mais Longe com Distância Euclidiana Quadrática e o Método de Ward com Distância Euclidiana, procuraram de certa forma distribuir as cidades do norte do Paraná em grupos que apresentaram as médias dos três anos muito próximas entre si.

Os dois Métodos apresentaram quase a mesma estrutura de distribuição, sendo diferenciado por algumas cidades que mudaram de grupo ou formaram grupos separados. As cidades de Itamaracá e Mirasselva que pertencia ao grupo três do método de Ward, se agruparam com as cidades do grupo cinco do método do Vizinho mais Longe. As cidades de Grandes Rios, Lunardelli, Jundiaí do Sul, Godoy Moreira, Lidianópolis, Rio Branco do Ivaí, Ariranha do Ivaí e Prado Ferreira, que no método de Ward compunham com outras cidades o grupo um, no método do Vizinho mais Longe se separaram e formaram o grupo três isoladamente, as demais cidades que compunham o grupo um no método de Ward formaram o grupo quatro no método do Vizinho mais Longe, somado com duas novas cidades: Santa Cecília do Pavão e São Pedro do Ivaí, que pertenciam ao grupo cinco no método de Ward. A cidade de Abatiá que pertencia ao grupo um no método de Ward passou a pertencer ao grupo cinco no método do Vizinho mais Longe. O grupo dois que apresentava as melhores médias no Método de Ward permaneceram as mesmas, sem se deslocarem do grupo um no Método do Vizinho mais Longe.

O Método que apresentou a melhor distribuição dos grupos foi o método do Vizinho mais Longe, que apesar da pouca diferença do resultado final do agrupamento, concluiu-se que o método do Vizinho mais Longe fez uma distribuição de médias mais homogêneas em relação ao outro método.

3 CONCLUSÃO

A análise de agrupamento é importante por permitir que seja possível agrupar objetos que sozinhos e isolados talvez não fosse possível de se analisar, possibilitando as descobertas de focos semelhantes sendo possível de intervir. Os dois métodos utilizados no trabalho (Método de Ward com Distância Euclidiana e o Método do Vizinho mais Longe com Distância Euclidiana Quadrática), procuraram agrupar as cidades com médias dos três anos mais homogêneas entre si, porém o método do Vizinho mais Longe, teve uma melhor distribuição de agrupamentos, principalmente com as cidades que apresentaram médias dos três anos mais baixas, facilitando a análise.

Pode se concluir através das análises realizadas com os dois métodos, que as cidades com as melhores notas, na sua maioria, possuem indústrias, não dependendo somente da agricultura, possuindo um melhor Índice de Desenvolvimento Humano (IDH), e conseqüentemente, esta melhoria é repassada para a educação, em forma de melhores qualidades de ensino as crianças e aos jovens. As cidades que apresentaram notas mais baixas, são cidades menores e dependentes da agricultura, possuem recursos econômicos mais escassos, tendo um repasse e investimento menor na educação, fazendo com que os jovens tenham que na maioria das vezes abandonarem os estudos para ajudarem as famílias, comprometendo os estudos destes. Através deste trabalho governos e dirigentes que queiram melhor homogeneidade do ensino fundamental no norte do Paraná, poderá posicionar-se da situação dos grupos para tentar intervir e melhorar a educação no estado.

REFERÊNCIAS

BARROSO, L.P.; ARTES, R. **Análise multivariada**: Minicurso do 10 Simpósio de Estatística Aplicada à Experimentação Agronômica – RBRAS, 48 Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria- SEAGRO. Lavras: UFLA, 2003

FREI, F. **Introdução à análise de agrupamento**: teoria e prática. São Paulo: UNESP, 2006

GATTI, Bernadete A. Avaliação Educacional no Brasil: pontuando uma História de ações. EccoS revista cientifica, junho ano/vol4, numero 001. Centro Universitário Nove de Julho. SP, Brasil disponível: <http://redalyc.uaemex.mx/pdf/715/71540102.pdf>. Acesso 16 de agos. 2010

HAIR, Joseph F.; ANDERSON, Rolph E.; TATHAN, Ronald L.; BLACK, William C. **Análise multivariada de dados**. Porto Alegre: Bookman, 2005

HISTÓRIA DA AVALIAÇÃO – disponível: <http://www.google.com.br/search?q=hist%C3%B3ria+da+avalia%C3%A7%C3%A3o+educacional+no+brasil&hl=pt-BR&client=firefoxa&sa=X&rls=org.mozilla:PTBR:official&tbs=tl:1,tll:1980,tlh:1989&ei=pz1gTMKJM4L-bc0YG5DQ&ved=0CF0QyQEoBg>. Acesso 04de agos. 2010

INEP – disponível: <http://www.inep.gov.br/institucional/historia.htm>. Acesso 04 de agos. 2010

MANLY, Bryan J. F. **Métodos estatísticos multivariados.** Porto Alegre: Bookman, 2008

MESSETTI, A.V.L. **Utilização de técnicas multivariadas na avaliação da divergência genética de girassol.** Faculdade de Ciências Agronômicas da Universidade Estadual Paulista - Campus de Botucatu - Área de Concentração em Energia na Agricultura. 2007

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte: UFMG, 2005

PROVA BRASIL- disponível: <http://provabrasil.inep.gov.br/index.php?option=com_content&task=view&id=15&Itemid=14>. Acesso 09 de agos. 2010

REIS, E. **Estatística Multivariada Aplicada**. Lisboa: Sílabo. 1997

ANEXOS

Anexo 1 – Quadro das Cidades do Norte do Paraná com as notas médias de proficiência em matemática e português dos anos de 2005, 2007, 2009

CÓDIGO	CIDADES	2005	2007	2009	MÉDIA
1	ABATIA	4,14	4,28	4,62	4,35
2	ALVORADA DO SUL	4,95	4,65	4,51	4,70
3	ANDIRA	4,41	5,00	4,75	4,72
4	APUCARANA	5,01	4,98	5,26	5,08
5	ARAPONGAS	4,92	4,87	5,06	4,95
6	ARAPUA	4,14	4,27	4,30	4,23
7	ARIRANHA DO IVAI	3,91	3,61	4,94	4,16
8	ASSAI	5,03	5,27	5,03	5,11
9	BANDEIRANTES	4,67	4,68	4,65	4,67
10	BARRA DO JACARE	4,32	4,23	4,48	4,34
11	BELA VISTA DO PARAISO	4,72	4,67	4,26	4,55
12	BOM SUCESSO	4,59	4,57	4,52	4,56
13	BORRAZOPOLIS	4,77	4,42	4,75	4,65
14	CAFEARA	3,98	4,70	4,49	4,39
15	CALIFORNIA	5,12	4,88	5,05	5,02
16	CAMBARA	4,40	4,47	4,56	4,48
17	CAMBE	4,79	4,93	5,08	4,93
18	CAMBIRA	4,60	4,62	4,79	4,67
19	CARLOPOLIS	4,49	4,55	4,40	4,48
20	CENTENARIO DO SUL	4,59	4,32	4,48	4,46
21	CONGONHINHAS	4,75	4,59	4,50	4,61
22	CONSELHEIRO MAIRINCK	4,60	4,54	5,17	4,77
23	CORNELIO PROCOPIO	4,83	4,85	5,27	4,98
24	CRUZMALTINA	4,56	4,09	4,31	4,32
25	FAXINAL	4,54	4,38	4,61	4,51
26	FIGUEIRA	4,59	4,81	4,28	4,56
27	FLORESTOPOLIS	4,24	3,98	4,08	4,10

CÓDIGO	CIDADES	2005	2007	2009	MÉDIA
28	GODOY MOREIRA	3,78	3,96	4,13	3,96
29	GRANDES RIOS	3,97	3,83	4,25	4,02
30	GUAPIRAMA	4,36	4,75	5,00	4,70
31	GUARACI	4,75	4,74	5,17	4,89
32	IBAITI	4,59	4,65	4,91	4,71
33	IBIPORA	4,65	4,84	4,71	4,73
34	ITAMBARACA	4,39	4,69	4,67	4,58
35	IVAIPORA	4,49	4,52	4,79	4,60
36	JABOTI	4,86	4,60	4,73	4,73
37	JACAREZINHO	4,54	4,53	4,63	4,57
38	JAGUAPITA	4,76	4,66	5,25	4,89
39	JANDAIA DO SUL	4,51	4,40	4,83	4,58
40	JAPIRA	4,30	4,11	4,18	4,20
41	JARDIM ALEGRE	4,16	4,76	4,89	4,60
42	JATAIZINHO	4,31	4,30	4,62	4,41
43	JOAQUIM TAVORA	5,19	5,49	5,57	5,42
44	JUNDIAI DO SUL	3,67	3,86	3,85	3,79
45	KALORE	4,39	4,67	5,25	4,77
46	LEOPOLIS	4,37	4,50	4,73	4,53
47	LIDIANOPOLIS	3,81	4,09	4,09	4,00
48	LONDRINA	4,98	5,01	5,13	5,04
49	LUNARDELLI	3,90	3,58	4,01	3,83
50	LUPIONOPOLIS	4,28	4,81	4,77	4,62
51	MARILANDIA DO SUL	4,97	5,07	4,97	5,00
52	MARUMBI	4,25	4,44	4,75	4,48
53	MAUA DA SERRA	4,40	4,28	4,97	4,55
54	MIRASELVA	4,31	4,70	4,68	4,56
55	NOVA AMERICA DA COLINA	4,87	4,32	4,75	4,65
56	NOVA FATIMA	4,87	4,96	5,23	5,02
57	NOVA SANTABARBARA	4,20	4,63	5,05	4,63
58	NOVO ITACOLOMI	4,68	5,14	5,22	5,01

CÓDIGO	CIDADES	2005	2007	2009	MÉDIA
59	PINHALAO	4,46	4,45	4,60	4,50
60	PITANGUEIRAS	5,06	5,44	5,17	5,23
61	PORECATU	4,31	4,43	4,61	4,45
62	PRADO FERREIRA	4,02	3,62	4,53	4,06
63	PRIMEIRO DE MAIO	4,41	4,46	4,45	4,44
64	QUATIGUA	4,35	4,93	4,77	4,68
65	RANCHO ALEGRE	4,35	4,28	4,69	4,44
66	RIBEIRAO CLARO	4,41	4,89	4,81	4,70
67	RIBEIRAO DO PINHAL	3,92	4,13	4,30	4,12
68	RIO BOM	4,73	4,74	4,45	4,64
69	RIO BRANCO DO IVAI	3,51	4,00	4,52	4,01
70	ROLANDIA	4,95	4,92	5,05	4,97
71	ROSARIO DO IVAI	3,97	4,16	4,39	4,17
72	SABAUDIA	5,18	5,17	5,43	5,26
73	SALTO DO ITARARE	3,83	4,53	4,12	4,16
74	SANTA AMELIA	3,89	5,36	5,58	4,94
75	SANTA CECILIA DO PAVAO	4,43	4,38	4,29	4,37
76	SANTA MARIANA	4,26	4,29	4,66	4,40
77	SANTANA DO ITARARE	4,58	5,00	4,68	4,75
78	SANTO ANTONIO DA PLATINA	4,95	4,91	4,84	4,90
79	SANTO ANTONIO DO PARAISO	4,37	4,93	4,66	4,65
80	SAO JERONIMO DA SERRA	3,94	4,69	4,30	4,31
81	SAO JOAO DO IVAI	4,17	4,21	4,46	4,28
82	SAO JOSE DA BOA VISTA	4,62	4,87	5,00	4,83
83	SAO PEDRO DO IVAI	4,43	4,32	4,43	4,39
84	SAO SEBASTIAO DA AMOREIRA	4,30	4,63	4,88	4,60
85	SAPOPEMA	4,34	4,79	5,45	4,86
86	SERTANEJA	4,24	5,27	5,17	4,89

CÓDIGO	CIDADES	2005	2007	2009	MÉDIA
87	SERTANOPOLIS	4,71	4,75	4,76	4,74
88	SIQUEIRA CAMPOS	4,88	4,98	5,05	4,97
89	TAMARANA	4,56	5,12	4,79	4,83
90	TOMAZINA	4,15	4,33	4,38	4,28
91	URAI	4,69	4,76	5,00	4,82
92	WENCESLAU BRAZ	4,68	4,70	5,14	4,84

Anexo 2 – Programas do R

Pacote Rcmdr

> local({pkg <- select.list(sort(.packages(all.available = TRUE))) if(nchar(pkg)) library(pkg, character.only=TRUE)})

Carregando pacotes exigidos: tcltk

Loading Tcl/Tk interface ... done

Carregando pacotes exigidos: car

Carregando pacotes exigidos: MASS

Carregando pacotes exigidos: nnet

Carregando pacotes exigidos: survival

Carregando pacotes exigidos: splines

Versão do Rcmdr 1.6-0

Anexando pacote: 'Rcmdr'

The following object(s) are masked from package:tcltk : tclvalue

Carregando pacotes exigidos: RODBC

norte <- sqlQuery(channel = 1, select * from [Plan2$])

Método de Ward Distancia Euclidiana

Agrupamentos <- hclust(dist(model.matrix(~-1 + F2+F3+F4, norte)) , method="ward")

plot(Agrupamentos, main= "Agrupamento das cidades do norte do Paraná ", xlab="Observação Números Representa cidades do Norte do Paraná", sub="Metodo =Ward; Distancia=Euclidiana ")

Método do Vizinho mais Longe Distancia Euclidiana Quadrática

Agrupamentos <- hclust(dist(model.matrix(~-1 + F2+F3+F4, norte))^2 , method="complete")

```
plot(Agrupamentos, main= "Agrupamento das cidades do norte do Paraná ", xlab="Observação Números Representa cidades do Norte do Paraná", sub="Metodo =Vizinho mais Longe; Distancia=Euclidiana Quadrática")
```