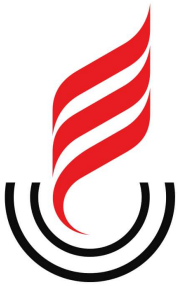

Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Aline Carla da Silva

AVALIAÇÃO DO IDEB MUNICIPAL DE CAMPINA GRANDE POR UM MODELO MULTINÍVEL BAYESIANO

Campina Grande
Abril de 2016.

Aline Carla da Silva

AVALIAÇÃO DO IDEB MUNICIPAL DE CAMPINA GRANDE POR UM MODELO MULTINÍVEL BAYESIANO

Monografia apresentada ao Curso de Especialização em Estatística Aplicada do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de especialista em Estatística.

Orientador:

Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros

Campina Grande

Abril de 2016.

S586a Silva, Aline Carla da

Avaliação do IDEB municipal de Campina Grande por um modelo multinível Bayesiano [manuscrito] / Aline Carla da Silva. - 2016.

65 p. : il. color.

Digitado.

Monografia (Estatística Aplicada) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2016.

"Orientação: Prof. Drº Kleber Napoleão Nunes de Oliveira Barros, Departamento de Estatistica".

1. IDEB. 2. Análise de Agrupamento. 3. Regressão Multinível. 4. Inferência Bayesiana. I. Título.

21. ed. CDD 519

Aline Carla da Silva

AVALIAÇÃO DO IDEB MUNICIPAL DE CAMPINA GRANDE POR UM MODELO MULTINÍVEL BAYESIANO

Monografia apresentada ao Curso de Especialização em Estatística Aplicada do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de especialista em Estatística.

Aprovado em: 20 / 04 / 2016

Banca Examinadora

Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros
Orientador

Prof. Dr. Mácio Augusto de Albuquerque

Prof. Dr. Silvio Fernando Alves Xavier Júnior

Dedico à minha família
e amigos.

Resumo

A partir dos dados oficiais retirados do site do Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP), que diz respeito ao Índice de Desenvolvimento da Educação Básica (IDEB) de escolas municipais de Campina Grande em 2013, realizamos uma Análise de Agrupamento para separar as observações com características semelhantes e dividi-las em grupos e uma Análise de Regressão Linear Múltipla com a técnica *Stepwise* para a seleção de variáveis significativas do modelo. Em cada grupo distinto foi selecionado um modelo de Regressão Multinível com Inferência Bayesiana nos parâmetros utilizando o método *Backward*, a fim de estimar parâmetros e encontrar um melhor modelo, para cada grupo, que aumente o valor do IDEB. Para isto, foram utilizados os Softwares RStudio e OpenBUGS. A partir dos resultados obtidos observa-se que um modelo com 6 parâmetros é suficiente para alcançar o objetivo em questão.

Palavras-chave: IDEB, Análise de Agrupamento, Regressão Multinível, Inferência Bayesiana.

Abstract

By the official data taken from the site of the National Institute for Research and Studies Anísio Teixeira (INEP), as regards to the Index of Development of Basic Education (IDEB) of municipal schools in Campina Grande in 2013, we conducted a cluster analysis to separate observations with similar characteristics and divide them into groups, a multiple linear regression analysis with *Stepwise* technique to select the significance variables in the model. In each distinct group was selected a Multilevel Regression Model with Bayesian Methods in parameters using a *Backward* method, in order to estimate the parameters and find a better model for each group, who increases the value of IDEB. For this purpose, has been used the Softwares RStudio and OpenBugs. Results showed that a model with 6 parameters is sufficient to achieve the goal in question.

Keywords: IDEB, Cluster Analysis, Multilevel Regression, Bayesian Methods.

Sumário

1 Introdução

A partir do final do século XX, as políticas de avaliação externa da educação começaram a ganhar destaque no cenário nacional e internacional. No Brasil, desde então, as políticas públicas federais passaram a enfatizar a aplicação, aos alunos, de testes padronizados nacionalmente, compreendida como instrumento adequado para se conhecer e promover uma educação de qualidade no país (NASCIMENTO; SILVA, 2014).

O estabelecimento do Índice de Desenvolvimento da Educação Básica (IDEB) consiste em um dos eixos centrais do Plano de Desenvolvimento da Educação (PDE). Sua apresentação à sociedade brasileira ocorreu em abril de 2007 e deu-se no contexto do lançamento do Plano de Aceleração do Crescimento (PAC). O índice é medido a cada dois anos e objetiva que, o país, a partir do alcance das metas municipais e estaduais, obtenha nota 6 no ano de 2022, que corresponde à qualidade do ensino em países desenvolvidos (SILVA, 2009).

Tomando como população as escolas municipais de Campina Grande em 2013 que possuiam o 5^o ano e estavam aptas a realizar o IDEB, objetivamos estimar os parâmetros que aumentem o valor desse índice, levando em consideração que esses parâmetros são variáveis contidas dentro de todas as unidades observadas (escolas), objetiva-se também levar os resultados obtidos para as unidades observadas para que o modelo encontrado seja posto em prática e o IDEB seja melhorado efetivamente.

No decorrer do trabalho, definiremos alguns conceitos das avaliações educacionais utilizadas à nível nacional mostrando também como é feito o cálculo do IDEB, faremos as revisões das técnicas de Regressão Linear Múltipla, Multinível e Bayesiana, Análise Multivariada e alguns métodos de escolha de modelos, necessárias para o estudo em questão; apresentaremos a metodologia empregada no estudo e mostraremos os resultados de todas as aplicações feitas no RStudio e OpenBUGS ambos para sitema operacional Windows além das discussões pertinentes; por fim apresentaremos as devidas conclusões para o estudo.

2 Revisão de literatura

Primeiramente vamos entender alguns conceitos sobre a educação à nível nacional e algumas especificações da cidade de Campina Grande e rever algumas definições de métodos e técnicas necessárias para a análise.

2.1 Avaliação e Mensuração de Escolas

No decorrer dos anos de 1990, o tema da qualidade da educação passou a ganhar importância na sociedade brasileira. A noção de qualidade da educação, segundo Gusmão (2010), apresenta diferentes significados, sendo defendidas duas visões distintas de qualidade. Uma que valoriza fundamentalmente a aprendizagem, principalmente a medida em testes padronizados, que juntamente com o acesso e a permanência formaria o tripé da qualidade da educação e outra que prioriza uma noção de qualidade mais ampla, ao enfatizar os processos e outros elementos da aprendizagem que não se restrinjam aos conteúdos disciplinares.

As políticas educacionais, implantadas no país a partir dos anos 1990, adotaram a primeira perspectiva de qualidade, o que impulsionou a criação de iniciativas de avaliação externa da educação. Atualmente, é tão importante medir a qualidade do ensino, que um aluno ao decorrer do Ensino Fundamental, passa por no mínimo 4 avaliações de porte nacional.

2.1.1 Avaliações Nacionais da Educação

A partir de 1995, segundo Freitas (2004), a qualidade de educação passa a ser objeto de regulação federal, cuja viabilidade exigira o aporte de um sistema de informações educacionais conjugado a um sistema nacional de avaliação, considerados ambos elementos estratégicos da boa-governança educacional no país. A LDB/1996 (Lei de Diretrizes e Bases da Educação - define e regulariza a organização da educação brasileira com base nos princípios presentes na Contituição Federal) define, no artigo 9^o, que caberá à União

coletar, analisar e disseminar informações sobre a educação, inciso V, e assegurar processo nacional de avaliação do rendimento escolar no ensino fundamental, médio e superior, em colaboração com os sistemas de ensino, objetivando a definição de prioridades e a melhoria da qualidade do ensino, inciso VI (BRASIL, 1996).

A partir de então, o governo federal inicia a implantação de um modelo padronizado de avaliação em larga escala, apresentado como mecanismo capaz de monitorar a qualidade da educação brasileira em todos os níveis e de contribuir para a sua elevação. Para o estabelecimento dessa proposta, o Inep/MEC, de acordo com Freitas (2004), teve de enfrentar o desafio de fazer com que diferentes atores, sobretudo equipes das secretarias estaduais e municipais de ensino, professores e gestores escolares, incorporassem a avaliação externa em seu cotidiano, percebendo-a como apoio para a melhoria da qualidade do ensino.

A criação de um sistema de avaliação e a geração de informações sobre o sistema escolar nacional partem do pressuposto de que levantar e tornar público informações sobre o desempenho dos sistemas de escolares, contribuiria para a melhoria da qualidade e excelência das instituições escolares (WAISELFISZ, 1993). A criação das iniciativas de avaliação e a divulgação de seus resultados têm gerado o que muitos autores, como Afonso (2009) e Freitas (2007), denominam de "quase mercado em educação", ao induzir a competitividade no sistema educacional e a comparação dos resultados das escolas, servindo para definir o repasse de recursos e o pagamento de bônus. Nesse modelo, segundo Afonso (2009), as escolas públicas passam a se organizar tendo como referência princípios da empresa privada, mesmo que tenham como objetivo central o lucro.

2.1.2 SAEB

O Sistema de Avaliação da Educação Básica (SAEB) tem como principal objetivo avaliar a educação básica brasileira e contribuir para a melhoria de sua qualidade e para a universalização do acesso à escola, oferecendo subsídios concretos para a formulação, reformulação e o monitoramento das políticas públicas voltadas para a educação básica (INEP, 2007). O SAEB realizou sua primeira avaliação em âmbito nacional em 1990, aplicando provas de conhecimentos por amostragem a alunos nos vários estados. A partir de 1995, os levantamentos passaram a concentrar-se nos alunos de 4^a e 8^a séries do ensino fundamental (nomenclaturas utilizadas até 2009, a partir de 2010 passam a ser 5^o e 9^o anos) e 3^a série do ensino médio. Além de aplicar testes, o SAEB reúne informações sobre a origem familiar dos alunos e seus hábitos e condições de estudo e sobre as práticas pedagógicas dos professores e sobre as formas de gestão da escola. O SAEB é composto

por três avaliações externas em larga escala:

Figura 1: Composição do SAEB
Fonte: Portal INEP/MEC(2014).

- **Avaliação Nacional da Educação Básica - ANEB:** abrange, de maneira amostral, alunos das redes públicas e privadas do país, em áreas urbanas e rurais, matriculados no 5^o ano e 9^o ano do Ensino Fundamental e na 3^a série do Ensino Médio, tendo como principal objetivo avaliar a qualidade, a equidade e a eficiência da educação brasileira.

- **Avaliação Nacional do Rendimento Escolar - ANRESC (também denominada "Prova Brasil"):** trata-se de uma avaliação censitária envolvendo os alunos do 5^o ano e 9^o ano do Ensino Fundamental das escolas públicas das redes municipais, estaduais e federal, com o objetivo de avaliar a qualidade do ensino ministrado nas escolas públicas. Participam desta avaliação as escolas que possuem, no mínimo, 20 alunos matriculados nas séries/anos avaliados.

- **A Avaliação Nacional da Alfabetização - ANA:** avaliação censitária envolvendo os alunos do 3^o ano do Ensino Fundamental das escolas públicas, com o objetivo principal de avaliar os níveis de alfabetização e letramento em Língua Portuguesa, alfabetização Matemática e condições de oferta do Ciclo de Alfabetização das redes públicas. A ANA foi incorporada ao SAEB pela Portaria n^o 482, de 7 de junho de 2013.

A ANEB e a ANRESC/Prova Brasil são realizadas bianualmente, enquanto a ANA é de realização anual.

2.1.3 IDEB

O IDEB é um indicador de qualidade educacional que combina informações de desempenho em exames padronizados (Prova Brasil ou SAEB) obtido pelos estudantes ao final

das etapas de ensino (5^o e 9^o anos do ensino fundamental e 3^a série do ensino médio) com informações sobre rendimento escolar (aprovação) (INEP, 2007).

O IDEB foi criado pelo Decreto n^o 6.094, de 24 de abril de 2007, como um dos eixos centrais do Plano de Desenvolvimento da Educação (PDE). Segundo o Decreto este índice consiste em um indicador objetivo de qualidade da educação. Sabe-se que, no Brasil, a questão do acesso à escola não é mais um problema, já que quase a totalidade das crianças ingressa no sistema educacional. Entretanto, as taxas de repetência dos estudantes são bastante elevadas, assim como a proporção de adolescentes que abandonam a escola antes mesmo de concluir a educação básica. Outro indicador preocupante é a baixa proficiência obtida pelos alunos em exames padronizados (Censo Escolar, 2007).

Como o IDEB é resultado do produto entre o desempenho e do rendimento escolar (ou o inverso do tempo médio de conclusão de uma série) então ele pode ser interpretado da seguinte maneira: para uma escola **A** cuja média padronizada da Prova Brasil, 5^o ano, é 5,0 e o tempo médio de conclusão de cada série é de 2 anos, a rede/escola terá o IDEB igual a 5,0 multiplicado por 1/2, ou seja, $\mathbf{IDEB = 2{,}5}$. Já uma escola **B** com média padronizada da Prova Brasil, 5^o ano, igual a 5,0 e tempo médio para conclusão igual a 1 ano, terá $\mathbf{IDEB = 5{,}0}$.

Indicadores educacionais como o IDEB são desejáveis por permitirem o monitoramento do sistema de ensino do País. Sua importância, em termos de diagnóstico e norteamento de ações políticas focalizadas na melhoria do sistema educacional, está focada em:

a) detectar escolas e/ou redes de ensino cujos alunos apresentem baixa performance em termos de rendimento e proficiência;

b) monitorar a evolução temporal do desempenho dos alunos dessas escolas e/ou redes de ensino.

As autoridades educacionais podem, por exemplo, financiar programas para promover o desenvolvimento educacional de redes de ensino em que os alunos apresentam baixo desempenho. Assim, monitorar as redes financiadas, para verificar se elas apresentam uma melhora de desempenho, é fundamental. Aliás, o financiador poderia estipular previamente o avanço desejado no indicador como contrapartida para a liberação de recursos, como no caso das metas.

2.1.4 Sistema de ensino em Campina Grande

Com a Lei Municipal n^o 3771/1999, institui-se o Sistema Municipal de Ensino de Campina Grande e tem sua configuração definida compreendendo as instituições de Educação Infantil e de Ensino Fundamental (incluindo Educação de Jovens e Adultos - EJA e Atendimento Educacional Especializado - AEE) mantidas pelo poder público municipal, as instituições de Educação Infantil criadas e mantidas pela iniciativa privada e os órgãos municipais de educação.

No ano de 2013 - ano em que foi realizado o último IDEB com resultado já publicado no Diário Oficial da União - o Município de Campina Grande contava com 257 unidades escolares de ensino infantil e fundamental, sendo 35 de Educação Infantil e 122 de Educação Fundamental. Entre essas 122 unidades, 79 possuiam 5^o ano que atendiam aos pré-requisitos para a aplicação da prova (como ter um número de alunos igual ou superior a 20, por exemplo) e apenas 11 possuiam 9^o ano que atendiam aos pré-requisitos (Censo Escolar, 2007).

Dessa forma, para o objetivo em questão nesse estudo, serão analisadas as 79 unidades escolares do município de Campina Grande que realizaram a prova do IDEB no ano de 2013.

2.2 Inferência Bayesiana

Métodos bayesianos estão incrementando sua popularidade nas ciências como meio de inferência probabilística (MALAKOFF, 1999). Dentre suas vantagens encontram-se a habilidade de incluir informação prévia, a facilidade da incorporação desta num contexto formal de decisão, o tratamento explícito da incerteza e a capacidade de assimilar novas informações em contextos adaptativos.

Estes métodos se tornaram mais populares após o progresso do hardware que possibilitou o desenvolvimento das técnicas de amostragens por simulações computacionais; esta técnica tem sido aplicada em vários campos do conhecimento (bioestatística, epidemiologia, engenharia, ciências da computação, econometria, ciências sociais e outros) (COSTA, 2004).

A modelagem bayesiana se inicia com a especificação de um modelo probabilístico completo, através da distribuição conjunta das quantidades observáveis e não observáveis, $f(x)$, do problema. A distribuição de probabilidade *a priori* $p(\theta)$ reflete o conhecimento prévio sobre os parâmetros. A Figura 2 apresenta o efeito desta distribuição de

probabilidade sobre a distribuição *a posteriori*, $p(\theta \mid x)$.

Uma distribuição "plana" indica pouca informação prévia e influencia levemente a distribuição a posteriori do parâmetro. Neste caso, a informação contida nos dados (verossimilhança) é a dominante. A Figura 2-b mostra a situação contrária, na qual a distribuição a priori contem informação suficiente para modificar substancialmente a distribuição a posteriori, gerando uma distribuição unimodal. Deve-se escolher a forma funcional da priori e o valor dos seus hiperparâmetros. Nos chamados modelos hierárquicos, os modelos nos quais estes hiperparâmetros são tratados como variáveis, sendo estimados a partir dos dados.

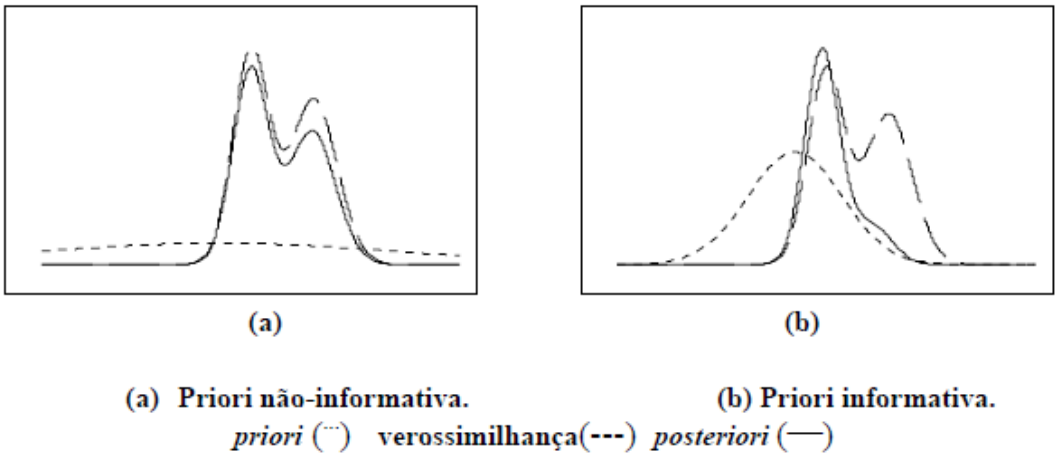

Figura 2: Efeito da distribuição *a priori*
Fonte: Técnicas Bayesianas, 2004.

2.2.1 Modelo de Regressão Linear Bayesiano

Como uma introdução aos modelos de regressão bayesiana, considera-se o modelo de regressão linear mais simples, com distribuições a priori informativas (COSTA, 2004).

Seja o modelo de regressão linear normal múltipla, definido como:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

onde $\mathbf{y}$ é um vetor $(n \times 1)$ de variáveis dependentes; $\mathbf{X}$ representa uma matriz $(n \times k)$ de variáveis exploratórias e $\boldsymbol{\varepsilon}$ é um vetor $(n \times 1)$ de variáveis aleatórias independentes, normalmente distribuídas e de variância constante σ^2.

Os parâmetros a serem estimados neste modelo são β e σ^2. Assumindo estes dois parâmetros como independentes, a distribuição a priori conjunta verifica:

$$p(\beta, \sigma^2) = p(\beta).p(\sigma^2) \tag{2.1}$$

Completa-se o modelo com as prioris independentes para β e σ^2,

$$\beta \sim N(\mathbf{r}, \mathbf{T}) \tag{2.2}$$

$$\sigma^2 \sim Inv - \chi^2(\nu, s^2) \tag{2.3}$$

onde $\mathbf{r}$ é um vetor $(k \times 1)$ contendo as médias a priori; $\mathbf{T}$ é uma matriz $(k \times k)$ contendo as variâncias e covariâncias a priori e ν e s^2 são os graus de liberdade e o fator de escala respectivamente. O modelo poderia ser completado com qualquer distribuição que expresse adequadamente o conhecimento anterior à observação dos dados. Porém, as expressões (2.2) e (2.3) possuem propriedades analíticas atrativas por serem distribuições a priori conjugadas.

Na sequência são apresentadas as expressões analíticas para as prioris:

$$p(\beta) = \frac{\mid T^{-1} \mid^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} e^{\frac{-1}{2}.[(\beta - r)^T T^{-1}(\beta - r)]}$$

$$p(\sigma^2) = \frac{(\nu s^2)^{\frac{\nu}{2}} (\sigma^2)^{\frac{-(\nu+2)}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} e^{\frac{-\nu s^2}{2\sigma^2}}$$

onde $\Gamma(\frac{\nu}{2})$ é a função gamma.

A função verossimilhança ou densidade condicional dos dados é:

$$L(\beta, \sigma^2) = p(y \mid X, \beta, \sigma^2) = \frac{e^{\frac{-1}{2\sigma^2}.[(y - X\beta)^T (y - X\beta)]}}{(2\pi\sigma^2)^{\frac{n}{2}}}.$$

Seguindo a metodologia bayesiana, combina-se a função de verossimilhança com as prioris $p(\beta)$ e $p(\sigma)$ para produzir a densidade a posteriori, e obtém-se:

$$p(\beta, \sigma^2 \mid y, X) \propto \frac{\mid T^{-1} \mid^{1/2}}{(2\pi)^{(k+n)/2}} . \frac{(\nu s^2)^{\nu/2} (\sigma^2)^{-(\nu+n+2)/2}}{2^{\nu/2} \Gamma(\nu/2)} . e^{-(\nu s^2)/2\sigma^2} . e^{-R/2} . e^{-1/2.[(\beta - \bar{\beta})^T V^{-1} (\beta - \bar{\beta})]} \tag{2.4}$$

onde:

$$V = \left(T^{-1} + \frac{X^T X}{\sigma^2} \right)^{-1}$$

$$\bar{\beta} = \left(T^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1} . \left(T^{-1} r + \frac{X^T y}{\sigma^2}\right)$$

$$R = \frac{y^T y}{\sigma^2} + r^T T^{-1} r - \bar{\beta}^T V^{-1} \bar{\beta}$$

Em geral, a inferência bayesiana exige a integração de funções multidimensionais complexas. Inferência exata será possível somente se estas integrais puderem ser determinadas analiticamente. Porém, na maioria dos casos é necessário efetuar aproximações (COSTA, 2004).

2.3 Seleção de Modelos

Diversos métodos de seleção de modelo são citados na literatura, neste trabalho utilizaremos o Coeficiente de Determinação R^2 - por ser usado nos métodos de seleção de variáveis *Backward, Forward e Stepwise* -, Critério de Informação de Akaike - por penalizar modelos com maior número de parâmetros - e Critério de Informação por Deviância - bastante utilizado quando se trata de simulações.

2.3.1 Coeficiente de Determinação - R^2

Geralmente usado no método de todas as regressões possíveis, o coeficiente de determinção quantifica a proporção da variabilidade da variável resposta que é explicada por um modelo de aproximação qualquer, é dado por:

$$R^2 = 1 - \frac{SQRes}{SQTotal}$$

O R^2 pode assumir valores no intervalo [0,1], sendo que valores próximos de 1 denotam uma boa relação entre a variável resposta e as p variáveis preditoras, indicando um bom ajuste. Por outro lado, valores próximos ou iguais a 0 denotam que o modelo não é superior a média amostral (DRAPER; SMITH, 1998).

Apesar do caráter relativo do coeficiente de determinação, parece lógico usá-lo como critério de seleção de modelos, já que, ao contrário do teste F do modelo, fornece uma "medida"do quanto o modelo é superior à média amostral. Sendo assim, se forem considerados dois modelos concorrentes cujos teste F dos modelos rejeitassem a hipótese nula optaria-se pelo R^2 de maior valor, ou com melhor quantidade de ajuste.

2.3.2 Critério de Informação de Akaike - AIC

O valor do AIC é simples de ser obtido para os casos de estimação por mínimos quadrados, como é o caso do ajuste de modelos de regressão, e para os casos de análises baseadas na estimativas de verossimilhança de uma forma geral (BURNHAM; ANDERSON, 2004).

Sua fórmula é dada por:

$$AIC = -2logL + 2n$$

sendo L a função de verossimilhança e n o número de parâmetros ajustados.

2.3.3 Critério de Informação de Desvio - DIC

O DIC é composto pela média a posteriori do desvio penalizado pelo número de parâmetros do modelo. Este critério é atrativo pois pode ser incorporado durante a simulação de Monte Carlo. Semelhante aos outros critérios, é uma aproximação assintótica para amostras grandes e é válido quando a distribuição a posteriori é aproximadamente uma distribuição normal multivariada (ARRABAL, 2012). Valores baixos para o DIC indicam melhores ajustes. O DIC é obtido por:

$$DIC = \bar{D}(\beta, M_i) + p_{di}$$

em que $p_{di} = \bar{D}(\beta, M_i) - D(\bar{\beta}, M_i)$ mede a complexidade do modelo i. O critério sugere uma comparação entre o desvio médio e o desvio aplicado na média a posteriori.

2.3.4 Métodos de seleção de variáveis

Como o número de regressões possíveis cresce com o aumento do número de parâmetros, foram propostos os métodos de regressão passo a passo, que embora não sejam ótimos, requerem um tempo computacional bem menor do que o de todas as regressões possíveis (DEMÉTRIO, 2008).

- **Método do passo atrás (*Backward*)** Consiste em ajustar inicialmente o modelo completo e a seguir eliminar variáveis, uma a uma, com menor correlação parcial com a resposta Y, menor diminuição no R^2 ou menor diminuição significativa no teste F parcial ou no teste t parcial, de acordo com algum critério de parada.

  Os passos, baseados no teste F parcial, a serem seguidos são:

1- Ajustar o modelo completo com l variáveis e obter $SQRes_c$ com $n-l$ graus de liberdade;

2- Para cada uma das l variáveis do modelo completo do Passo **1**, considerar o modelo reduzido, com a retirada de uma variável e calcular $SQRes_r$ com $n-l+1$ graus de liberdade e $F = \frac{SQRes_r - SQRes_c}{QMRes_c}$;

3- Obter F_{min};

4- Comparar F_{min} com F_{tab} (percentil da tabela de F com 1 e $n-l$ graus de liberdade a um nível de significância α, usa-se, em geral, $\alpha = 0,10$):

(i) se $F_{min} > F_{tab}$, não eliminar nenhuma variável e parar o processo, ficando o modelo completo com l variáveis;
(ii) se $F_{min} < F_{tab}$, eliminar a variável com F_{min} e voltar ao Passo **1** com novo modelo completo com $l = l - 1$ variáveis.

- **Método do passo a frente (*Forward*)** Consiste em incluir inicialmente no modelo a variável com maior coeficiente de correlação simples com a variável resposta e a seguir variáveis, uma a uma, com maior correlação parcial com a resposta Y, maior aumento no R^2 ou maior aumento significativo no teste F parcial ou no teste t parcial, de acordo com algum critério de parada.

  Os passos, baseados no teste F parcial, a serem seguidos são:

  1- Ajustar o modelo reduzido com m variáveis e obter $SQRes_r$ com $n-m$ graus de liberdade;

  2- Para cada uma das variáveis não pertencentes ao modelo do Passo **1**, considerar o modelo completo, com a adição de uma variável extra e calcular $SQRes_c$ com $n-m-1$ graus de liberdade e $F = \frac{SQRes_r - SQRes_c}{QMRes_c}$;

  3- Obter F_{max};

  4- Comparar F_{max} com F_{tab} (percentil da tabela de F com 1 e $n-m-l$ graus de liberdade a um nível de significância α, usa-se, em geral, $\alpha = 0,10$):

  (i) se $F_{max} > F_{tab}$, incluir variável com F_{max} e voltar ao Passo **l** com novo modelo reduzido com $m = m + 1$ variáveis;
  (ii) se $F_{max} < F_{tab}$, não incluir a variável com o F_{max} e parar o processo, ficando o modelo completo com m variáveis.

- **Método do passo a frente passo atrás (*Stepwise*)** Consiste na mistura dos dois anteriores. Os passos, baseados no teste F parcial, a serem seguidos são:

1- Ajustar o modelo reduzido com m variáveis e obter $SQRes_r$ com $n-m$ graus de liberdade;

2- Para cada uma das variáveis não pertencentes ao modelo do Passo **1**, considerar o modelo completo, com a adição de uma variável extra e calcular $SQRes_c$ com $n-m-1$ graus de liberdade e $F = \frac{SQRes_r - SQRes_c}{QMRes_c}$;

3- Obter F_{max};

4- Comparar F_{max} com F_{tab} (percentil da tabela de F com 1 e $n-m-l$ graus de liberdade a um nível de significância α, usa-se, em geral, $\alpha = 0,10$):

(i) se $F_{max} > F_{tab}$, incluir a variável com F_{max} e passar para o Passo **5** com modelo completo com $l = m+1$ variáveis;
(ii) se $F_{max} < F_{tab}$, não incluir a variável com o F_{max} e passar para o Passo **5** com modelo completo com $l = m$ variáveis;

5- Ajustar o modelo completo com l variáveis e obter $SQRes_c$ com $n-l$ graus de liberdade;

6- Para cada uma das l variáveis do modelo completo do Passo **4**, considerar o modelo reduzido, com a retirada de uma variável e calcular $SQRes_r$ com $n-l+1$ graus de liberdade e: $F = \frac{SQRes_r - SQRes_c}{QMRes_c}$;

7- Obter F_{min};

8- Comparar F_{min} com F_{tab} (percentil da tabela de F com 1 e $n-l$ graus de liberdade a um nível de significância α, usa-se, em geral, $\alpha = 0,10$):

(i) se $F_{min} > F_{tab}$, não eliminar nenhuma variável e voltar ao Passo **1** com novo modelo reduzido com $m = l$ variáveis e parar o processo se no Passo **4** nenhuma variável for incluída;
(ii) se $F_{min} < F_{tab}$, eliminar a variável com F_{min} e voltar ao Passo **1** com novo modelo reduzido com $m = l-1$ variáveis.

3 Material e Métodos

A partir dos dados oficiais do IDEB - publicados pelo INEP - foram selecionadas 79 escolas municipais de Campina Grande que possuiam 5^o ano e realizaram a prova do IDEB em 2013.

Para a análise dos dados utilizaremos os softwares Rstudio e OpenBUGS para ambiente Windows. Sendo softwares estatísticos e de fácil acesso, possuem as ferramentas necessárias para as análises.

3.1 O cálculo do IDEB

A forma geral do IDEB é dada por (3.1):

$$IDEB_{ji} = N_{ji}P_{ji} \quad \text{com} \quad 0 \leq N_j \leq 10; \;\; 0 \leq P_j \leq 1 \;\; e \;\; 0 \leq IDEB_j \leq 10 \tag{3.1}$$

em que,
i = ano do exame (SAEB e Prova Brasil) e do Censo Escolar;
N_{ji} = média da proficiência em Língua Portuguesa e Matemática, padronizada para um indicador entre 0 e 10, dos alunos da unidade j, obtida em determinada edição do exame realizado ao final da etapa de ensino;
P_{ji} = indicador de rendimento baseado na taxa de aprovação da etapa de ensino dos alunos da unidade j.

Em (3.1), a média de proficiência padronizada dos estudantes da unidade j, N_{ji}, é obtida a partir das proficiências médias em Língua Portuguesa e Matemática dos estudantes submetidos a determinada edição do exame realizado ao final da etapa educacional considerada (Prova Brasil ou Saeb). A proficiência média é padronizada para estar entre zero e dez, de modo que $0 \leq IDEB \leq 10$. N_{ji} é obtida de acordo com (3.2).

$$N_{ji} = \frac{n_{ji}^{lp} + n_{ji}^{mat}}{2} \quad e \quad n_{ji}^{\alpha} = \frac{S_{ji}^{\alpha} - S_{inf}^{\alpha}}{S_{sup}^{\alpha} - S_{inf}^{\alpha}} 10 \tag{3.2}$$

em que,

n_{ji}^{α} = proficiência na disciplina α, obtida pela unidade j, no ano i, padronizada para valores entre 0 e 10;
α = disciplina (Matemática ou Língua Portuguesa);
S_{ji}^{α} = proficiência média (Língua Portuguesa ou Matemática), não padronizada, dos alunos da unidade j obtida no exame do ano i;
S_{inf}^{α} = limite inferior da média de proficiência (Língua Portuguesa ou Matemática) do SAEB 1997 (ano de referência);
S_{sup}^{α} = limite superior da média de proficiência (Língua Portuguesa ou Matemática) do SAEB 1997 (ano de referência).

Para as unidades escolares (ou redes) que obtiverem a $S_{ji}^{\alpha} < S_{inf}^{\alpha}$, a proficiência média é fixada em S_{inf}^{α}. Por sua vez, aquelas unidades que obtiverem $S_{ji}^{\alpha} > S_{sup}^{\alpha}$ têm o desempenho fixado em S_{sup}^{α}. A Tabela 1 apresenta a média e o desvio padrão das proficiências dos alunos da 4^a e da 8^a série do ensino fundamental (EF) e da 3^a série do ensino médio (EM) no SAEB de 1997. Posteriormente, a Tabela 2 apresenta os valores dos limites inferiores e superiores utilizados na padronização das proficiências médias em Língua Portuguesa e Matemática dos alunos da 4^a e da 8^a série do ensino fundamental e da 3^a série do ensino médio.

Tabela 1: SAEB 1997 - Proficiências médias e desvio padrão

Série	Matemática		Língua Portuguesa	
	Média	Desvio Padrão	Média	Desvio Padrão
4^a do EF	190,8	44	186,5	46
8^a do EF	250,0	50	250,0	50
3^a 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

$$S_{inf}^{\alpha} = \bar{X}_{\alpha} - (3DP) \quad e \quad S_{sup}^{\alpha} = \bar{X}_{\alpha} + (3DP). \tag{3.3}$$

Esses limites, inferiores e superiores, apresentados na Tabela 2, são usados para calcular todos os IDEBs, 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.

Tabela 2: Limite superior e inferior das proficiências

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

Fonte: SAEB 1997 - INEP/MEC.

O indicador de rendimento, P_j , é obtido conforme (2.4), onde a proporção de aprovados em cada uma das séries da etapa considerada, p^r , é calculada diretamente do Censo Escolar. Se p^r $(r = 1, 2, ..., n$, em que n é o número de séries com taxa de aprovação positiva) é a taxa de aprovação da r-ésima série da etapa educacional considerada, então o tempo médio de duração da série é:

$$T_{ji} = \sum_{r=1}^{n} \frac{1}{p^r} = \frac{n}{P_{ji}} \tag{3.4}$$

Em (3.4), P_{ji} é a taxa média de aprovação na etapa educacional no ano i. Nota-se que, na ausência de evasão durante a etapa e em equilíbrio estacionário, $\frac{n}{P_{ji}}$ dá o tempo médio para conclusão de uma etapa para os estudantes da unidade $j(T_{ji})$.

Se P é o inverso do tempo médio para conclusão de uma série, então, $P_{ji} = \frac{1}{T_{ji}}$. Deste modo, temos que $IDEB_{ji} = \frac{N_{ji}}{T_{ji}}$, ou seja, o indicador fica sendo a pontuação no exame padronizado ajustada pelo tempo médio (em anos) para conclusão de uma série naquela etapa de ensino (Nota Técnica: IDEB, 2007).

3.2 Modelos de Regressão Linear Múltipla e Multinível

A análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis, sendo uma das mais utilizadas na análise de dados (Barros et al. 2008). No estudo em questão trataremos da Regressão Linear Múltipla e multinível especificamente.

3.2.1 Regressão Linear Múltipla

- **Modelo estatístico**

Em análise de regressão linear múltipla existe uma relação funcional entre uma variável dependente e k variáveis independentes, conforme elucidado na equação (3.5).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i, \tag{3.5}$$

onde Y_i é a variável resposta; $\beta_0, \beta_1, \ldots, \beta_k$ são parâmetros desconhecidos; $X_1, \ldots, X_k$ são variáveis regressoras e ε_i é o erro aleatório associado ao modelo.

Em termos matriciais, o modelo de regressão linear múltipla é dado por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

sendo,

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{n1} \\ 1 & x_{21} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \ e \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

em que, $\mathbf{Y}$ é um vetor de dimensão $n \times 1$ da variável aleatória Y, $\mathbf{X}$ é a matriz de dimensões $n \times k$, $\boldsymbol{\beta}$ é o vetor de dimensão $p \times 1$ de parâmetros desconhecidos e $\boldsymbol{\varepsilon}$ é o vetor de dimesão $n \times 1$ de variáveis aleatórias não observáveis.

- **Análise de variância e teste F**

A análise de variância (ANOVA) é baseada na decomposição da soma de quadrados e graus de liberdade associados a variável resposta Y, e na utilização do teste F para verificar se as variáveis independentes conjuntamente contribuem significativamente para explicá-la (RÊGO, 2012).

Em resumo, o procedimento é descrito na Tabela 3.

Tabela 3: Análise de variância

Fonte de Variação	GL	SQ	QM	F
Regressão	k	$SQReg$	$\frac{SQReg}{k}$	$\frac{QMReg}{QMRes}$
Resíduo	$n-p$	$SQRes$	$\frac{SQRes}{n-p}$	–
Total	$n-1$	$SQTot$	–	–

onde,

- $SQRes = \mathbf{Y'Y} - \boldsymbol{\beta'X'Y}$;

- $SQTotal = \mathbf{Y'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$;
- $SQReg = \hat{\boldsymbol{\beta}}\mathbf{'X'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$;
- e define-se as hipóteses para o teste F como:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \beta_j \neq 0 \text{ para pelo menos um, } j = i, 2, \cdots, k. \end{cases}$$

A estatística de teste será:

$$F = \frac{QMReg}{QMRes} \sim F_{(k,n-p)},$$

em que, k é o número de variáveis independentes e $p = K + 1$. Encontrando os valores de F e fixando um nível α de significância, pode-se decidir se:

$F_{calculado} > F_{tabelado}$, rejeita-se a hipótese H_0 e conclui-se ao nível α de significância que há indícios de regressão;

$F_{calculado} < F_{tabelado}$, não rejeita-se a hipótese H_0 e conclui-se ao nível α de significância que não há indícios de regressão.

3.2.2 Regressão Multinível

O ajuste de modelos multiníveis é baseado na estimação dos mínimos quadrados generalizados (GLS). Dessa forma, escreve-se um modelo contendo parâmetros fixos e aleatórios, como se segue:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta},$$

onde $\mathbf{X}$ é a matriz associada com o vetor de parâmetros fixos, $\mathbf{Z}$ é a matriz associada com o vetor de parâmetros aleatórios e $\mathbf{Y}$ é o vetor de respostas. A estimação dos parâmetros é feita iterativamente através dos mínimos quadrados generalizados (IGLS), ajustando-se modelos de regressão para as partes fixa e aleatória (GOLDSTEIN, 1995).

Suponha um simples modelo de componente de variância com dois níveis:

$$\mathbf{Y}_{ij} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_{ij} + \mathbf{u}_{0j} + \mathbf{r}_{ij}.$$

Através do GLS, estimam-se os coeficientes fixos,

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y},$$

onde $\mathbf{V}$ é a matriz de variância e covariância de $\mathbf{Y}$, com resíduos normais. Esse proce-

dimento iterativo é iniciado a partir de uma estimativa razoável de mínimos quadrados ordinários (OLS) dos coeficientes fixos.

Tendo estimado os parâmetros fixos, $\boldsymbol{\beta}$, extrai-se um vetor residual, $\mathbf{Y'} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, que é utilizado para estimar os parâmetros aleatórios do modelo. Calcula-se em seguida, $E(\mathbf{Y}*) = \mathbf{Y'Y'}^T = \mathbf{V}$, construindo o vetor $\mathbf{Y} * * = Vec(\mathbf{Y}*)$ para ser utilizado como variável resposta na equação de regressão para estimar os parâmetros aleatórios, $\theta^c = (Z^T V *^{-1} Z)^{-1} Z^T V^{-1} * Y * *$. Nesta equação, $V*$ é o produto de *Kronecker* de $V (V* = V \otimes V$.

Assim, alterna-se entre estimar o vetor de parâmetros fixos e aleatórios até a convergência do modelo. Para os modelos não-lineares é utilizado a quasi-verossimilhança (GOLDSTEIN, 1995).

- **Modelos Lineares Multiníveis com 2 Níveis**

O modelo para análise no nível 1 pode ser denotado por:

$$Y_{ij} = \beta_{0j} + \beta_{ij}(X_{ij} - X^b_{.j}) + r_{ij}, \tag{3.6}$$

onde r_{ij} tem distribuição normal, $N(0, \sigma^2)$, com n_i unidades no nível 1 dentro de cada J-ésima unidade no nível 2 e Y_{ij} é uma função do conjunto de características individuais das unidades no nível 1.

Utilizando o modelo (3.6) no nível 1, podemos estudar a(s) associação(ões) da(s) variável(is) X_{ij} com a resposta Y_{IJ} dentro de uma população de unidades no nível 2.

Cada unidade no nível 2 é descrita pelo par de valores $(\beta_{0,j}, \beta_{1,j})$, que podem ser preditos através das características das unidades no nível 2. Por exemplo, considere a variável preditora W_j. Assim, o modelo no nível 2 será:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \tag{3.7}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \tag{3.8}$$

Substituindo as equações (3.7) e (3.8) em (3.6), produz-se a equação de predição para a resposta:

$$Y_{ij} = \gamma_{00} + \gamma_{01} Y_j + \gamma_{10}(X_{ij} - X^b_j) + \gamma_{11} W_j (X_{ij} - X^b_{.j}) + u_{0j} + u_{ij}(X_{ij} - X^b_{.j}) + r_{ij}$$

Este é um exemplo simples de um modelo linear multinível. Note que a equação acima

não é um típico modelo linear padrão assumindo mínimos quadrados ordinários (OLS), pois os erros têm uma estrutura mais complexa. Tais erros, u_{0j} e u_{1j}, são dependentes dentro de cada j-ésima unidade no nível 2.

E esses erros têm variância desiguais porque u_{0j} e u_{1j} variam ao redor das unidades do nível superior (2), e u_{1j} também varia de acordo com os valores de $(X_{ij} - X^b_{.j})$ das unidades no nível 1.

Assim, os modelos lineares multiníveis fornecem uma maior flexibilidade na formulação de modelos explicativos nas equações nas pesquisas em diferentes áreas do conhecimento: educacional, geográfica, sociológica, epidemiológica, demográfica e econômica (SANTOS et al. 2000).

3.3 Análise Multivariada

A Estatística Multivariada inclui os métodos de análise das relações de múltiplas variáveis dependentes e/ou múltiplas variáveis independentes, quer se estabeleçam ou não relações de causa/efeito entre estes dois grupos. São também incluídos na estatística multivariada os métodos de análise das relações entre indivíduos por duas ou mais variáveis.

Entre as várias possibilidades de aplicações da Estatística Multivariada destaca-se a Análise de Agrupamento, onde o propósito é identificar objetos que sejam similares o bastante para serem agrupados em um mesmo grupo (FERREIRA, 2008).

3.3.1 Análise de Agrupamento

Em análise de agrupamento, os objetos dentro de um grupo devem estar muito próximos uns dos outros e os grupos devem ser bastante diferentes entre si. Cada observação é atribuída a um único grupo apenas, embora alguns métodos de agrupamentos não hierárquicos como Fuzzy das C-médias permitam que um elemento amostral seja classificado em mais de uma população (MINGOTI, 2005).

A separação dos objetos para determinados grupos (agrupamento) é feita por meio de ligações. Os tipos de ligações mais comuns são: Ligações Simples (vizinho mais próximo), Ligações Completas (vizinho mais distante), Método das Médias das Distâncias (utilizado neste trabalho), Método do Centróide, Método de Ward. Vamos ver detalhadamente os três primeiros tipos de ligações:

- **Ligação simples (vizinho mais próximo)**

Nas ligações simples o agrupamento é feito juntando-se dois grupos com menor distância ou maior similaridade. Uma vez formado o novo grupo, por exemplo, (AB), na ligação simples, a dissimilaridade entre (AB) e algum outro grupo C é calculado:

$$d_{(AB)C} = min\{d_{AC}, d_{BC}\}.$$

Os resultados obtidos são dispostos graficamente em um diagrama em árvore ou dendrograma que possui uma escala para se observar os níveis.

- **Ligação completa (vizinho mais distante)**
  Na ligação completa o procedimento é muito semelhante ao da ligação simples, com uma única exceção. O algoritmo aglomerativo começa determinando a menor distância d_{ik}, constrói-se a matriz de distâncias $D = (d_{ik})$ e os grupos vão se juntando. Se A e B são dois grupos de um único elemento, tem-se (A, B) como novo grupo. A distância entre (A, B) e outro grupo C é dada por:

$$d[(A, B), C] = max\{d_{(AC)}, d_{(BC)}\}$$

- **Método das Médias das Distâncias**
  Na ligação média consideramos a distância média, segundo a figura abaixo, como a distância entre agrupamentos:

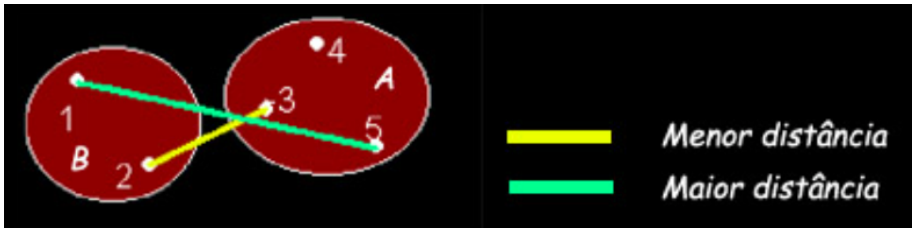

Figura 3: Esquema da distância média entre agrupamentos
Fonte: The cyclops project, 2016.

A fórmula para cálculo da distância média d_{media} é dada pela média das distâncias entre todos os pares de pontos:

$$\bar{d} = \frac{d_{(1,3)} + d_{(1,4)} + \ldots + d_{(2,5)}}{6}$$

Para montar o dendograma, procedemos unindo sempre grupos que apresentem a menor distância de acordo com uma das três regras acima.

3.3.2 Índice de Rand

O índice de Rand permite comparar duas partições com número de grupos não necessariamente iguais. Basicamente, este índice baseia-se no número de pares de parcelas que foram atribuídos da mesma maneira em cada uma das partições, ou seja, baseia-se no número de pares de parcelas concordantes (tipo I e II). Assim, temos o índice de Rand, designado por IR e é dado por (RAND, 1971):

$$IR = \frac{\binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}[\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2]}{\binom{n}{2}}$$

Verifica-se $0 \leq IR \leq 1$, tomando o valor 0 quando as duas partições não têm qualquer semelhança (ou seja, quando uma partição é constituída por um só grupo com todos as parcelas, e a outra é constituída por n grupos com 1 parcela cada) e o valor 1 quando o acordo entre as duas partições é completo.

3.4 OpenBugs

O OpenBugs é um software livre que permite simular distribuições à posteriori através do uso de algoritmos MCMC (ALBERT, 2007). A linguagem utilizada é similar à do R, mas o OpenBugs tem a seguinte estrutura:

- Especificação da distribuição da variável resposta;
- Especificação das distribuições a priori para os parâmetros;
- Leitura do banco de dados;
- Especificação dos valores iniciais (optativo).

Para exemplificar, veremos um exemplo de modelo linear com uma covariável. Primeiramente, escrevemos o algoritmo do modelo - Figura 4:

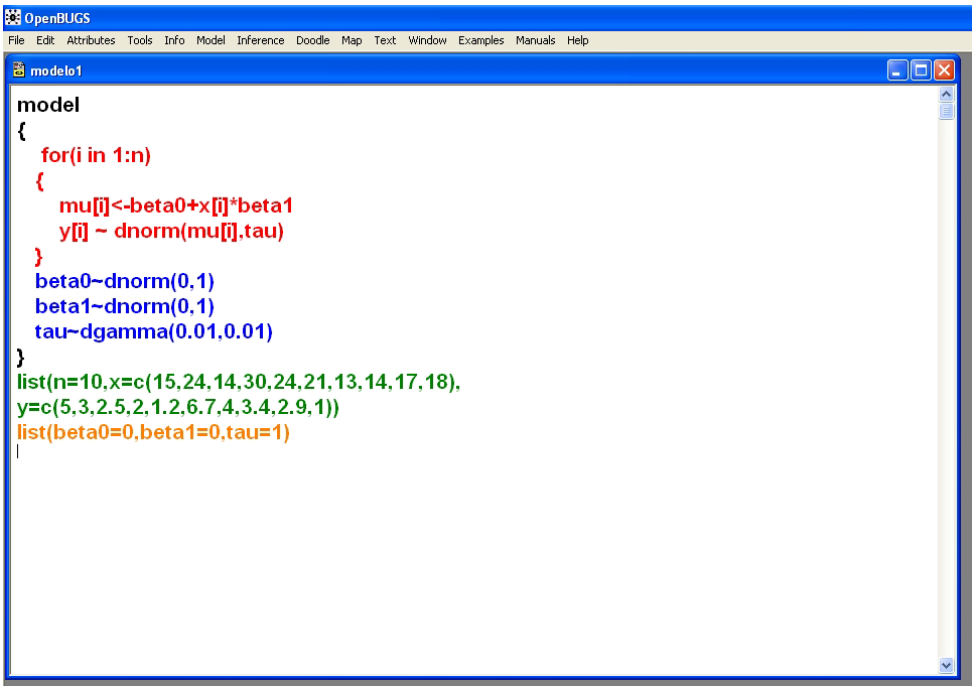

Figura 4: Modelo linear com uma covariável.

Vamos ao menu Model/Specification e clicamos em *check model*:

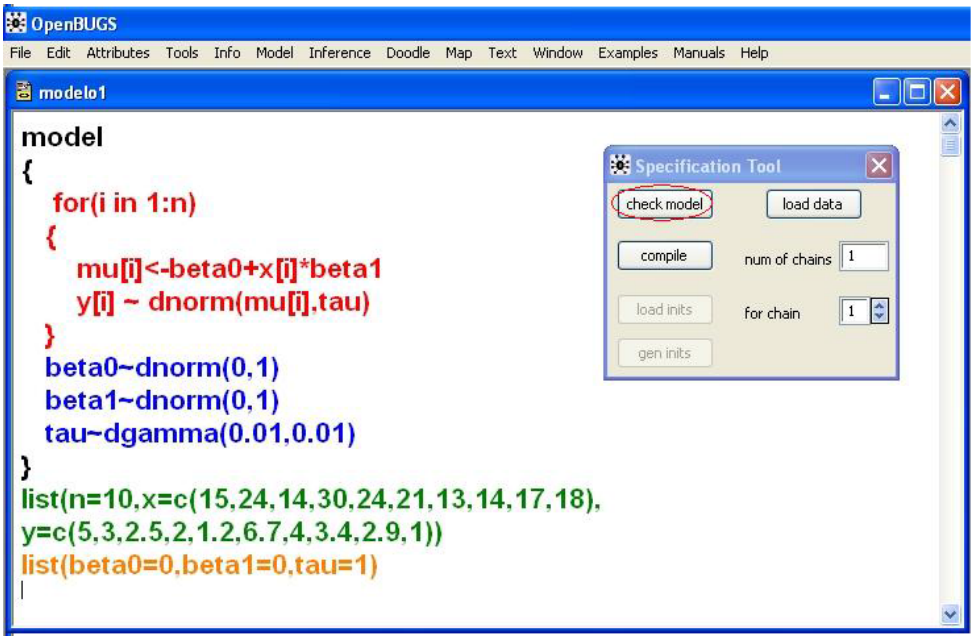

Figura 5: Checagem do modelo.

Se estiver tudo correto, aparecerá embaixo a mensagem *model is syntactically correct*. Na mesma janela, selecionamos os dados e clicamos em *load data*:

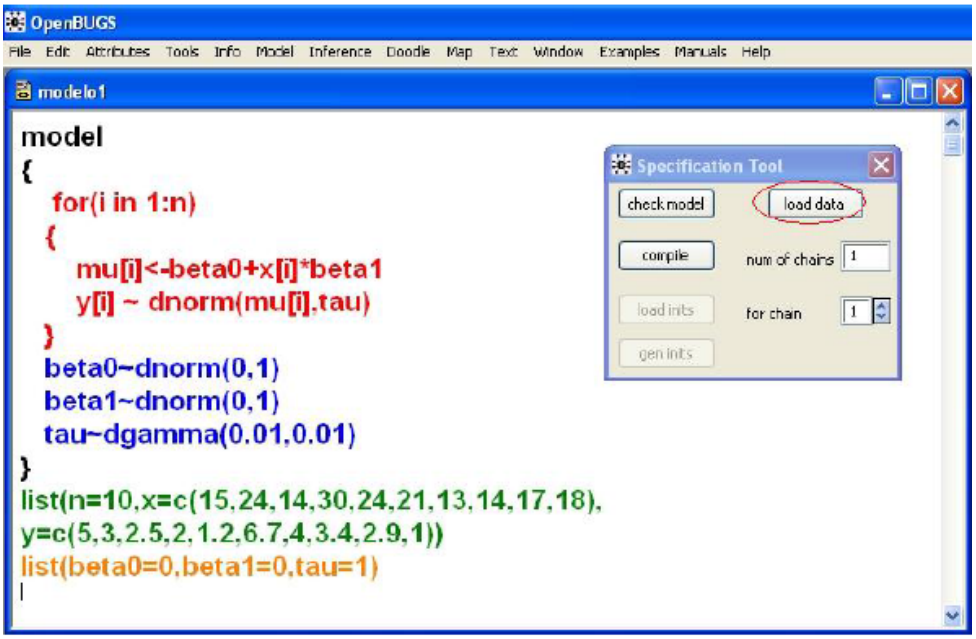

Figura 6: Carregando os dados do modelo.

Se estiver tudo correto, aparecerá embaixo a mensagem *data loaded*. Na mesma janela, clicamos em *compile*:

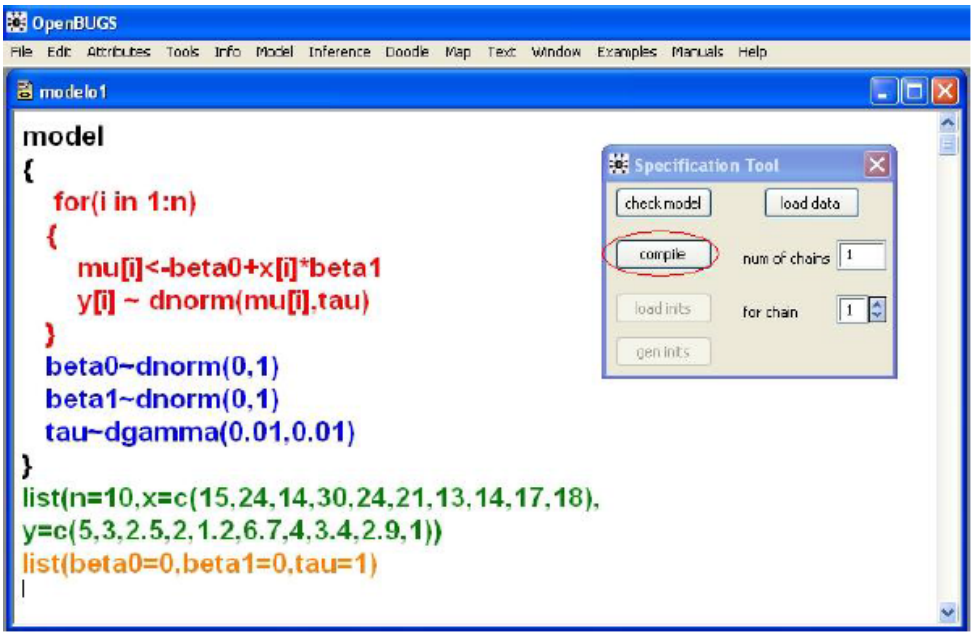

Figura 7: Compilando o modelo.

Se estiver tudo correto, aparecerá embaixo a mensagem *model compiled*. Para especificar os valores iniciais dos parâmetros, seleciona-se a lista inicial e clicamos em *load inits*.

Se não desejamos especificar valores iniciais, clicamos em *gen inits*:

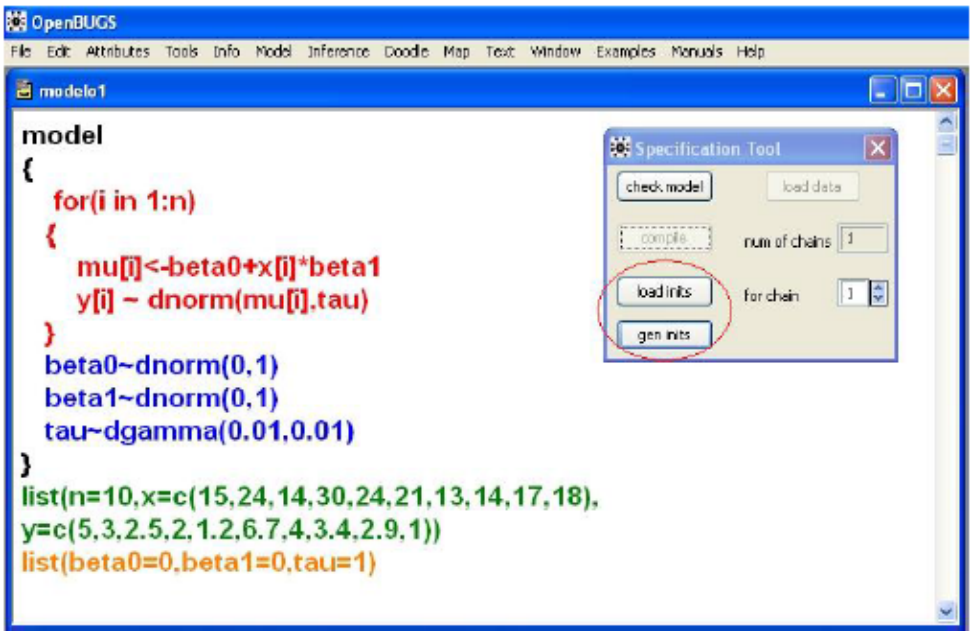

Figura 8: Especificando valores iniciais.

Deverá aparecer a mensagem *model is initialized*. Uma vez especificado o modelo, as cadeias começam a ser simuladas. Para isso, vamos ao menu model/update, especificamos o *burn-in* e clicamos em *update*:

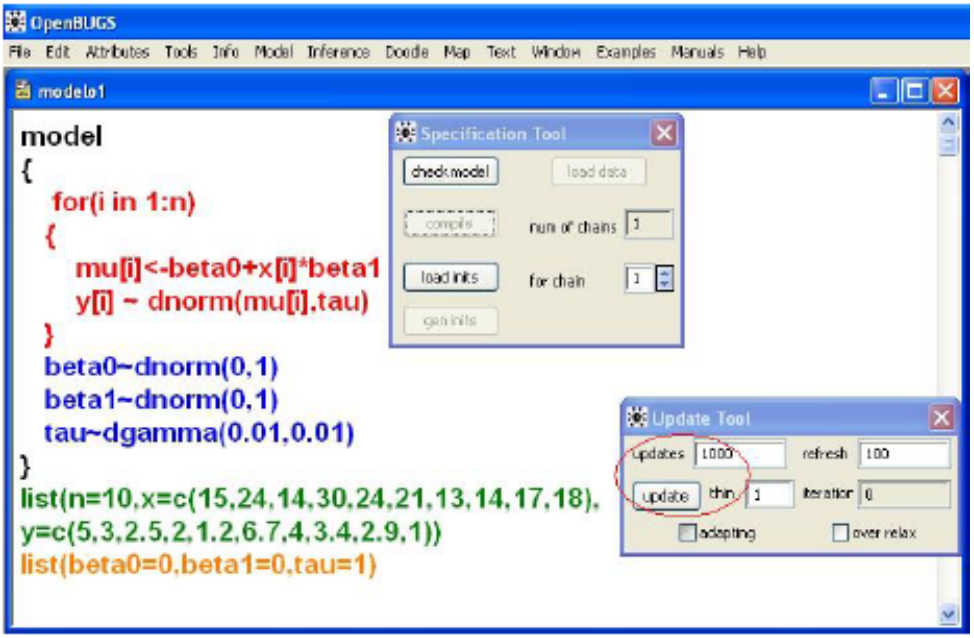

Figura 9: Simulando as cadeias do modelo.

Uma vez feito o *burn-in*, começa a simulação dos valores para nossa amostra de interesse. Para isso, vamos ao menu Inference/Samples.

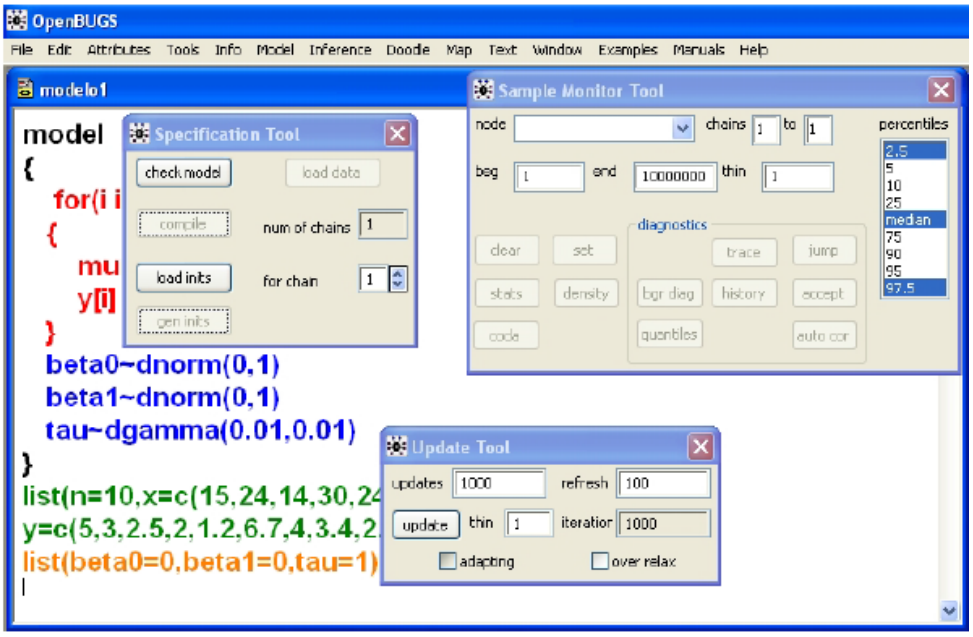

Figura 10: Simulando valores para a amostra de interesse.

Na opção *node* é introduzido o nome dos parâmetros de interesse. Se o nome é válido, clica a opção *set*. Repetir o processo para todos os parâmetros:

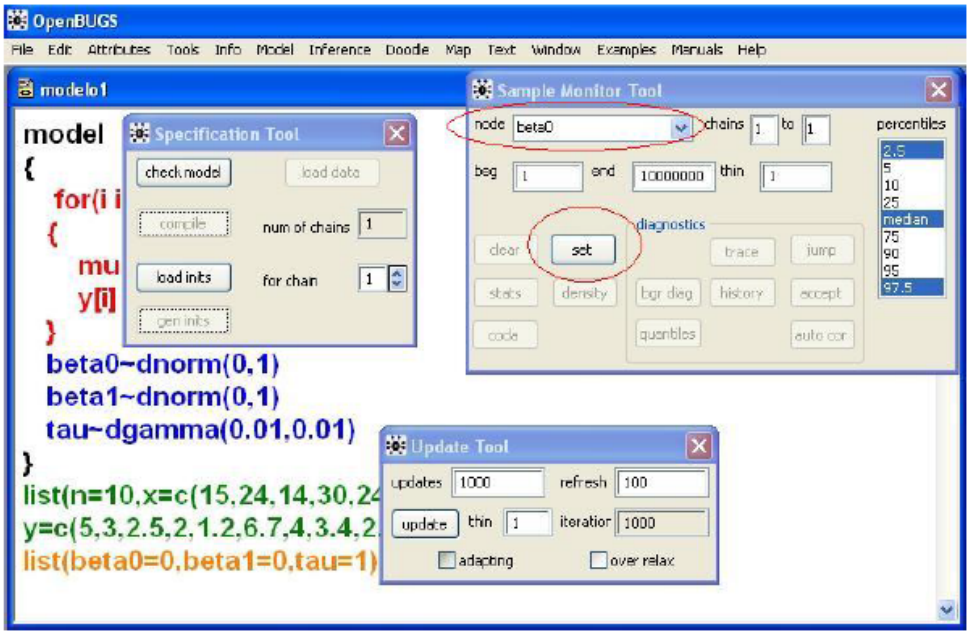

Figura 11: Parâmetros de interesse.

Uma vez introduzidos todos os parâmetros, colocamos * na opção *node*. Isso indica que queremos informação sobre todos os parâmetros introduzidos:

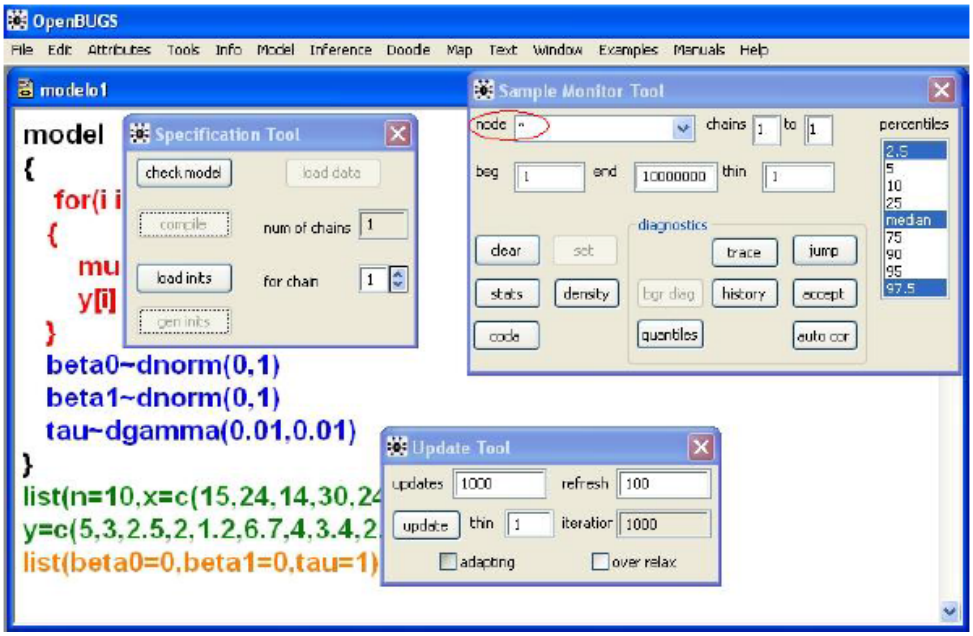

Figura 12: Informação sobre os parâmetros.

Agora já se pode simular valores das distribuições à posteriori de nossos parâmetros. Seleciona-se o tamanho de amostra desejada na janela UPDATE TOOL anterior e clicamos em *update*.

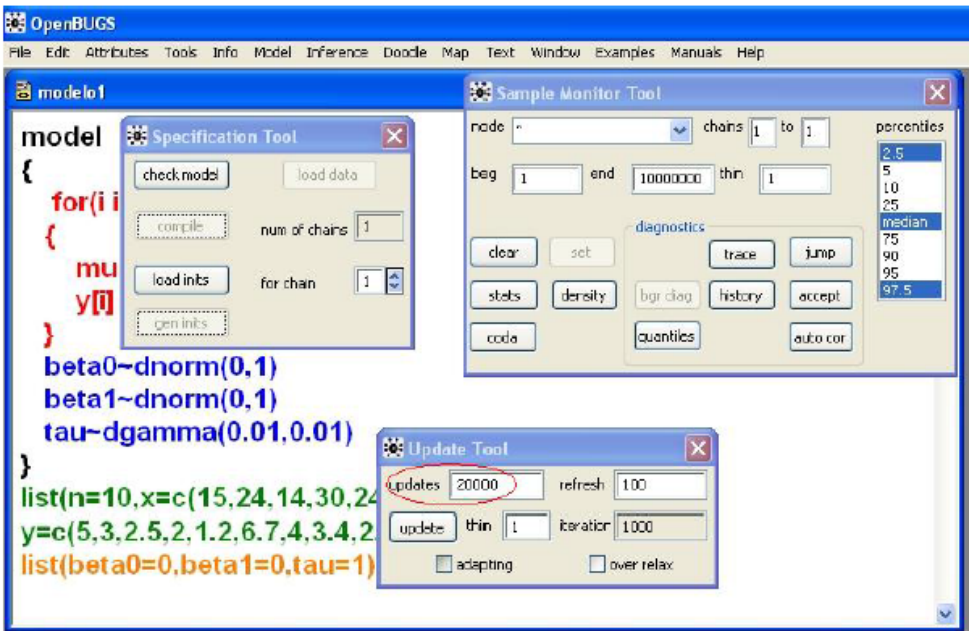

Figura 13: Selecionando o tamanho da amostra.

Agora pode ser pedido um resumo para cada parâmetro da amostra simulada. Na janela SAMPLE MONITOR TOOL clicamos em *stats* e aparecerá uma janela similar a:

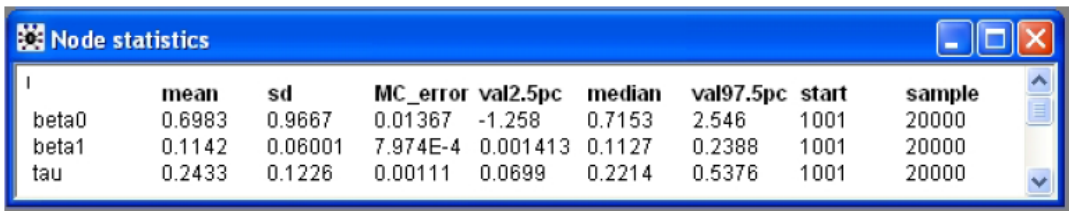

Figura 14: Resumo dos parâmetros da amostra simulada.

Dentro da mesma janela SAMPLE MONITOR TOOL, aparecem outras opções para a nossa amostra à posteriori simulada. Algumas delas são: *history*: mostra a série de todos os valores simulados para cada parâmetro; *accept*: mostra a série com a probabilidade de aceitação para iteração; *coda*: entrega os valores simulados para cada um dos parâmetros em cada iteração. Útil para utilizá-los em outros programas.

3.5 Estrutura dos Dados

Tomando cada escola como uma observação, foram avaliadas 14 variáveis: número de funcionários da escola, número de alunos da escola, acessibilidade nas dependências da escola, biblioteca dentro das dependências, laboratório de informática também dentro das dependências, com acesso à internet ou não, número de computadores para uso dos alunos, sala de leitura aberta em período integral, área de lazer nas dependências da escola, atividade complementar tornando o ensino integral e as taxas de distorção idade-série, aprovação, reprovação e abandono.

Usando a medida de Rand foi obtido o número de grupos igual a 5, após uma análise de agrupamento as 79 escolas foram separadas em 5 grupos com características semelhantes, para que o modelo com parâmetros estimados seja o mais eficiente possível dentro da realidade de cada grupo distinto.

Com uma análise de regressão linear múltipla e a técnica *Stepwise* foram selecionadas as variáveis com significância estatística igual ou inferios a $\alpha = 0,10$ para o modelo e descartadas as variáveis que não possuiam significância estatística.

3.6 Modelo Empregado

A partir das variáveis significativas deu-se início à seleção do modelo para cada grupo distinto utilizando a técnica *Backward* no método da Regressão Multinível com Inferência Bayesiana nos parâmetros.

Para cada modelo encontrado, foram observados os valores do Critério de Informação do Desvio (DIC) e a quantidade de parâmetros estimados (p_{Di}), sendo escolhido o modelo com menores valores.

4 Resultados e Discussão

Primeiramente realizamos uma análise multivariada, por se tratar de um estudo de relações de múltiplas variáveis, com aplicação da análise de agrupamento - para que as observações com características em comum sejam analisadas de modo semelhante - utilizando as 14 variáveis (número de funcionários, número de alunos, acessibilidade nas dependências, biblioteca, laboratório de informática, acesso à internet, número de computadores, sala de leitura, área de lazer, atividade complementar e as taxas de distorção idade-série, aprovação, reprovação e abandono).

A partir do índice de Rand e do índice de Rand ajustado, observamos que o número de grupos ideal foi no intervalo de 5 - 6 onde os índices se aproximam do valor 1, indicando um acordo quase completo entre as partições.

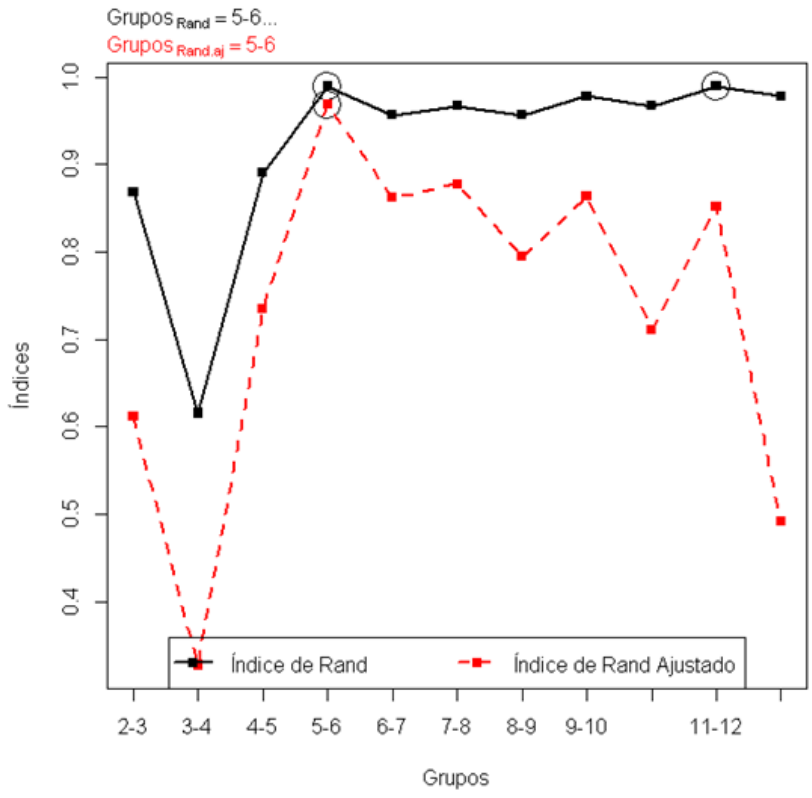

Figura 15: Índice de Rand e índice de Rand ajustado.

Para melhor serem trabalhados, foi escolhida a quantidade de 5 grupos distintos. As observações foram divididas com o método das médias das distâncias que tem como vantagem evitar valores extremos e levar em consideração toda a informação dos grupos, segundo Reis (1997) um grupo passa a ser definido como um conjunto de indivíduos no qual cada um tem mais semelhanças - em média - com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo.

Nesse caso em questão, uma das características em comum dentro dos grupos e incomum entre os grupos é a localização das escolas: o 1^o grupo é caracterizado por escolas rurais, o 2^o grupo tem sua maioria formada por escolas da região norte, o 3^o grupo possui escolas mais afastadas no centro da cidade - na região sul, o 4^o grupo possui mais observações na região leste e o 5^o grupo na região oeste. Outra característica similar dentro dos grupos e dissimilar entre eles é o uso dos laboratórios de informática e do acesso dos alunos à internet, tão como o número de alunos - escolas mais centrais tendem a ser maiores em quantidade de alunos. A divisão dos grupos é facilmente observada a partir do Dendograma - anexo C.

Em seguida, realizamos a Regressão Linear Múltipla e selecionamos as variáveis significativas com o método *Stepwise*, as variáveis **número de funcionários** e **número de alunos** foram transforamdas em classes de 0 (contendo de 12 à 29 funcionários e 20 à 39 alunos), 1 (contendo de 30 à 49 funcionário e 40 à 92 alunos) e 2 (contendo de 50 à 84 funcionários e 93 à 116 alunos), para serem melhor analisadas. Os resultados se encontram na Tabela 4.

Tabela 4: Variáveis selecionadas

	Estimativa	Erro Padrão	t-valor	Pr(>|t|)
Intercepto	-1,612886	0,772110	-2,089	0,04025 *
β_1	-0,209081	0,124172	-1,684	0,09655 .
β_2	0,190580	0,120882	1,577	0,11928
β_3	-0,390582	0,190860	-2,046	0,04437 *
β_4	0,207438	0,109649	1,892	0,06253 .
β_5	0,028093	0,009376	2,996	0,00375 **
β_6	6,719862	0,853173	7,876	2,61e-11 ***
Significância:	*** < 0,001	** < 0,01	* < 0,05	. < 0,1

A técnica indicou o melhor modelo como sendo aquele em que se verifica evidências ao nível de 0,1% de probabilidade que a variável β_6 (taxa de aprovação) é significativa, ao nível de 1% de probabilidade que a variável β_5 (número de computadores) é sinificativa, ao nível de 5% de probabilidade que a variável β_3 (atividade complementar) é significativa,

ao nível de 10% de probabilidade que as variáveis β_1 e β_4 (acessibilidade nas depêndencias e número de alunos) são significativas e que a variável β_2 (área de lazer) não é significativa. Os demais coeficientes das variáveis foram descartados pela técnica *Stepwise*.

Tendo os grupos distintos selecionados e as variáveis significativas para um modelo geral, partiu-se para a seleção dos modelos de cada grupo com o método *Backward*; começando com todas as variáveis existentes no modelo, testando as exclusões de cada variável e utilizando um critério de comparação para a escolha do modelo, excluindo a variável (se necessário) que melhore o modelo, esse procedimento é repetido até que essa melhoria não seja possível. O modelo de partida foi:

$$\mu_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{5i}x_{5i} + \beta_{6i}x_{6i} + \epsilon_i.$$

com $i = 1, 2, 3, 4, 5$. Onde β_1 =acessibilidade nas depêndencias, β_2 =área de lazer, β_3 =atividade complementar, β_4 =número de alunos (categorizado em três cateorias: 0 - menos de 20 alunos; 1 - entre 20 e 40; 2 - mais de 40), β_5 =número de computadores para uso dos alunos e β_6 =taxa dos aprovados. Os modelos selecionados se encontram na Tabela 5.

Como a diminuição de $0,1$ do modelo com menor DIC com 5 parâmetros não é satisfatória, o melhor modelo encontrado tem média $\mu_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{6i}x_{6i} + \epsilon_i$. Suas estimativas se encontram nas Tabelas 6 e 7.

Tabela 5: Modelos selecionados

Parâmetros	Modelos	DIC	p_{Di}
7	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{5i}x_{5i} + \beta_{6i}x_{6i}$	138,1	28,93
6	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{5i}x_{5i}$	178,5	21,4
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{6i}x_{6i}$	122,6	21,99
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{5i}x_{5i} + \beta_{6i}x_{6i}$	150,1	23,72
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{4i}x_{4i} + \beta_{5i}x_{5i} + \beta_{6i}x_{6i}$	131,9	25,12
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{5i}x_{5i} + \beta_{6i}x_{6i}$	130,2	23,27
	$\beta_{0i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{5i}x_{5i} + \beta_{6i}x_{6i}$	143,4	23,44
5	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i}$	176,7	18,91
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{6i}x_{6i}$	139,7	19,06
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{4i}x_{4i} + \beta_{6i}x_{6i}$	123,9	20,99
	$\beta_{0i} + \beta_{1i}x_{1i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{6i}x_{6i}$	122,5	18,72
	$\beta_{0i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \beta_{4i}x_{4i} + \beta_{6i}x_{6i}$	132,1	17,43

Observamos, com a Tabela 6, que a variância do modelo está baixa, $\hat{\sigma}^2 = 0,2151$, o que indica que o modelo foi bem ajustado. Os $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_3$ não foram significativos para nenhum dos grupos. O $\hat{\beta}_2$ foi significativo apenas para o 4^o grupo, $\hat{\beta}_{24} = 0,768$, indicando que o aumento de uma unidade dessa variável (área de lazer) aumenta o valor do IDEB em 0,77 no grupo 4, isto é, as escolas do grupo 4 que posuem área de lazer possuem

IDEB 7% superior, não descartando a possibilidade dos demais serem zero com 95% de credibilidade. O $\hat{\beta}_4$ do 2^o grupo foi significativo - $\hat{\beta}_{42} = 0,7018$ - indicando que o aumento de uma unidade dessa variável (número de alunos divididos em 3 classes) aumenta em 0,7 o valor do IDEB do grupo 2, e nao se descarta a possibilidade dos demais serem zero com com 95% de credibilidade. O $\hat{\beta}_6$ foi significativo para todos os grupos indicando que o aumento de uma unidade dessa variável (taxa de aprovação) aumenta em 6,42 o valor do IDEB para o 1^o grupo, em 6,6 para o 2^o grupo, em 6,52 para o 3^o grupo, em 6,53 para o 4^o grupo e em 5,67 para o 5^o grupo. Assim, por exemplo, para o 5^o grupo, uma escola com 100% de aprovação tem IDEB 57% superior à uma escola hipotética do mesmo grupo que tenha aprovação zero.

Tabela 6: Estimativas dos parâmetros do modelo

Parâmetro	Estimativas	Desvio	q 2,5%	Mediana	q 97,5%
σ^2	0,2151	0,05745	0,1488	0,2094	0,3094
β_{01}	-2,212	3,5	-14,0	-1,689	2,347
β_{02}	-1,721	0,956	-3,54	-1,711	0,1099
β_{03}	-3,061	4,573	-17,54	-1,785	1,081
β_{04}	-1,349	2,298	-6,2	-1,292	3,126
β_{05}	-1,039	1,04	-2,965	-1,062	0,8268
β_{11}	0,1364	0,3865	-0,6137	0,1346	0,8988
β_{12}	-0,3336	0,1938	-0,7182	-0,3323	0,04287
β_{13}	-0,1693	0,1907	-0,5358	-0,1717	0,2162
β_{14}	-0,6112	0,4518	-1,52	-0,6031	0,2561
β_{15}	-0,9308	0,3076	-1,533	-0,9288	-0,3317
β_{21}	0,1677	0,4655	-0,7666	0,1763	1,081
β_{22}	0,2108	0,1815	-0,1459	0,2118	0,5636
β_{23}	0,191	0,2174	-0,2335	0,1906	0,607
β_{24}	0,768	0,399	0,006209	0,7594	1,582
β_{25}	0,04389	0,2694	-0,4919	0,04498	0,5693
β_{31}	0,5139	3,299	-3,63	0,02809	12,32
β_{32}	-0,1594	0,1977	-0,5494	-0,1578	0,228
β_{33}	1,394	4,447	-2,269	0,1102	15,82
β_{34}	0,276	2,096	-3,691	0,1874	5,072
β_{35}	0,6785	0,5845	-0,4234	0,6784	1,808
β_{41}	-0,0512	0,4372	-0,9271	-0,05147	0,8024
β_{42}	0,7018	0,1771	0,3543	0,7009	1,047
β_{43}	0,1564	0,223	-0,285	0,1557	0,5849
β_{44}	0,2065	1,697	-2,164	0,1407	2,826
β_{45}	-0,3199	0,2113	-0,7318	-0,3206	0,1025
β_{61}	6,423	1,316	3,798	6,418	9,078
β_{62}	6,591	1,084	4,53	6,584	8,64
β_{63}	6,524	1,058	4,52	4,542	8,512
β_{64}	6,529	1,278	4,003	6,542	9,028
β_{65}	5,674	1,062	4,639	6,581	8,552

Na Tabela 7 verificamos que γ_6 é significativo, que indica que uma maior taxa de aprovação dá uma vantagem de 3,35 no IDEB em relação às outras variáveis que não diferem de zero. Os efeitos de u_0, u_1, u_2, u_3 e u_4 não são significativos para os grupos e o efeito de u_6 é significativo para os 5 grupos com com 95% de credibilidade, ou seja, o efeito da taxa de aprovação no 1^o grupo é 3,06, no 2^o grupo é 3,24, no 3^o grupo é 3,17, no 4^o grupo é 3,18 e no 5^o grupo é 3,22.

Tabela 7: Estimativas dos parâmetros do modelo

Parâmetro	Estimativa	Desvio	q 2,5%	Mediana	q 97,5%
γ_0	7,516	11,87	-1,908	-0,1495	31,45
γ_1	-0,1092	0,8575	-1,112	-0,1809	0,931
γ_2	0,8641	2,972	-0,7144	0,1792	14,08
γ_3	0,2678	1,605	-2,063	0,1096	5,522
γ_4	0,2703	1,367	-1,008	0,08309	3,807
γ_6	3,353	0,7769	1,94	3,329	5,025
u_{01}	-9,728	13,43	-43,84	-1,62	1,028
u_{02}	-9,237	11,98	-33,37	-1,573	0,5267
u_{03}	-10,58	15,11	-48,19	-1,635	1,011
u_{04}	-8,865	12,31	-36,89	-1,411	1,246
u_{05}	-8,555	11,86	-32,84	-1,026	1,157
u_{11}	0,2456	0,8994	-0,9352	0,2963	1,507
u_{12}	-0,2244	0,8708	-1,314	-0,1478	0,8287
u_{13}	-0,06003	0,8687	-1,147	0,009131	1,004
u_{14}	-0,502	0,9548	-1,854	-0,4284	0,7682
u_{15}	-0,8216	0,9043	-2,042	-0,7366	0,2739
u_{21}	-0,6964	3,03	-13,92	-0,01578	1,14
u_{22}	-0,6533	2,975	-13,81	0,0316	0,9662
u_{23}	-0,6732	2,983	-13,89	0,002532	0,9824
u_{24}	-0,09609	2,955	-13,13	0,5525	1,731
u_{25}	-0,8203	2,98	-13,89	-0,1399	0,8493
u_{31}	0,2461	2,884	-3,65	-0,07064	9,58
u_{32}	-0,4274	1,612	-5,726	-0,2647	1,955
u_{33}	1,127	4,337	-2,769	0,03488	15,17
u_{34}	0,008258	2,899	-6,678	0,05727	5,508
u_{35}	0,4108	1,636	-4,614	0,5043	2,963
u_{41}	-0,3215	1,435	-3,957	-0,125	1,159
u_{42}	0,4315	1,37	-3,031	0,607	1,759
u_{43}	-0,114	1,382	-3,548	0,07128	1,226
u_{44}	-0,06382	1,751	-2,8	0,03727	2,306
u_{45}	-0,5903	1,39	-4,072	-4,4016	0,7293
u_{61}	3,069	1,172	0,558	3,11	5,277
u_{62}	3,238	0,9953	1,135	3,243	5,2
u_{63}	3,171	0,943	1,235	3,201	4,996
u_{64}	3,176	1,189	0,6618	3,212	5,454
u_{65}	3,221	1,002	1,044	3,26	5,11

A Figura 16 traz os IDEBs preditos com bandas de 95% de confiança onde os pontos

são IDEBs observados. Podemos observar que o modelo foi bem ajustado aos parâmetros pois as observações se encontram dentro dos limites de aceitação e próximos à linha dos valores preditos, não havendo nenhum *outlier* aparente.

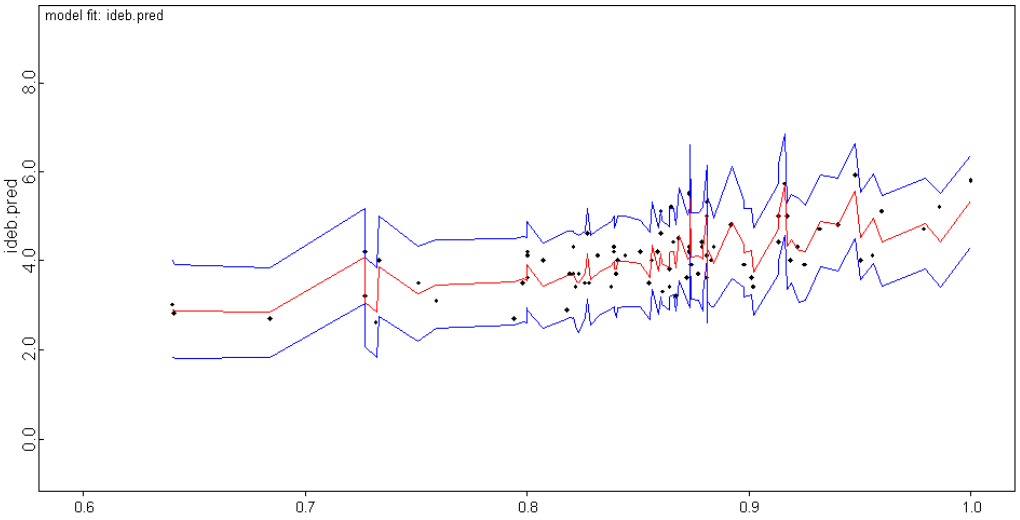

Figura 16: Limites de credibilidade para as observações preditas.

Por fim, a Figura 17 mostra os valores dos IDEBs preditos com o modelo escolhido, onde a média geral do IDEB é a linha e os box-plots são apresentados para o IDEB esperado de cada escola.

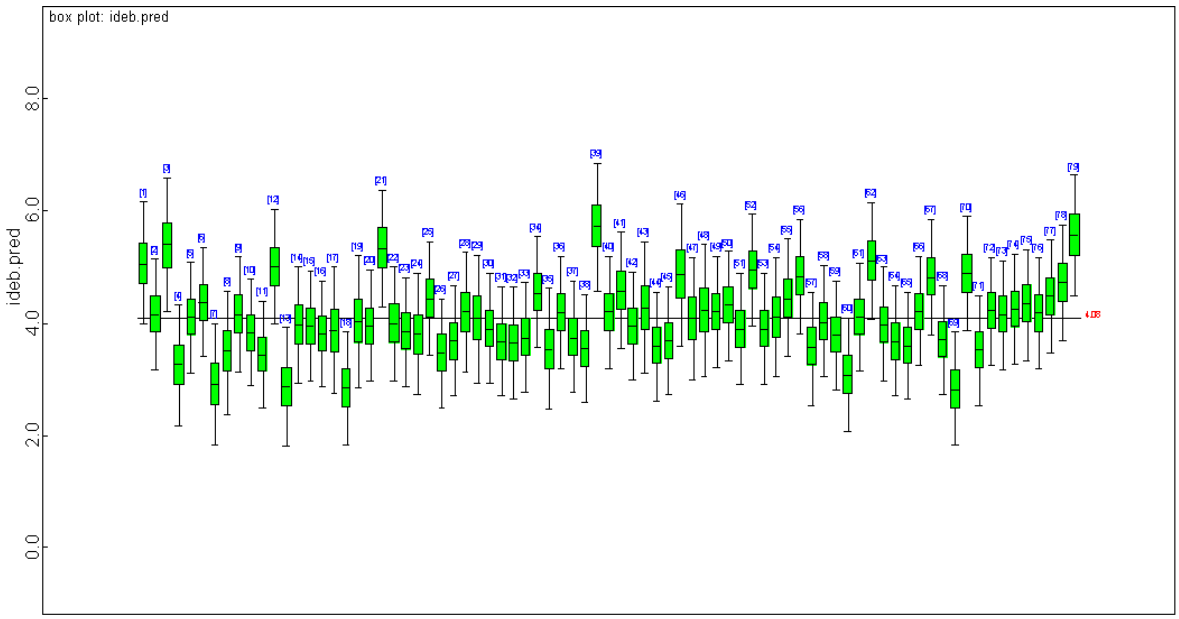

Figura 17: Boxplots dos valores preditos dos IDEBs.

Com os resultados obtidos, fica claro que um dos caminhos diretos para o aumento do IDEB do município de Campina Grande é o aumento da taxa de aprovação nas escolas. Segundo Silva et al. (2015) um aumento de 10% nas despesas públicas em educação, ocasiona uma melhora de 1,9% no IDEB. No entanto, sabe-se que aumentar a receita corrente municipal é algo muito difícil e independe da gestão pública local. Um outro caminho seria melhorando a infraestrutura das escolas com áreas de lazer e aumentando a capacidade física das mesmas, o que corrobora com Osio (2013) onde se comprova que as más condições das escolas afetam o desempenho dos alunos implicando em um rendimento escolar mais baixo. Os resultados expostos no texto concordam com Silva et al. (2015) e Osio (2013) já que a melhoria do IDEB pode ser obtida com o aumento da área de lazer e da taxa de aprovação, que por sua vez, são obtidas com o aumento dos gastos na educação. Também se ressalta nos estudos feitos pelo Instituto de Pesquisa Econômica Aplicada - IPEA (2010) que o futuro do Brasil depende do sucesso dos investimentos na política de educação para elevar a qualidade do desempenho dos alunos.

Importante considerar, ainda, que não são só as variáveis empíricas que influenciam no desempenho dos alunos acerca do IDEB, em nível nacional. Rabelo (2011) afirma que o processo não é neutro e sofre influência das relações sociais estabelecidas dentro e fora do contexto escolar e sugere aprofundar o caráter de diagnóstico das avaliações incluindo itens abertos, de resposta construída pelos estudantes, que avaliam competências distintas das associadas a itens objetivos.

5 Conclusões

Primeiramente, conclui-se, com o uso da Análise de Agrupamento e Índice de Rand, que as escolas do município de Campina Grande possuem características que permitem ser divididas em 5 grupos distintos, onde as características são semelhantes dentro dos grupos, esse conhecimento pode ser útil para possíveis estudos, avaliações e mudanças que visem a melhoria das escolas; implementação de áreas de lazer em um determinado grupo e atividade complementar em outro grupo - por exemplo - atendendo as necessidades específicas de cada um.

Após as análises de Regressão Múltipla verificamos que, das 14 variáveis utilizadas, 6 foram significativas para o estudo: acessibilidade nas dependências da escola, área de lazer nas dependências da escola, atividade complementar tornando o ensino integral, número de alunos da escola, número de computadores para uso dos alunos e taxa de aprovação. Independente de qual grupo esteja.

Aplicando o método de *Backward* em um modelo de Regressão Multinível com Inferência Bayesiana nos Parâmetros concluimos que o modelo com 6 parâmetros é suficiente para aumentar o valor do Índice de Desenvolvimento da Educação Básica (IDEB) nas escolas municipais de campina Grande, ou seja, visando esse objetivo, as escolas podem se ater ou investir mais nas variáveis: β_2= área de lazer, β_3= atividade complementar, β_4= número de alunos e β_6= taxa de aprovados, para um aumento efetivo do índice.

As análises aqui apresentadas têm o objetivo de nortear ações pedagógicas e auxiliar gestores escolares em suas tomadas de decisões visando o melhoramento da qualidade de ensino, medida a partir do IDEB. Os resultados são referentes ao IDEB aplicado no 5^o ano do ensino fundamental, lembramos que o IDEB também se aplica aos alunos do 9^o ano do ensino fundamental, ficando - sua análise - como sugestão para trabalhos futuros, tal como, a análise do IDEB do estado da Paraíba em comparação com os demais estados do Brasil.

Referências

AFONSO, A. J. **Avaliação educacional:** regulação e emancipação: para uma sociologia das políticas avaliativas contemporâneas. 4. ed. São Paulo. Cortez. 2009.

ALBERT, J. **Bayesian computation with R.** *Springer Science & Business Media*, 2007.

ARRABAL, C. T. **Estimação Clássica e Bayesiana para relação espécie-área com Distribuições Truncadas no Zero**. 2012.

COELHO-BARROS, E. A. et al. **Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos.** *Revista Colombiana de Estadística*, v. 31, n. 1, p. 111-129, 2008.

BURNHAM, K. P.; ANDERSON, D.R. Multimodel inference: Understanding AIC and BIC in model selection, **Sociological Methods Reseach.** London, v. 33, n. 2. p. 261-304. 2004.

BRASIL. Presidência da República. Constituição da república federativa do Brasil (1988). Disponível em: <http://www.planalto.gov.br/ccivil_ 03/Constituicao/Constitui%C3%A 7ao.htm>. Acesso em: 9 dez. 2015.

COSTA, W. U. Técnicas Bayesianas para Engenharia Elétrica. Minas Gerais, jan. de 2004.

DEMÉTRIO, C.G.B.; ZOCCHI, S.S.; **Modelos de Regressão**. Departamento de Ciências Exatas, ESALQ, USP, Piracicaba, 2008.

DRAPER, N. R.; SMITH, H. **Applied regression analysis by example.** 3. ed. New York: Wiley, 1998. 706p.

EDUCACENSO. **Censo Escolar da Educação Básica.** 2007. Disponível em: <http:// portal.inep.gov.br/basica-censo>. Acesso em: 9 dez. 2015.

FERREIRA, D. F. **Estatística Multivariada.** 1. ed. Lavras: UFLA, 2008.

FREITAS, D. N. T. **Avaliação da educação básica e ação normativa federal.** Cadernos de Pesquisa, v. 34, n. 123, p. 663-689, 2004.

FREITAS, D. N. T. **A avaliação da educação básica no Brasil:** dimensão normativa, pedagógica e educativa. Campinas, SP: Autores Associados, 2007.

GOLDSTEIN, H. **Multilevel statistical models.** 2. ed. London: Institute of Education, University of London. 1995.

GUSMÃO, J. B. B. **Qualidade da Educação no Brasil:** consenso e diversidade de significados. 2010. Dissertação (Mestrado). Faculdade de Educação da Universidade de São Paulo. São Paulo. 2010. 180p.

INEP. **Índice de Desenvolvimento da Educação Básica (Ideb).** 2007. Disponível em: <http://portal.inep.gov.br/web/portal-ideb/o-que-e-o-ideb>. Acesso em: 9 dez. 2015.

INEP. **Sistema de Avaliação da Educação Básica (Saeb).**Disponível em: <http://portal.inep.gov.br/web/saeb/aneb-e-anresc>. Acesso em: 9 dez. 2015.

EM DESENVOLVIMENTO, IPEA Brasil. **Estado, planejamento e políticas públicas.** Brasília: IPEA, edição, 2010.

MALAKOFF, D. **Bayes offers a 'new' way to make sense of numbers.** *Science*, v. 286, n. 5444, p. 1460-1464, 1999.

MARQUES, J. M.; CHAVES, A. **Análise de Agrupamentos (*CLUSTER*).** 2015.

MATELUNA, D. I. G. **Uso do WinBugs/OpenBugs.** Jul, 2012. São Paulo. 41p. Notas de aula.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Editora UFMG, 2005.

NASCIMENTO, M. R.; SILVA, A. F. Desdobramentos do índice de desenvolvimento da educação báscia (IDEB) nas práticas de docentes de uma escola municipal do agreste paraibano. **A crise do capitalismo e seus impactos na educação pública brasileira**, Campina grande, 2014. Disponível em: <http://www.escavador.com/sobre/7819349/andreia-ferreira-da-silva>. Acesso em: 30 nov. 2015.

OSIO, M. M. G. **Análise de modelos de regressão multiníveis simétricos.** São Carlos - SP. 2013.

RABELO, M. L. **ANÁLISE COMPARATIVA DOS PROCESSOS DE AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA.** In: *Colóquio de Matemática da Região Centro-Oeste*, 2. Mato Grosso. 2011.

RAND, W. M. **Objective criteria for the evaluation of clustering methods.** *Jour-*

nal of the American Statistical association, v. 66, n. 336, p. 846-850, 1971.

RÊGO, N. L. **Modelo de regressão linear múltipla com variável dummy: um estudo de caso.** 2014.

REIS, E. **Estatística multivariada aplicada**. Lisboa: Edições Silabo, 1997. 342p.

SANTOS, C. A. S. T. et al. **Modelagem multinível.** *Sitientibus*, Feira de Santana, n. 22, p. 89-98, 2000.

DA SILVA, A. B. et al. **Determinantes do IDEB: um estudo empírico com a Receita Corrente Líquida Municipal e a Despesa Pública em Educação.** 2015.

SILVA, A. F. Índice de desenvolvimento da educação básica (IDEB): avaliação estandardizada, organização escolar e trabalho docente. **Comunicação, Estado e políticas educacionais**, Campina Grande, 2009. Disponível em: <http://www.anaisdosimposio.fe.ufg.br/up/248/o/1.3._ 32>. Acesso em: 30 nov. 2015.

THE CYCLOPS PROJECT. **Análise de Agrupamentos.** Disponível em: <http: inf.ufsc.br/ patrec/agrupamentos.html>. Acesso em: 11 mar. 2016.

WAISELFISZ, J. **Sistemas de avaliação do desempenho escolar e políticas públicas.** *Ensaio.* Rio de Janeiro, v. 1, n. 1. 1993.

Apêndices

APÊNDICE A - Script para análise no software R

```
# pacotes necessarios #
library(cluster)
library(vegan)
library(ecodist)
library(MASS)
library(pvclust)
library(fpc)

    options(digits=3)

    # carregar banco de dados#

    # construindo o dendograma e separando grupos #
m =cbind(acess, bib, labinf, sleitura, alazer, atcomplem, funct, alunost, internet,
pcs, distorc, aprov, reprov, aband)
HCidx=function(m,k,hang,texto,method,distanc,obj,border,auto,alpha,arr,press)
is.int = function(num)
if( ceiling(num) != floor(num) ) out = FALSE
else out = TRUE
return(out)
if(alpha<0 | alpha>1)
alarm()
stop('alpha deve estar entre 0 e 1')
if(is.numeric(m)==FALSE)
alarm()
stop('m deve ser numérico!')
if( k<1 | is.numeric(k)==FALSE | (is.int(k)==FALSE & auto==TRUE) )alarm()
stop('Valor de k inválido!')
```

```
if(method<1 | method>7 | is.numeric(method)==FALSE | is.int(method)==FALSE)alarm()
stop('Método Inválido!')
if(distanc<1 | distanc>13 | is.numeric(distanc)==FALSE |
is.integer(distanc)==TRUE)alarm()
stop('Distância Inválida!')
distancia = function(distanc)
switch( distanc,
'1' = 'euclidean', '2' = 'maximum',
'3' = 'manhattan', '4' = 'canberra',
'5' = 'binary', '6' = 'minkowski',
'7' = 'bray-curtis', '8' = 'mahalanobis',
'9' = 'jaccard', '10'= 'simple',
'11'= 'difference', '12'= 'sorensen',
'13'= 'Partial')
metodo = function(method)
switch( method,
'1' = 'ward', '2' = 'single',
'3' = 'complete', '4' = 'average',
'5' = 'mcquitty', '6' = 'median',
'7' = 'centroid')
titulo = function(title)
switch( title,
'1' = 'Método Ward', '2' = 'Método Ligação Simples',
'3' = 'Método Ligação Completa', '4' = 'Método Média das Distâncias',
'5' = 'Método Mcquitty', '6' = 'Método Mediana',
'7' = 'Método Centróide')
N = nrow(m) ; p = ncol(m)
# if(N>p)  t = t(m) ; n = p
# if(N<=p) t = t(m) ; n = p
t = t(m) ; n = p
if( (distanc >= 1 & distanc <= 6)==TRUE )
dist = dist(t, method=distancia(distanc))
if( (distanc >= 7 & distanc <= 13)==TRUE )
dist = distance(t, method=distancia(distanc))
s = as.vector(dist)
hc = hclust(dist, method=metodo(method) ) ; print(hc);
title = method
```

```
if(obj=='TRUE' || obj=='T')
windows()
par(bty='o',mar=c(4.4, 4.3, 2.0, 0.5),font=6,cex.axis=1.2,cex.lab=1.5,fg=1,lwd=2)
if( is.character(texto) )
plot(hc,hang=hang,sub='',main=texto,xlab='Parcelas',ylab='Altura',lwd=2)
if( !is.character(texto) )
plot(hc,hang=hang,sub='',main=titulo(title),xlab='Parcelas',ylab='Altura',lwd=2)
if(obj=='FALSE' || obj=='F')
windows()
par(bty='o',mar=c(4.4, 4.3, 2.0, 0.5),font=6,cex.axis=1.2,cex.lab=1.5,fg=1,lwd=2)
if( is.character(texto) )
plot(hc,hang=hang,sub='',main=texto,xlab='Parcelas',ylab='Altura',lwd=2)
if( !is.character(texto) )
plot(hc,hang=hang,sub='',main=titulo(title),xlab='Parcelas',ylab='Altura',lwd=2)
plot(hc,hang=hang,sub='',main=titulo(title),xlab='Parcelas',ylab='Altura',lwd=2,lab
els=FALSE)
if(auto=='FALSE' || auto=='F')
h=locator(1)$y
list = rect.hclust(hc, h=h, border="black")
crisp = cutree(hc, h=h)
k = max(crisp)
card = numeric(k)
cat('\ n————————————————————————————\ n')
cat('corte: h =', h)
cat('\ n————————————————————————————-')
if(auto=='TRUE' || auto=='T')
list = rect.hclust(hc, k=k, border="black")
crisp = cutree(hc, k=k)
card = numeric(k)
# ———————————— silhueta ————————————-
slt = silhouette(crisp,dist)
windows()
plot(slt,main=,xlab=expression("Silhueta "*s[i]))
# ——————————————————————————————————-
# cluster.stats(dist, crisp1, crisp2)
mat = matrix(nrow=n-2,ncol=6)
idx = function(crisp1,crisp2)
```

```
x = crisp1 ; y = crisp2
xx = outer(x, x, "==");
yy = outer(y, y, "==") ; upper = row(xx) < col(xx) ; xx = xx[upper];
yy = yy[upper] ; a = sum(as.numeric(xx & yy));
b = sum(as.numeric(xx & !yy)) ; c = sum(as.numeric(!xx & yy));
d = sum(as.numeric(!xx & !yy)) ; ni = (b + a) ; nj = (c + a);
M = a + b + c + d ; q = (ni * nj)/M;
rand.aj=(a - q)/((ni + nj)/2 - q) ; rand = (a + d) / M
return(list(rand=rand,rand.aj=rand.aj))
for(i in 2:(n-2))
crisp1 = as.vector(cutree(hc, k=i))
slt = silhouette(crisp1,dist)
mat[i-1,2] = mean(slt[,3])
crisp2 = as.vector(cutree(hc, k=i+1))
mat[i-1,3] = i ; mat[i-1,4] = i+1
mat[i-1,5] = idx(crisp1,crisp2)$rand
mat[i-1,6] = idx(crisp1,crisp2)$rand.aj
colnames(mat) = c("N.Grupos","Silhueta","G1","G2","Rand","Rand-aj.")
# rownames(mat) = c(rep(,9) )
mat[,1] = seq(2,n-1)
crisp1 = as.vector(cutree(hc, k=n-1))
slt = silhouette(crisp1,dist)
max.silh = max(slt[,3]) ; min.silh = min(slt[,3])
mat[n-2,2] = mean(slt[,3])
max.idx = which(mat[,2] == max(mat[,2]), arr.ind = TRUE)
windows()
plot(mat[,1],mat[,2],xlab="Número de grupos",ylab="Silhueta média",
type="n",pch=22,lwd=2,bg=2,col=1)
points(mat[max.idx,1],mat[max.idx,2],pch=1,cex=3,col="black")


    matplot(mat[,1],mat[,2],type="o",pch=22,lwd=2,bg=2,col=1,add=T)
mtext( bquote(bar(s)[max]==.(max(slt[,3]) )), adj=0)
mtext( bquote(grupos==.( mat[max.idx,1] )), adj=1)
xx = NULL
for(i in 1:(n-3))
xx[i] = paste(mat[i,3],mat[i,4],sep=-")
```

```
xx = as.factor(xx)
y1 = as.vector(mat[-(n-2),5]) ; y2 = as.vector(mat[-(n-2),6])

    rand.max = which(mat[,5]==max(mat[-(n-2),5]), arr.ind=T)
rand.aj.max = which(mat[,6]==max(mat[-(n-2),6]), arr.ind=T)

    windows()
matplot(1:(n-3),cbind(y1,y2),xlab="Grupos",ylab="Índices",
type="n",xaxt="n", pch=15, lwd=2)
if(length(rand.max)==1)
mtext( bquote(Grupos[ Rand]==.(
as.character( xx[rand.max]) )), adj=0, line=1.2)
if(length(rand.max)>1)
mtext( bquote(paste(
Grupos[ Rand]==.(as.character( xx[rand.max] )),"...") ), adj=0, line=1.2)

    if(length(rand.aj.max)==1)
mtext( bquote(Grupos[ Rand.aj]==.( as.character(xx[rand.aj.max[1]]) )),
adj=0,col=2,line=0)
if(length(rand.aj.max)>1)
mtext( bquote(paste(
Grupos[ Rand.aj]==.(as.character( xx[rand.aj.max] )),"...") ),
adj=0, line=0, col=2)
points(rand.max,y1[rand.max],pch=1,cex=3,col="black")
points(rand.aj.max,y2[rand.aj.max],pch=1,cex=3,col="black")
matplot(1:(n-3),cbind(y1,y2),xlab="Grupos",ylab="?ndices",
type="o",xaxt="n", pch=15, lwd=2, add=T)
axis(1, at=1:(n-3), labels=xx)
legend("bottom",ncol=2,
legend=c("Índice de Rand","Índice de Rand Ajustado"),
col=c(1,2),pch=c(15,15),lwd=c(2,2),lty=c(1,2))
mat = round(mat,arr)
mat[rand.max,5] = paste(mat[rand.max,5],"*",sep=)
mat[rand.aj.max,6] = paste(mat[rand.aj.max,6],"*",sep=)
mat[max.idx,2] = paste(mat[max.idx,2],"*",sep=)
```

```
mat[(n-2),3:6] =
mat[-max.idx,2] = paste( mat[-max.idx,2], , sep=)
mat[-rand.max,5] = paste( mat[-rand.max,5], , sep=)
mat[-rand.aj.max,6] = paste( mat[-rand.aj.max,6], , sep=)
mat = data.frame( mat, stringsAsFactors=T, row.names=NULL)
cat(-————————————————————————-\ n ")
cat("Silhueta – Rand – Rand Ajustado \ n")
cat(-————————————————————————\ n")
print(mat)
cat(-————————————————————\ n \ n")


    distmat = as.matrix(dist)
for(c in 1:k) card[c] = length(list[[c]])
alt = vector()
for(i in 1:n)
if(i==1) alt[1] = min( distmat[,1][distmat[,1]!=0], Inf )
if(i==n) alt[n] = min( distmat[n,][distmat[n,]!=0], Inf )
else
alt[i]=min(distmat[,i][distmat[,i]!=0],distmat[i,][distmat[i,]!=0],Inf)
for(i in 1:k)
l = 1 ; q = card[i] ; obj = list[[i]] ; d = alt[obj]
if(q>2)
var = var(d) ; sd = sqrt(var) ; suma = matrix(ncol=11,nrow=1)
suma[,7] = sd ; suma[,8] = var
summar = summary(d) ; cv = sd*100/summar[4]
suma[,10] = sd/sqrt(length(d)) ; suma[,11] = IQR(d)
for(l in 1:6) suma[,l] = as.numeric(summar[l])
suma[,9] = cv
suma = data.frame(suma,row.names=")
colnames(suma) = c('Min.','1st Qu.','Median','Mean','3rd
Qu.','Max.','sd','var','cv(%)','Ep','IQR')
cat('\ n \ n————- Grupo', i,'——— [',q,']
objetos————————')
out1 = matrix(nrow=1,ncol=card[i] ) ; out2=out1
out1[1,] = obj ; out2[1,] = alt[ list[[i]] ]
rownames(out1) = c('obj')
```

```
colnames(out1) = rep(”,each=card[i])
rownames(out2) = c('alt')
colnames(out2) = rep(”,each=card[i])
print(out1) ; print(out2)
cat('\ n')
print(suma)
cat('————————————————————————————\ n')
if(q<=2)
if(q==2)
cat('\ n \ n————- Grupo', i,'——— [',q,']
objetos————————\ n')
out1 = matrix(nrow=1,ncol=card[i] ) ; out2=out1
out1[1,] = obj ; out2[1,] = alt[ list[[i]] ]
rownames(out1) = c('obj')
colnames(out1) = rep(”,each=card[i])
rownames(out2) = c('alt')
colnames(out2) = rep(”,each=card[i])
print(out1) ; print(out2)
if(q==1)
cat('\ n \ n————- Grupo', i,'——— [',q,']
objeto————————')
out1 = matrix(nrow=1,ncol=card[i] ) ; out2=out1
out1[1,] = obj ; out2[1,] = alt[ list[[i]] ]
rownames(out1) = c('obj')
colnames(out1) = rep(”,each=card[i])
rownames(out2) = c('alt')
colnames(out2) = rep(”,each=card[i])
print(out1) ; print(out2)
cat('————————————————————————————\ n')
rm(d)
# ————————————sumário matriz similaridade————————————
var = var(s) ; sd = sqrt(var) ; out = matrix(ncol=11,nrow=1)
out[,7] = sd ; out[,8] = var ; out[,10] = sd/sqrt(length(s))
summar = summary(s) ; cv = sd*100/summar[4]
for(l in 1:6) out[,l] = as.numeric(summar[l])
out[,9] = cv ; out[,11] = IQR(s)
out = data.frame(out,row.names=”)
```

```
colnames(out) = c('Min.','1st Qu.','Median','Mean','3rd
Qu.','Max.','sd','var.','cv(%)','Ep','IQR')
cat('\n\n----------sumário matriz similaridade----------\n')
print(out) ; rm(out)
cat('------------------------------------------------\n')
# ----------sumário matriz cofenética----------
c = cophenetic(hc) ; cof = c
c.mean = mean(c) ; c.sd = sd(c)
s.mean = mean(s) ; s.sd = sd(s)
cc = (c.sd/c.mean)/(s.sd/s.mean)
c = as.vector(c) ; corr.s.c = cor(c, s)
var = var(c) ; sd = sqrt(var) ; out = matrix(ncol=11,nrow=1)
out[,7] = sd ; out[,8] = var ; out[,10] = sd/sqrt(length(c))
summar = summary(c) ; cv = sd*100/summar[4]
for(l in 1:6) out[,l] = as.numeric(summar[l])
out[,9] = cv ; out[,11] = IQR(c)
out = data.frame(out,row.names=")
colnames(out) = c('Min.','1st Qu.','Median','Mean','3rd
Qu.','Max.','sd','var.','cv(%)','Ep','IQR')
cat('\ n\ n----------sumário matriz cofenética----------\ n') print(out) ;
rm(out)
cat('------------------------------------------------\ n')
cat('\ n\n----------\ n')
cat(' ANOVA\ n')
mg = sum(alt)/n ; vg = var(alt)
medias = vector() ; vars = vector()
for(i in 1:k)
medias[i] = mean( alt[ list[[i]] ] )
vars[i] = var( alt[ list[[i]] ] )
dif = (alt - mg)^2 ; SQTot = sum(dif)
dif = (medias-mg)^2
for(i in 1:k) dif[i] = dif[i]*card[i]
SQEntre = sum(dif) ; QMEntre = SQEntre/(k-1)
for(i in 1:k)
aux = ( alt[ list[[i]] ] - medias[i] )^2
dif[i] = sum(aux)}
SQDentro = sum(dif) ; QMDentro = SQDentro/(n-k)
```

```
Fcal = QMEntre/QMDentro ; Ef = SQEntre/SQDentro
R2 = SQEntre/SQTot ; lambda = SQDentro/(SQEntre+SQDentro)
Ftab = qf(p=1-alpha, df1=k-1, df2=n-k, lower.tail = T)
pvalor = pf(q=Fcal, df1=k-1, df2=n-k, lower.tail = F)
xsup = qf(p=0.99, df1=k-1, df2=n-k, lower.tail = T)
prob = function(a, b, c=0, d=0, xlim, las=1, p1, p2,
curve.col, col1, col2, pos, arr)
if(a==0) a = 0.0000000000001
coord1.x = c(a, seq(a, b, 0.01), b)
coord1.y = c(c, df(seq(a, b, 0.01),p1,p2), d)
coord2.x = c(b, seq(b, xsup, 0.01), xsup)
coord2.y = c(c, df(seq(b, xsup, 0.01),p1,p2), d)
area = pf(q=b,df=p1,df2=p2)-pf(q=a,df=p1,df2=p2)
area = round(area,arr)
m = df(((a+b)/2),p1,p2)
a = 0 ; b = round(b, arr)
par(las=las,mar=c(5,4,4,2)+0.1,font=2,cex.axis=1,cex.lab=1,fg=1,lwd=1)
curve(df(x,p1,p2),xlim=xlim,ylab=expression(
italic(f)[x](x)),n=2000,
main=paste('F(',p1,',',p2,')'),
col=coord2.x,coord2.y,col=col2)
text((a+b)/2, m, labels=paste(area), font=2, pos=pos)
windows()
prob(a=0,b=Ftab,c=0,d=0,xlim=c(0,xsup),p1=k-1,p2=n-k,
curve.col=1,col1="white", col2="gray", pos=3, arr=arr)
if(pvalor>alpha)
legend("topright",col=c("green","white"),lty=c(1,1),legend=sapply(
c(bquote(Fcal == .(Fcal)), bquote(p-valor == .(pvalor))), as.expression))
if(pvalor>=0.01 && pvalor<0.05)
legend("topright",col=c("green","white"),lty=c(1,1),legend=sapply(
c(bquote(Fcal == .(Fcal)), bquote(p-valor == .(pvalor)* "* ")),
as.expression))
if(pvalor>=0.001 && pvalor<0.01)
legend("topright",col=c("green","white"),lty=c(1,1),legend=sapply(
c(bquote(Fcal == .(Fcal)), bquote(p-valor == .(pvalor)* "** ")),
as.expression))
if(pvalor<0.001)
```

```
legend("topright",col=c("green","white"),lty=c(1,1),legend=sapply(
c(bquote(Fcal == .(Fcal)), bquote(p-valor == .(pvalor)* "*** ")),
as.expression))
sign = function(pvalor,alpha)
if(pvalor>alpha)
conc =
if(pvalor>=0.01 && pvalor<0.05)
conc = "*"
if(pvalor>=0.001 && pvalor<0.01)
conc = "**"
if(pvalor<0.001)
conc = "***"
return(conc)
abline(v=Fcal, col='black')
out = matrix(0,nrow=3,ncol=8)
out[1,1]=round(SQEntre,arr) ; out[2,1]=round(SQDentro,arr)
out[3,1]=round(SQTot,arr)
out[1,2]=k-1 ; out[2,2]=n-k ; out[3,2]=n-1
out[1,3]=round(SQEntre/(k-1),arr)
out[2,3]=round(SQDentro/(n-k),arr)
out[1,4]=round(out[1,3]/out[2,3],arr)
out[1,5]=round(Ftab,arr) ; out[1,6]=round(pvalor,arr)
out[1,7]=sign(pvalor=pvalor,alpha=alpha)
out[1,8]=round(Ef,arr)
out = data.frame(out)
colnames(out) = c("SQ","GL","QM","F","Ftab","pvalue","sign","Efic")
rownames(out) = c("Entre","Dentro","Total")
out[3,3]=NA ; out[2:3,4:7]=NA ; out[-1,7:8]=NA
cat('--------------------------------------------------------\n')
print(out,na.print=)
cat('\ n')
if(Fcal<=Ftab) cat('Como Fcal(',Fcal,') <= Ftab(',Ftab,'), aceitamos H0 a',
100*alpha,'%.\ n\ n' )
if(Fcal>Ftab) cat('Como Fcal(',Fcal,') > Ftab(',Ftab,'), rejeitamos H0 a',
100*alpha,'%.\ n\ n' )
if(pvalor>alpha) cat("p-valor(",pvalor,") > alpha(",alpha,")\ n\ n")
if(pvalor<=alpha) cat("p-valor(",pvalor,") <= alpha(",alpha,")\ n\ n")
```

```
cat('R2 =', R2,'\ n')
cat('lambda =', lambda,'\ n')


    cat('-----------------------------------\ n\ n')
if(press==TRUE || press==T)
return(list(cardinalidade=card,grupos=k,alturas=alt,corr.s.c=corr.s.c,
cc=cc,Dist=dist,Cof=cof))
else
return(list(cardinalidade=card,grupos=k,alturas=alt,corr.s.c=corr.s.c,
cc=cc))
rm(list=c(k,hang,method,obj,border,auto,alpha,matriz,c,hc,R2,lambda,
card,alt,mf,summ,dist,distanc,d,l,list,suma,c,matriz,out1,out2,
dif,c.mean,c.sd,s.mean,s.sd,cc,sum.c,sum.s,alpha1,texto,
distorc,s,var,sd,out,summar,cv,out,coord1.x,coord1.y,
coord2.x,coord2.y,corr.s.c,cc,distmat,cof,press))}


    HCidx(m=m,k=5,hang=-1,texto=,method=4,distanc=1,obj=T,border=3,
auto=T,alpha=0.05,arr=3,press=F)
dist1<-distance(m,"euclidean") # euclidean bray-curtis man-hattan mahalanobis jaccard
simple difference sorensen Partial
clust1<-hclust(dist1,"average") # "ward", "single", "complete", "average", "mcquitty",
"median"or "centroid"
plot(clust1, labels=nome, cex=0.5, hang=-0.5, main=)
rect.hclust(clust1, k=5, border="red")


    grupo1 = cbind(
sort(clust1$order[1:6]),1
)
grupo2 = cbind(
sort(clust1$order[7:34]),2
)
grupo3 = cbind(
sort(clust1$order[35:59]),3
)
grupo4 = cbind(
```

```
sort(clust1$order[60:64]),4
)
grupo5 = cbind(
sort(clust1$order[65:79]),5
)


    grupos = rbind(grupo1, grupo2, grupo3, grupo4, grupo5)


    # seleção de variaveis#
aj = lm(ideb acess+bib+labinf+sleitura+alazer+atcomplem+funct+alunost+internet+pcs
+distorc+aprov+reprov)
aj = step(aj)
summary(aj)


    # modelo escolhido #
aj1 = lm(ideb acess+atcomplem+alunost+pcs+aprov)
summary(aj1)
```

APÊNDICE B - Script para análise no software OpenBugs

```
model
for( i in 1 : n )
ideb[i]   dnorm(mu[i], tau)


    # mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a4[gr[i]]*alunost[i] + a5[gr[i]]*pcs[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a4[gr[i]]*alunost[i] + a5[gr[i]]*pcs[i]


    mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a4[gr[i]]*alunost[i] + a6[gr[i]]*aprov[i]


    # mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a5[gr[i]]*pcs[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a4[gr[i]]*alunost[i] + a5[gr[i]]*pcs[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a3[gr[i]]*atcomplem[i] +
a4[gr[i]]*alunost[i] + a5[gr[i]]*pcs[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a2[gr[i]]*alazer[i] + a3[gr[i]]*atcomplem[i] +
a4[gr[i]]*alunost[i] + a5[gr[i]]*pcs[i] + a6[gr[i]]*aprov[i]


    # mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a4[gr[i]]*alunost[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a4[gr[i]]*alunost[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a3[gr[i]]*atcomplem[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a2[gr[i]]*alazer[i] +
a4[gr[i]]*alunost[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a1[gr[i]]*acess[i] + a3[gr[i]]*atcomplem[i] +
a4[gr[i]]*alunost[i] + a6[gr[i]]*aprov[i]
# mu[i] <- a0[gr[i]] + a2[gr[i]]*alazer[i] + a3[gr[i]]*atcomplem[i] +
a4[gr[i]]*alunost[i] + a6[gr[i]]*aprov[i]
```

```
    ideb.pred[i]   dnorm(mu[i], tau)


    for(j in 1:5)
a0[j] <- g0 + u0[j]
u0[j]   dnorm(mu.a0, tau.a0)
a1[j] <- g1 + u1[j]
u1[j]   dnorm(mu.a1, tau.a1)
a2[j] <- g2 + u2[j]
u2[j]   dnorm(mu.a2, tau.a2)
a3[j] <- g3 + u3[j]
u3[j]   dnorm(mu.a3, tau.a3)
a4[j] <- g4 + u4[j]
u4[j]   dnorm(mu.a4, tau.a4)
a5[j] <- g5 + u5[j]
u5[j]   dnorm(mu.a5, tau.a5)
a6[j] <- g6 + u6[j]
u6[j]   dnorm(mu.a6, tau.a6)
g0   dnorm(mu.a0, tau.a0)
g1   dnorm(mu.a1, tau.a1)
g2   dnorm(mu.a2, tau.a2)
g3   dnorm(mu.a3, tau.a3)
g4   dnorm(mu.a4, tau.a4)
g5   dnorm(mu.a5, tau.a5)
g6   dnorm(mu.a6, tau.a6)


    mu.a0   dnorm(0.0, 1.0E-3)
mu.a1   dnorm(0.0, 1.0E-3)
mu.a2   dnorm(0.0, 1.0E-3)
mu.a3   dnorm(0.0, 1.0E-3)
mu.a4   dnorm(0.0, 1.0E-3)
mu.a5   dnorm(0.0, 1.0E-3)
mu.a6   dnorm(0.0, 1.0E-3)


    tau.a0   dgamma(1, 1)
```

tau.a1 dgamma(1, 1)
tau.a2 dgamma(1, 1)
tau.a3 dgamma(1, 1)
tau.a4 dgamma(1, 1)
tau.a5 dgamma(1, 1)
tau.a6 dgamma(1, 1)

tau dgamma(0.1, 0.1)
sigma2 <- 1/tau

list(tau = 1, tau.a0 = 1, tau.a1=1)

list(n = 79,
ideb = c(5, 5.1, 5.5, 3.5, 3.7, 3.9, 3, 3.7, 4.6, 3.4, 4, 5.3, 2.8, 4, 4.2, 4.2, 4,
2.7, 4.4, 4.3, 5.8, 4.1, 3.2, 3.7, 5.1, 3.1, 3.7, 3.9, 4.2, 4.2, 3.7, 3.5, 4.1, 4,
3.4, 3.9, 3.4, 3.6, 5.7, 5.2, 4.5, 3.3, 4.1, 3.6, 4.3, 4.8, 4, 4.3, 4.3, 5, 3.4, 4.1,
4.1, 4.2, 5.2, 4.7, 3.5, 3.6, 4.2, 3.2, 3.9, 5, 4.3, 2.9, 3.5, 4.4, 4.8, 3.5, 2.6,
4.7, 2.7, 3.6, 4.6, 4.3, 4, 3.8, 4, 4.4, 5.9),
acess = c(1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1,
1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0,
1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0),
bib = c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,
0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0),
labinf = c(1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1,
0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1),
sleitura = c(1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0,
1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1),
alazer = c(1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0,
1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0),
atcomplem = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0,

1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
func = c(33, 50, 19, 29, 34, 32, 32, 45, 39, 16, 18, 31, 62, 73, 30, 46, 29, 47, 44,
29, 15, 44, 24, 17, 31, 23, 58, 52, 29, 26, 84, 55, 26, 17, 26, 23, 21, 55, 43, 13,
32, 47, 19, 28, 34, 23, 30, 19, 53, 36, 25, 37, 33, 29, 18, 28, 19, 20, 25, 12, 40,
23, 40, 16, 24, 53, 70, 47, 30, 20, 23, 57, 48, 68, 52, 33, 45, 44, 25),
alunos = c(20, 66, 21, 29, 21, 40, 33, 93, 35, 34, 23, 55, 51, 48, 35, 35, 23, 37, 51,
20, 31, 57, 29, 27, 31, 56, 54, 70, 20, 54, 55, 33, 45, 22, 40, 53, 36, 39, 116,
25, 65, 34, 34, 23, 26, 31, 39, 31, 43, 39, 42, 50, 51, 46, 21, 47, 24, 27, 35, 25,
17, 40, 41, 26, 24, 58, 45, 45, 28, 37, 28, 35, 63, 41, 47, 60, 50, 30, 64),
internet = c(1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1,
1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1,
0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1),
pcs = c(21, 0, 19, 12, 0, 12, 16, 16, 17, 0, 0, 0, 13, 18, 0, 8, 20, 2, 5, 0, 15, 17,
0, 19, 11, 9, 9, 5, 5, 10, 10, 0, 10, 9, 5, 10, 0, 11, 0, 0, 0, 10, 5, 0, 9, 16,
20, 5, 10, 10, 14, 0, 12, 15, 0, 16, 9, 0, 0, 0, 0, 16, 10, 0, 0, 0, 13, 11, 0, 15, 8,
10, 0, 8, 2, 0, 18, 17, 17),
distorc = c(0.14, 0.29, 0.25, 0.4, 0.34, 0.2, 0.49, 0.24, 0.31, 0.11, 0.24, 0.2, 0.42,
0.4, 0.18, 0.26, 0.44, 0.51, 0.26, 0.16, 0.05, 0.26, 0.2, 0.27, 0.22, 0.32, 0.31,
0.26, 0.21, 0.34, 0.28, 0.44, 0.36, 0.18, 0.22, 0.18, 0.27, 0.35, 0.23, 0.33, 0.28,
0.4, 0.19, 0.22, 0.28, 0.23, 0.36, 0.18, 0.23, 0.12, 0.2, 0.22, 0.21, 0.27, 0.19,
0.14, 0.33, 0.19, 0.2, 0.35, 0.36, 0.17, 0.21, 0.28, 0.24, 0.24, 0.22, 0.24, 0.14,
0.18, 0.29, 0.34, 0.33, 0.3, 0.3, 0.35, 0.26, 0.17, 0.09),
aprov = c(0.913, 0.86, 0.873, 0.751, 0.877, 0.898, 0.64, 0.823, 0.827, 0.864, 0.807,
0.881, 0.641, 0.841, 0.851, 0.839, 0.733, 0.684, 0.879, 0.884, 1, 0.844, 0.867, 0.84,
0.96, 0.759, 0.821, 0.925, 0.873, 0.8, 0.819, 0.855, 0.832, 0.95, 0.822, 0.898,
0.902, 0.8, 0.916, 0.865, 0.868, 0.861, 0.881, 0.881, 0.821, 0.892, 0.883, 0.922,
0.873, 0.917, 0.838, 0.956, 0.8, 0.727, 0.986, 0.979, 0.828, 0.872, 0.859, 0.727,
0.874, 0.881, 0.839, 0.818, 0.798, 0.866, 0.94, 0.826, 0.732, 0.932, 0.794, 0.901,
0.86, 0.879, 0.856, 0.864, 0.919, 0.913, 0.948),
reprov = c(0.087, 0.125, 0.117, 0.171, 0.092, 0.088, 0.238, 0.147, 0.144, 0.136,
0.174, 0.119, 0.256, 0.132, 0.143, 0.078, 0.237, 0.205, 0.091, 0.104, 0, 0.121, 0.123,
0.111, 0.04, 0.159, 0.179, 0.068, 0.116, 0.152, 0.131, 0.1, 0.129, 0.04, 0.178,
0.097, 0.098, 0.16, 0.084, 0.135, 0.132, 0.097, 0.097, 0.119, 0.148, 0.091, 0.073,
0.078, 0.127, 0.071, 0.131, 0.044, 0.164, 0.25, 0, 0.021, 0.112, 0.112, 0.107, 0.206,
0.092, 0.119, 0.128, 0.182, 0.194, 0.108, 0.021, 0.103, 0.24, 0.06, 0.187, 0.058,
0.118, 0.09, 0.129, 0.088, 0.076, 0.087, 0.052),
aband = c(0, 0.015, 0.01, 0.079, 0.031, 0.014, 0.122, 0.03, 0.029, 0, 0.02, 0, 0.103,

0.027, 0.006, 0.083, 0.03, 0.112, 0.03, 0.012, 0, 0.035, 0.01, 0.05, 0, 0.082, 0,
0.007, 0.01, 0.048, 0.05, 0.044, 0.039, 0.01, 0, 0.005, 0, 0.04, 0, 0, 0, 0.042,
0.022, 0, 0.032, 0.017, 0.044, 0, 0, 0.012, 0.031, 0, 0.035, 0.023, 0.014, 0, 0.06,
0.016, 0.034, 0.072, 0.034, 0, 0.033, 0, 0.007, 0.026, 0.039, 0.072, 0.028, 0.007,
0.019, 0.042, 0.022, 0.031, 0.015, 0.048, 0.005, 0, 0),
funct= c(1, 2, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 2, 2, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0,
2, 2, 0, 0, 2, 2, 0, 0, 0, 0, 0, 2, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 2, 1, 0, 1, 1, 0,
0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 2, 2, 1, 1, 0, 0, 2, 1, 2, 2, 1, 1, 1, 0),
alunost = c(0, 1, 0, 0, 0, 1, 0, 2, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1,
1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1),
gr = c(4, 2, 4, 3, 2, 3, 5, 5, 5, 2, 2, 2, 3, 5, 2, 3, 4, 2, 1, 2, 5, 5, 2, 4, 3, 3,
3, 1, 1, 3, 3, 2, 3, 3, 1, 3, 2, 3, 2, 2, 2, 3, 1, 2, 3, 5, 4, 1, 3, 3, 5, 2, 3, 5,
2, 5, 3, 2, 2, 2, 2, 5, 3, 2, 2, 2, 3, 3, 2, 5, 3, 3, 2, 3, 2, 2, 5, 5, 5)
)

APÊNDICE C - Dendograma das escolas divididas em grupos

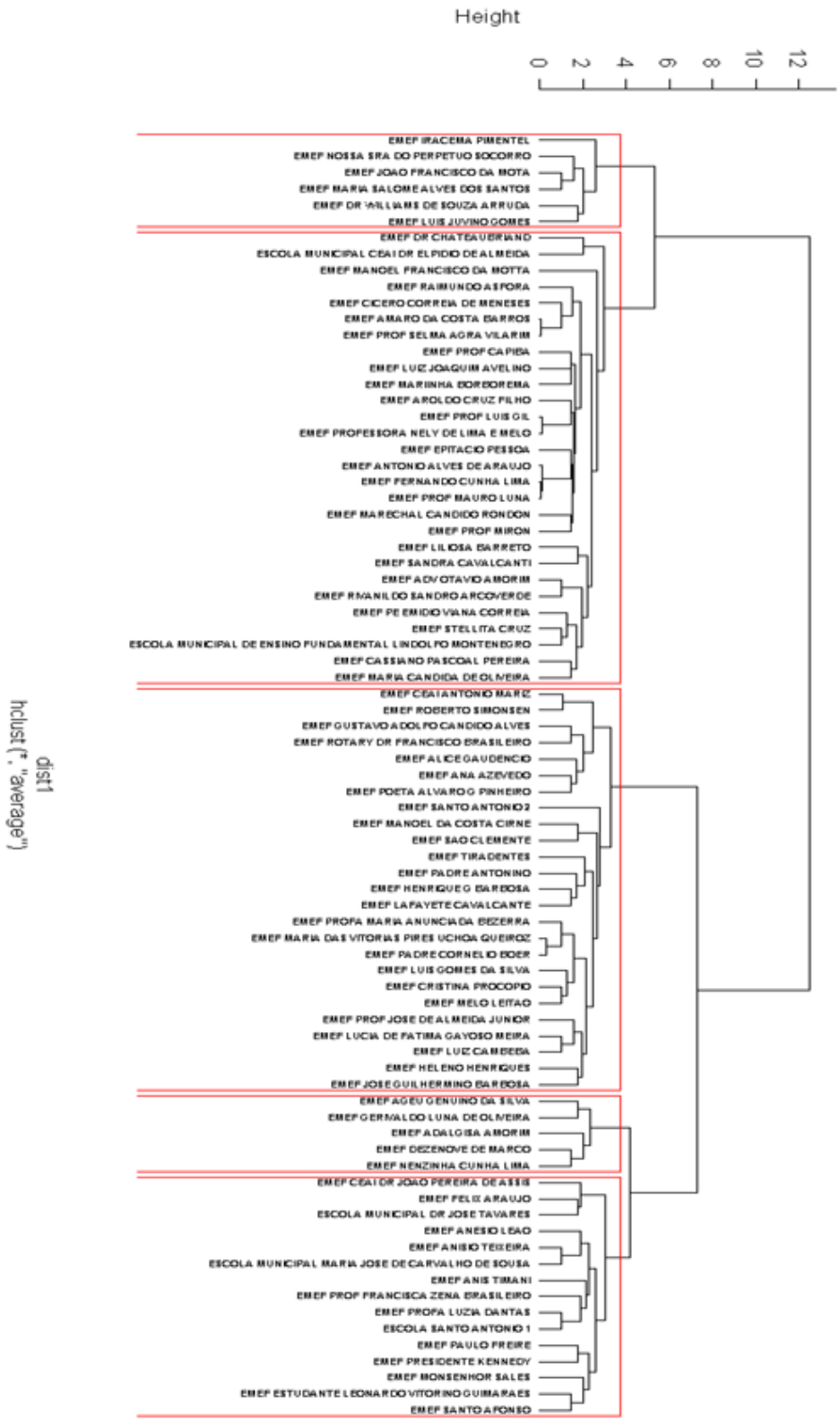

Figura 18: Dendograma do método das médias das distâncias.