REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 2675-3243

84 anos do IBGE

volume 77
número 242
julho / dezembro 2019

IBGE
Instituto Brasileiro de Geografia e Estatística

Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

Volume 77 - número 242 - jul/dez 2019

ISSN 2675-3243

R. Bras. Estat., Rio de Janeiro, v.77, n.242, p. 1-98, jul/dez 2019

Instituto Brasileiro de Geografia e Estatística – IBGE

@IBGE. 2019

Revista Brasileira de Estatística, ISSN 2675-3243

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE
Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.
Os originais para publicação deverão ser submetidos para o site http://rbes.net.br.
Os artigos submetidos à RBE não devem ter sido publicados em outros periódicos.
A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Informações Adicionais

Revista Brasileira de Estatística/IBGE - v.1, n.1 (jan/mar. 1940) – Rio de Janeiro: IBGE, 1940.v. trimestral (1940-1986), semestral (1987-).
Continuação de: Revista de economia e estatística. Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989). Co-edição com a Associação Brasileira de Estatística a partir do v.58.
ISSN publicação online 2675-3243, a partir de 2019.
I. Estatística – Periódicos. II. IBGE. III. Associação Brasileira de Estatística.

Gerência de Biblioteca e Acervos Especiais CDU 31(05)
RJ-IBGE/88-05 (ver. 2009) PERIÓDICO

Sumário

Nota do Editor

É com grande satisfação que anunciamos a publicação do volume 77 (número 242) da Revista Brasileira de Estatística, relativo ao período de Julho-Dezembro de 2019. Este volume traz quatro artigos com variada contribuição, em que são abordados temas envolvendo propostas de metodologia e aplicações diversas.

O primeiro artigo, de autoria de Cristiely Gomes Pires e Helton Saulo, traz um estudo da qualidade do ar em relação ao poluente atmosférico PM10, considerando uma avaliação estatística, por meio de cartas de controle por atributos, em que a concentração do poluente segue uma distribuição Birnbaum-Saunders. O segundo artigo, de autoria Fidel Henrique Fernandes e Renato Santos da Silva, traz uma análise e a discussão do desempenho dos estimadores μ e λ da distribuição simplex. O terceiro artigo, de autoria de Camila Cristina Lopes, Francisco Cribari-Neto e Tatiene Correia de Souza, traz a aplicação de um modelo de regressão beta para avaliar e comparar as magnitudes dos impactos que os gastos com programas assistenciais exerceram sobre as proporções de votos válidos recebidos pelo candidato governista nos segundos turnos das eleições presidenciais de 2006, 2010 e 2014. O quarto artigo, de autoria de Luiz Geraldo Mendes, traz um estudo que possibilita o mapeamento e análise da distribuição espacial das escolas públicas e privadas na Zona Urbana do Município de Rondonópolis – MT, correlacionando vulnerabilidade social.

Agradeço muito pela colaboração dos Editores Executivos Marcel de Toledo Vieira e Paulo Justiniano Ribeiro Junior, pelo seu apoio total e irrestrito na RBE e que foi fundamental à retomada da revista. Um agradecimento especial ao Professor Pedro Luis do Nascimento Silva, pela ajuda em vários processos da revista nesta etapa de retomada. Finalmente, agradeço aos editores associados e a todos revisores, que anonimamente contribuíram para mais este número da Revista Brasileira de Estatística.

Desejo a todos que tenham uma excelente leitura.

José André de Moura Brito.

Editor Responsável.

Revista Brasileira de Estatística, V.77, N. 242, p.7-32, Julho/Dezembro 2019

MONITORAMENTO DA QUALIDADE DO AR USANDO CARTAS DE CONTROLE POR ATRIBUTOS BASEADAS NA DISTRIBUIÇÃO BIRNBAUM-SAUNDERS

Cristiely Gomes Pires

cristiely.gomes.p@gmail.com

Instituto de Matemática e Estatística, Universidade Federal de Goiás, Goiânia, Brasil

Helton Saulo

heltonsaulo@gmail.com

Departamento de Estatística, Universidade de Brasília, Brasília, Brasil

Resumo: Este trabalho apresenta um estudo da qualidade do ar em relação ao poluente atmosférico PM10. Uma avaliação estatística foi feita por meio de cartas de controle por atributos em que a concentração do poluente segue uma distribuição Birnbaum-Saunders. Utilizou-se o algoritmo proposto por Leiva et al. (2015) no monitoramento das estações do Estado de São Paulo localizadas em Osasco, Santa Gertrudes e Cubatão - Vila Parisi, em janeiro e agosto de 2015. Constatou-se elevados níveis de poluição do PM10, especialmente no inverno e na área industrial, e concluiu-se que se faz necessário, tendo em vista a saúde da população e do meio ambiente, ações corretivas para amenizar os níveis de concentração de poluentes.
Palavras-chave: Cartas de controle por atributos; Distribuição Birnbaum-Saunders; Estado de São Paulo; PM10.

Abstract: This work presents a study of air quality concerning the PM10 air pollutant. A statistical analysis was made by means of attribute control charts when the pollutant concentration follows a Birnbaum-Saunders distribution. An algorithm proposed by Leiva et al. (2015) was used to assess pollution levels in three monitoring stations located in the State of Sao Paulo, Brazil, more specifically, stations located in Osasco, Santa Gertrudes and Cubatao-Vila Parisi, in January and August of 2015. High levels of PM10 pollution were observed, especially in the winter and in the industrial area. The results showed that, due to the health of the population and environment, it is necessary to take corrective actions to reduce pollutants concentrations.
Keywords: Control charts by atributes; Birnbaum-Saunders distribution; State of São Paulo; PM10.

1.INTRODUÇÃO

O desenvolvimento das grandes indústrias no período da Revolução Industrial provocou um grande aumento da concentração de poluentes na atmosfera, ocasionando no século passado seguidos episódios de mortes em algumas cidades da Europa e Estados Unidos, e de pânico na Região Metropolitana de São Paulo devido aos fortes odores, causando mal-estar e lotando os serviços médicos de emergência da região (Pinto, 2013; CETESB, 2014).

Indústrias, incineradores e veículos automotores associados ao clima, geografia, uso do solo, tipologia das fontes e às condições de emissão e dispersão local, são os principais fatores que contribuem para a concentração de poluentes no ar. Informações sistemáticas produzidas pelo monitoramento do ar nas áreas urbanas respaldam ações de fiscalização, controle e gestão da qualidade do ar.

Partículas ultrafinas de material sólido e líquido que ficam suspensas no ar, na forma de poeira, neblina, aerossol, fumaça e fuligem, são denominadas de material particulado (PM). Com especificações físicas de diâmetro aerodinâmico inferior ou semelhante a 10μm, subdividindo se em duas frações: PM2.5, a fração fina do PM com tamanho de até 2.5μm, e PM10, a fração grossa do PM com partículas de tamanho entre 2.5 e 10 μm. Neste estudo nos limitaremos às partículas inaláveis PM10 de fração grossa. Níveis de poluentes, como partículas inaláveis PM10, acima do padrão estabelecido podem desenvolver danos à vegetação, contaminação do solo e da água, além de sérios problemas na saúde da população, tais como, irritação dos olhos, nariz e da garganta, bronquite, pneumonia, doenças respiratórias crônicas, câncer de pulmão, problemas cardíacos, etc. (Marques & Dos Santos, 2012; CETESB, 2014; Pinto, 2013).

O Estado de São Paulo a partir de outubro de 2013 passou a ter como referência os padrões de qualidade do ar determinados no Decreto Estadual n^o 59113 de 23/04/2013, e não mais os padrões federais determinados na resolução CONAMA n^o 03 de 28/06/90. O decreto estabeleceu uma revisão dos valores-guia para os poluentes atmosféricos como parte de um conjunto de metas gradativas e progressivas baseadas nas diretrizes estabelecidas pela Organização Mundial de Saúde (CETESB, 2015).

O Decreto Estadual de São Paulo estabeleceu o nível máximo de 120 $\mu g/m^3$ de concentrações de PM10 permitido em 24 horas, sendo que, o padrão anual em média é 40 $\mu g/m^3$. O decreto classificou também níveis de poluente

PM10 de magnitude 250 $\mu g/m^3$ como estado de atenção, 420 $\mu g/m^3$ como estado de alerta e 500 $\mu g/m^3$ como estado de emergência. Além disso, o monitoramento da qualidade do ar deve ser conduzido por cartas de controle, mensurando o número de vezes em que as concentrações de PM10 alcançam o máximo tolerado pelo decreto.

Cartas de controle frequentemente são empregadas em técnicas de monitoramento de processos, as medidas de proporções de uma característica da qualidade em amostras selecionadas são grafadas a partir do processo versus tempo. A carta é composta por uma linha central (LC) e dois limites de controle, superior e inferior (LSC e LIC). A linha central representa o processo sob controle, ou seja, quando não há fontes excepcionais da variabilidade presente. Quando houver fontes de variabilidades excepcionais o gráfico do processo irá apresentar pontos que estarão fora dos limites de controles, fornecendo assim, um sinal para investigação do processo e suporte para ações corretivas da fonte de variabilidade excepcional (Montgomery, 2005).

Modelos probabilísticos com assimetria positiva vêm sendo empregados para descrever as concentrações dos poluentes no ar. Uma variável aleatória contínua positiva que mensura o tempo de vida até a ocorrência de um acontecimento de interesse, e o modelo probabilístico associado é frequentemente chamado de distribuição de vida. Birnbaum & Saunders, 1969 propuseram uma distribuição de vida associada com espécimes de materiais expostos à fadiga, produto de um padrão cíclico de tensão e força. O modelo descreve o tempo total decorrido até que o dano cumulativo, produzido pelo desenvolvimento e crescimento de uma ruptura, ultrapasse um valor limite e culmine na falha do material. Além do modelo proposto por Birnbaum e Saunders, outros modelos probabilísticos foram utilizados como distribuições de vida por Nelson & Hahn, 1972; Chhikara & Folks, 1977; Kalbfleish & Prentice, 1980; Bartlett & Kendall, 1946; Gumbel, 1958 e Barlow e Proschan, 1965.

Segundo Leiva, 2015 a distribuição Birnbaum-Saunders (BS) pode ser aplicada em diferentes campos e áreas do conhecimento, mais recentemente em negócios, criptografia, economia, finanças, indústria, seguro, inventário, nutrição, psicologia, controle de qualidade e toxicologia. O autor ainda cita algumas das aplicações inéditas da distribuição Birnbaum-Saunders como a migração de falhas de metálicos em nano-circuitos devido ao calor num chip de computador, a acumulação de substâncias deletérias nos pulmões vindas da poluição atmosférica, a ingestão por seres humanos de produtos químicos tóxicos de resíduos industriais, a diminuição da biomassa na pesca ao longo do

tempo em uma determinada zona e a ocorrência de desastres naturais, como terremotos e tsunamis.

Leiva et al., 2015 sugerem o modelo BS para descrever dados de concentração de poluentes baseados nos trabalhos de Ott, 1990 e Desmond, 1985, que discutem as propriedades do modelo log-normal para concentração de poluentes e condução ao modelo BS, respectivamente. A partir das propriedades semelhantes entre os modelos log-normal e BS, o estudo de Leiva et al., 2015 faz uma analogia e afirma, sobre determinadas condições, a adequação do modelo BS para os dados de concentração de poluentes.

Mesmo com o aumento de metodologias baseadas no modelo BS existem poucos trabalhos sobre ferramentas de monitoramento da qualidade do ar envolvendo esse modelo e apenas um sobre cartas de controle por atributos, i.e. Leiva et al., 2015, o qual utilizaremos como base neste trabalho; ver Lio et al., 2010; Aslam et al., 2011; Leiva et al., 2014 e Saulo et al., 2015. Neste artigo iremos supor que a concentração de poluentes PM10 segue a distribuição Birnbaum-Saunders. A partir desta suposição construiremos cartas de controle para monitorar a qualidade do ar de estações localizadas no estado de São Paulo.

Esse trabalho está estruturado em seis seções tendo início por essa introdução. Na Seção 2 realiza-se uma discussão sobre a distribuição BS e em seguida apresenta um relato breve sobre cartas de controle. Na Seção 4 são apresentadas cartas de controle para a distribuição BS. Na Seção 5 os resultados são apresentados. Por fim, na última seção tem-se as considerações finais.

2. A DISTRIBUIÇÃO BIRNBAUM-SAUNDERS

Seja T uma variável aleatória com distribuição BS com parâmetros α e β, em que α é um parâmetro de forma e β um parâmetro de escala e também a mediana da distribuição. Então, a função de distribuição acumulativa associada pode ser escrita como

$$F_T(t) = P(T \leq t) = \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right)\right], t > 0, \quad (1)$$

em que

$\alpha = \frac{\sigma}{\sqrt{\omega\mu}} > 0$ e $\beta = \frac{\omega}{\mu} > 0.$

Diz-se que T segue uma distribuição BS, $T \sim$ BS(α,β). A função de densidade de probabilidade de T é

$f_T(t, \alpha, \beta) = F_T^{`}(t).$

Consequentemente,

$$f_T(t;\alpha,\beta) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\beta}+\frac{\beta}{t}-2\right)\right)\frac{1}{2\alpha}\left(\left(\frac{t}{\beta}\right)^{-1/2}+\left(\frac{t}{\beta}\right)^{-3/2}\right)$$

$$= \frac{1}{\sqrt{8\pi}}\exp\left(\frac{1}{\alpha^2}\right)\exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\beta}+\frac{\beta}{t}\right)\right)\frac{t^{-3/2}}{\alpha\beta^{1/2}}(t+\beta), \quad (2)$$

para $t > 0$, $\alpha > 0$ e $\beta > 0$.

Para caracterizar a distribuição BS dada em (2) vamos definir a média, variância, coeficientes de assimetria e curtose. Logo, considere a seguinte transformação monótona:

$$X = \frac{1}{2}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right),$$

isolando T temos,

$$T = \beta\left\{1 + 2X^2 + 2X(1+X^2)^{1/2}\right\}.$$

De (1) temos que $X \sim N(0, \frac{1}{4}\alpha^2)$, desta maneira, obtemos o valor esperado, a variância, a assimetria e a curtose dados respectivamente, por

$$E(T) = \beta\left(1+\frac{1}{2}\alpha^2\right) \qquad Var(T) = (\alpha\beta)^2\left(1+\frac{5}{4}\alpha^2\right),$$

$$\mu_3 = \frac{16\alpha^2(11\alpha^2+6)}{(5\alpha^2+4)^3} \quad \text{e} \quad \mu_4 = 3 + \frac{6\alpha^2(93\alpha^2+41)}{(5\alpha^2+4)^2},$$

em que μ_3 e μ_4 medem a assimetria e o grau de achatamento da distribuição, respectivamente.

Outras duas quantidades importantes da distribuição BS são a função geradora de momentos e a taxa de falhas, dadas a seguir. O r-ésimo momento é dado por

$$E\left(\left(\frac{T}{\beta}\right)^r\right) = \sum_{j=0}^{n} 2r_{2j} \sum_{k=0}^{j} j_k \frac{(2(r-j+k))!}{2^{r-j+k}(r-j+k)!}\left(\frac{\alpha}{2}\right)^{2(r-j+k)}. \quad (3)$$

A taxa de falhas é dada por

$$h_T(t) = -\frac{d}{dt}\log(1-F_T(t)) = \frac{f_T(t)}{1-F_T(t)} = \frac{\phi\left(\frac{1}{\alpha}\left(\sqrt{\frac{t}{\beta}}-\sqrt{\frac{\beta}{t}}\right)\right)t^{-3/2}(t+\beta)}{2\alpha\beta^{1/2}\Phi\left(-\frac{1}{\alpha}\left(\sqrt{\frac{t}{\beta}}-\sqrt{\frac{\beta}{t}}\right)\right)}, \quad (4)$$

em que ϕ representa a função densidade de probabilidade da normal padrão. A taxa de falha h_T da BS tem as seguintes propriedades: (a) é unimodal para qualquer valor de α, crescente para $t < t_c$, e decrescente para $t > t_c$, em que t_c denota seu ponto de mudança; (b) converge para $1/(2\alpha^2\beta)$ quando $t \to \infty$; e (c) cresce quando $\alpha \to 0$ Kundu et al., 2008; Leiva, 2015.

A taxa de falhas média é dada por

$$v(t) = \frac{1}{t}\int_0^t h(s)ds,$$

em que $v(t)$ é quase não-decrescente e aproxima-se de uma constante quando $t \to \infty$.

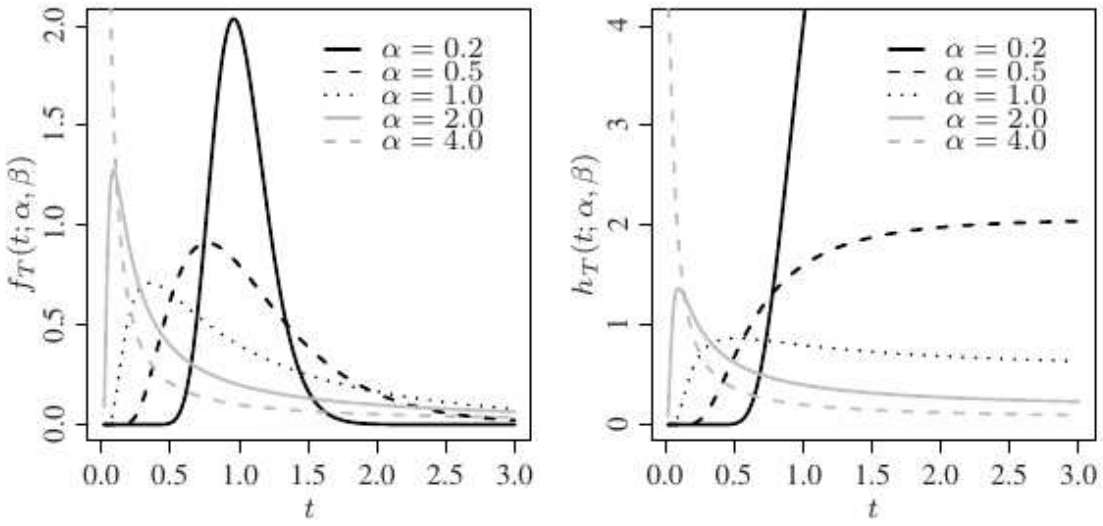
O parâmetro de escala $\beta > 0$ também é a mediana da distribuição BS e α é um parâmetro de forma. Quando α decresce para zero, a distribuição BS tende para a distribuição normal de média β e variância τ, em que $\tau \to 0$ quando $\alpha \to 0$.

A distribuição BS satisfaz as propriedades de proporcionalidade e reciprocidade; ver Leiva, 2015. Ou seja, a partir da função densidade de probabilidade da BS dada em (2), claramente se $T \sim \mathrm{BS}(\alpha, \beta)$, para todo $b > 0$, a variável aleatória $Y = bT$ pode ser modelado por uma distribuição BS com parâmetros α e $b\beta$. Além disso, a variável aleatória $Y = 1/T$ tem a mesma distribuição de T com o parâmetro β substituído por $1/\beta$. Especificamente, se $T \sim \mathrm{BS}(\alpha, \beta)$, então

(D1) $bT \sim \mathrm{BS}(\alpha, b\beta)$ para $b > 0$;
(D2) $\frac{1}{T} \sim \mathrm{BS}\left(\alpha, \frac{1}{\beta}\right)$; e
(D3) $Y = \log(T) \sim \log - \mathrm{BS}(\alpha, \log(\beta))$, no qual $\log - \mathrm{BS}(\alpha, \log(\beta))$ representa a distribuição logarítmica BS.

A Figura 1 (esquerda) mostra as densidades da distribuição BS para diferentes valores de α, com $\beta = 1$. Note que a distribuição BS é contínua, unimodal e possui assimetria positiva (assimetria à direita). Note também que, quando α tende a zero, a distribuição BS tende a ser simétrica em torno de β (a mediana da distribuição) e sua variabilidade é reduzida. Adicionalmente, quando α aumenta, a distribuição tem caudas mais pesadas. Desse modo, α altera não apenas sua forma, mas também sua assimetria e curtose. A Figura 1 (direita) mostra os gráficos da taxa de falha da BS para diferentes valores de α, com $\beta = 1$.

Figura 1 - Gráficos da densidade (esquerda) e taxa de falha (direita) do modelo BS



2.1 ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA

Seja $T_1, T_2, \ldots, T_n$ uma amostra aleatória de tamanho n da variável aleatória T com distribuição $\mathrm{BS}(\alpha, \beta)$, e seja $t_1, t_2, \ldots, t_n$ seus valores observados. A função de verossimilhança é dada por

$$L(\alpha,\beta) = \prod_{i=1}^{n} f(t_i;\alpha,\beta) \tag{5}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\alpha^2}\left(\frac{t_i}{\beta}+\frac{\beta}{t_i}-2\right)\right)\frac{1}{2\alpha}\left(\left(\frac{t_i}{\beta}\right)^{-1/2}+\left(\frac{t_i}{\beta}\right)^{-3/2}\right).$$

E o logaritmo da função de verossimilhança $l(\alpha,\beta)$ é dada da seguinte forma

$$l(\alpha,\beta) = \log\{L(\alpha,\beta)\}$$

$$= \sum_{i=1}^{n} log\left\{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\alpha^2}\left(\frac{t_i}{\beta}+\frac{\beta}{t_i}-2\right)\right)\frac{1}{2\alpha}\left(\left(\frac{t_i}{\beta}\right)^{-1/2}+\left(\frac{t_i}{\beta}\right)^{-3/2}\right)\right\}.$$

Por meio de maximização obtemos os estimadores de máxima verossimilhança de α e β, $\hat{\alpha}$ e $\hat{\beta}$, respectivamente. Para encontrar tais estimadores é necessário resolver o sistema de derivadas parciais

$$\dot{\ell}_\alpha = \frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = -\frac{2n}{\alpha^3} + \frac{1}{\alpha^3}\sum_{i=1}^{n}\left(\frac{t_i}{\beta}+\frac{\beta}{t_i}\right) - \frac{n}{\alpha} = 0, \tag{6}$$

$$\dot{\ell}_\beta = \frac{\partial \ell(\alpha,\beta)}{\partial \beta} = \frac{1}{2\alpha^2}\sum_{i=1}^{n}\left(\frac{t_i}{\beta^2}-\frac{1}{t_i}\right) - \frac{n}{2\beta} + \sum_{i=1}^{n}\frac{1}{t_i+\beta} = 0.$$

Portanto, temos

$$\hat{\alpha} = \left(\frac{s}{\hat{\beta}}+\frac{\hat{\beta}}{r}-2\right)^{1/2}, \tag{7}$$

em que

$$s = \frac{1}{n}\sum_{i=1}^{n} t_i, \quad r = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{t_i}\right)^{-1}, \tag{8}$$

são as médias aritmética e harmônica, respectivamente.

Para encontrar $\hat{\beta}$ é necessário resolver a equação:

$$\beta^2 - \beta(2r + \kappa(\beta)) + r(s + \kappa(\beta)) = 0,$$

em que

$$\kappa(\delta) = \left[\frac{1}{n}\sum_{i=1}^{n} (\delta + t_i)_{-1}\right]_{-1} \quad para \quad \delta > 0.$$

A distribuição conjunta assintótica dos estimadores de máxima verissimilhança $(\hat{\alpha}, \hat{\beta})$ é dada por

$$\sqrt{n}\begin{pmatrix}\hat{\alpha} - \alpha \\ \hat{\beta} - \beta\end{pmatrix} \sim N\left(\begin{matrix}0 \\ 0\end{matrix}, \begin{pmatrix}\frac{\alpha^2}{2} & 0 \\ 0 & \frac{\beta^2}{(0.25+\alpha^{-2}+I(\alpha))}\end{pmatrix}\right),$$

em que

$$I(\alpha) = 2\int_0^{\infty} \left(\frac{1}{1+1+(\alpha u)^2/2+\alpha u(1+(\alpha u)^2/4)^{1/2}} - \frac{1}{2}\right)^2 d\Phi(u).$$

2.2 ESTIMAÇÃO POR MOMENTOS MODIFICADOS

Estimadores de momentos convencionais não são recomendáveis para estimar os parâmetros da distribuição BS, pois eles podem não ser únicos e nem existirem Leiva, 2015. Alternativamente, podemos empregar o método dos momentos modificado, que fornece estimadores que sempre existem e são únicos (Leiva, 2015).

A seguir denota-se os estimadores de momentos modificados dos parâmetros α e β da distribuição BS e acrescenta-se suas distribuições assintóticas, que podem ser usadas para construir intervalos de confiança e testes de hipóteses para estes parâmetros.

Seja $T_1, T_2, \ldots T_n$ uma amostra aleatória de tamanho n da distribuição BS dada em (2), os estimadores de momentos modificados para α e β, que denotaremos por $\tilde{\alpha}$ e $\tilde{\beta}$ são

$$\tilde{\alpha} = \left\{2\left[\left(\frac{s}{r}\right)^{1/2} - 1\right]\right\}^{1/2}, \tag{9}$$

$$\tilde{\beta} = (sr)^{1/2},$$

em que r e s são como as médias aritmética e harmônica dadas em (8).

A distribuição assintótica conjunta de $\tilde{\alpha}$ e $\tilde{\beta}$ é dada por

$$\begin{pmatrix}\tilde{\alpha}\\ \tilde{\beta}\end{pmatrix} \sim N\left[\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{pmatrix}\frac{\alpha^2}{2n} & 0\\ 0 & \frac{(\alpha\beta)^2}{n}\left(\frac{1+\frac{3}{4}\alpha^2}{(1+\frac{1}{2}\alpha^2)^2}\right)\end{pmatrix}\right].$$

3. CARTAS DE CONTROLE

De acordo com Montgomery, 2005 as cartas de controle tem uma relação próxima com teste de hipóteses. Considere um teste de hipótese no qual H_0 testa $\mu_1 = \mu_0$ e H_1 testa $\mu_1 \neq \mu_0$, $\overline{x}$ como o estimador de μ_1 e μ_0 a média alvo do processo em questão, então, se $\overline{x}$ estiver entre os limites de controle, podemos concluir que a média do processo está sob controle e não rejeitamos H_0 e se $\overline{x}$ excede algum dos limites de controle, concluímos que a média do processo está fora de controle e rejeitamos H_0. Para um modelo geral de carta de controle suponha y uma estatística amostral, que mede a característica da qualidade de interesse, com média μ_y e desvio padrão σ_y, e k a distância dos limites de controle à linha central em unidades de desvio padrão, então o LIC, LC e LSC são calculados como:

$$LIC = \mu_y - k\sigma_y, \tag{10}$$

$$LC = \mu_y,$$

$$LSC = \mu_y + k\sigma_y.$$

Comumente se usa $k = 3$.

A partir da natureza dos dados no monitoramento da característica da qualidade pode-se classificar as cartas de controle por variáveis ou atributos. Para dados contínuos pode-se usar cartas para variáveis, como a média amostral $\overline{X}$, que monitora a centralidade, ou a amplitude (R), que monitora a dispersão, e para dados qualitativos usa-se cartas por atributos, para a fração de não conformes (p) ou para o número médio de não conformes (np). Neste estudo serão utilizadas cartas de controle baseadas no número médio de não conformidade (np) quando amostras de igual tamanho (n) são retiradas do processo (Leiva et al., 2015).

3.1 CARTAS DE CONTROLE POR ATRIBUTOS DE NP

Controle de processos por atributos é quando no monitoramento do processo atribuímos ao produto em inspeção a classificação “defeituoso” ou “não defeituoso”, mais recentemente a terminologia “conforme” e “não conforme” se

tornou mais popular. Servem para monitorar certa fração/número de itens defeituosos, ou seja, a quantidade de unidades do produto que não satisfazem uma ou mais das especificações daquele produto. São em geral menos informativos que os gráficos para variáveis, porém, são muito úteis nas indústrias de serviços e nos esforços de melhoria da qualidade não industrial (Costa et al., 2005; Montgomery, 2005).

Conforme Costa et al., 2005 orienta na construção de uma carta de np faz-se necessário de ante mão diagnosticar e eliminar as causas atribuíveis, que as unidades de inspeção sejam as mesmas para cada amostra e que as amostras sejam de tamanhos iguais. Tendo em vista que os fundamentos estatísticos das cartas de np são baseados na distribuição binomial, suponha p a probabilidade de que uma unidade não esteja de acordo com as especificações, ou seja, esteja não conforme, e que as unidades produzidas sejam independentes, então cada unidade produzida é uma realização da variável aleatória Bernoulli com parâmetro p.

Logo, se n unidades do produto são selecionadas e D representa o número de unidades não conformes nessa amostra, então D segue uma distribuição binomial np com função de probabilidade

$$P(D = d) = \binom{n}{d} p^d (1-p)^{n-d}, \quad d = 0,1 \dots, n, \tag{11}$$

com média $E(D) = np$ e variância $Var(D) = np(1-p)$.

Portanto, podemos calcular os limites de controle e a linha central, LIC, LSC e LC respectivamente, por

$$LIC = max\{0, np_0 - k\sqrt{np_0(1-p_0)}\},$$

$$LC = np_0,$$

$$LSC = np_0 + k\sqrt{np_0(1-p_0)}, \tag{12}$$

em que p_0 é o valor de p quando o processo está em controle estatístico.

Para a operação real da carta toma-se amostras subsequentes de tamanho n, calcula-se o número de não conformes por np_0 para cada amostra e grifa os na carta. Caso a carta não apresente um padrão não-aleatório ou np_0 esteja dentro dos limites de controle, considera-se que o processo opera sob controle estatístico. Caso apresente um padrão não aleatório ou np_0 esteja fora dos limites de controle, considera-se que o processo está fora de controle; ver (Montgomery, 2005).

4. CARTAS DE CONTROLE POR ATRIBUTOS PARA A BS

Leiva et al., 2015 e Saulo et al., 2015 propuseram uma metodologia baseada em cartas de controle para o monitoramento do risco ambiental quando a concentração de contaminantes segue uma distribuição BS. Os autores destacam a importância do monitoramento do risco ambiental para que a saúde humana seja protegida, argumentam o uso da carta de controle pela facilidade de interpretá-la e de ser atualizada sempre que são disponibilizados dados adicionais e defendem o emprego da distribuição BS, devido ao fato que dados de concentrações de contaminantes são frequentemente assimétricos (com assimetria positiva).

Para a implementação da carta de np baseada na BS parte-se de uma amostra $t_1, \ldots, t_n$ do poluente atmosférico T e verifica-se a probabilidade de exceder ou não um padrão t estabelecido pela regulamentação oficial. Considere a construção da carta de controle de np dado na Seção 3.1, assim, D representa o número de vezes que a concentração desse poluente excedeu o valor determinado. A probabilidade p calculada pela média da variável aleatória T com distribuição contínua BS(α, μ) temos que $D \sim Binomial(n, p)$ com equação dada em (11) e a carta pode ser construída como em (12). Na qual k é um coeficiente de controle e indica nível de alerta para $k = 2$ e nível de perigo para $k = 3$, p_0 é a fração de não conformidade correspondente a média μ_0 da variável aleatória T, quando a contaminação está em controle, e n é o tamanho da amostra.

Para estabelecer o algoritmo descrito em Leiva et al., 2015 neste estudo, será utilizado uma reparametrização da distribuição BS dada em (2) proposta por Santos-Neto et al., 2012, em que $T \sim BS(\alpha, \mu)$, ou seja, trocando a mediana β pela média $\mu = \frac{\beta}{2}[2 + \alpha^2]$. Foi considerado $t_0 = a\mu_0$, em que $a > 0$ é uma constante de proporcionalidade. Portanto, reparametrizando a distribuição BS em termos da média μ e expressando em função de t_0 e a, tem-se

$$F_T(t_0; \alpha, \mu) = \Phi\left[\frac{1}{\alpha}\xi\left(\frac{a[1+\alpha^2/2]}{\mu/\mu_0}\right)\right], \tag{13}$$

em que ξ é definido por

$$\xi(y) = \sqrt{y} - 1/\sqrt{y} = 2senh(log(\sqrt{y})) \tag{14}$$

Por conseguinte, quando um processo de monitorização está sob controle ($\mu = \mu_0$) para uma concentração de poluentes vindas da distribuição BS, então a partir de (13) a fração de não conformidade pode ser escrita como

$$p_0 = 1 - F_T(t_0; \alpha, a) = \Phi\left[-\frac{1}{\alpha}\xi(a[1 + \alpha^2/2])\right]. \tag{15}$$

Note que a especificação do ponto t_0 é equivalente a indicar o ponto de inspecção constante $a > 0$, porque $t_0 = a\mu_0$, sendo μ_0 a média alvo.

Adequando o algoritmo proposto por Leiva et al., 2015 para um critério da avaliação ambiental por meio de cartas de controle de np quando o contaminante atmosférico pode ser modelado por uma distribuição $T \sim BS(\alpha, \mu)$, tem-se:

1. Tome N subgrupos de tamanho n.

2. Colete n dados $t_1, \ldots, t_n$ da variável aleatória de interesse T para cada subgrupo.

3. Realize um estudo de autocorrelação para os dados recolhidos na etapa 2 para detectar possível dependência sazonal e/ou serial. Se qualquer dependência for detectada, ela deve ser removida usando técnicas adequadas e, em seguida, continuar com o Passo 4.

4. Determine a média alvo μ_0, a constante de inspeção a, e o coeficiente de controle k.

5. Conte em cada subgrupo de dados n o número d de vezes que t_i excede $t_0 = a\mu_0$, para $i = 1, \cdots, n$.

6. Calcule $LIC = \max\{0, n\hat{p}_0 - k\sqrt{n\hat{p}_0(1-\hat{p}_0)}\}$, $LC = n\hat{p}_0$ e $LSC = n\hat{p}_0 + k\sqrt{n\hat{p}_0(1-\hat{p}_0)}$, dados em (12), onde $\hat{p}_0 = \Phi[-(1/\hat{\alpha})\xi(a[1+\hat{\alpha}^2/2])]$, dado em (15) e a estimativa de α dado em (7) ou (9).

7. Declare o processo como fora de controle estatístico se $d \geq LSC$ ou $d \leq LIC$, ou como em controle estatístico se $LIC \leq d \leq LSC$.

Segundo Leiva et al., 2015 devem-se considerar para o algoritmo estabelecido acima, para o caso de monitoramento da qualidade do ar, as seguintes adaptações: Devem ser usadas, para o Passo 1, N subgrupos, por exemplo, N dias, e o tamanho da amostra n formado, por exemplo, como cada hora do dia; para o Passo 2, os dados podem corresponder as concentrações de PM10 em $(\mu g/m^3)$, que devem ser medidas para cada subgrupo com uma estação de monitorização; para Passo 4, d deve ser o número de vezes que as concentrações de PM10 (t_i) excede o valor crítico t_0, que deve ser proporcional a média alvo μ_0, em que t_0 é oriundo de uma diretriz oficial da qualidade do ar; para o Passo 5, apenas o LSC deve ser considerado, com a LIC correspondente sendo igual a zero, declare o nível de contaminação como fora de controle estatístico em uma estação de monitorização, se $d \geq LSC$ ou, de outra forma, como em controle estatístico.

5. APLICAÇÃO

Neste trabalho será analisado o poluente atmosférico PM10 para estações localizadas no estado de São Paulo no ano de 2015. Os dados foram adquiridos por meio do sistema Qualar (Sistema de Informações de Qualidade do Ar), da Companhia Ambiental do Estado de São Paulo (CETESB). Foram escolhidas 3 estações automáticas: uma para a Região Metropolitana de São Paulo (Osasco); uma no Interior (Santa Gertrudes); e uma na Baixada Santista (Cubatão - Vila Parisi). A seleção das estações foi feita de acordo com o grau de poluição segundo o relatório CETESB, 2015. Os meses escolhidos para análise foram agosto e janeiro, e utilizou-se como critérios de seleção as estações inverno e verão, devido as variações sazonais significativas das condições atmosféricas. Também foram considerados meses com 31 dias e com estação completa.

A escolha das três estações de monitoramento da qualidade do ar, em relação ao PM10, para São Paulo ocorreu em razão do estado apresentar áreas com diferentes características e vocações econômicas, o que demanda formas diferenciadas de monitoramento e controle da poluição. Os problemas de qualidade do ar na Região Metropolitana de São Paulo ocorrem especialmente em razão da poluição proveniente do grande número de veículos automotores, o interior paulista caracteriza-se pela intensa atividade agropecuária e a baixada santista pelo alto nível de industrialização com realce para Cubatão que é o principal polo industrial do estado (CETESB, 2014).

Segundo o relatório Qualidade do Ar no Estado de São Paulo da CETESB de 2015, a rede de monitoramento do estado contava com 57 estações automáticas fixas, 1 estação móvel e 29 pontos de monitoramento manual divididos em 13 Unidades de Gerenciamento de Recursos Hídricos (UGRHIs) localizadas nas unidades vocacionais do tipo industrial, em industrialização e agropecuária. 2015 foi um ano também atípico para a qualidade ambiental, devido as chuvas superiores às respectivas médias climatológicas observadas nos meses de maio, julho e setembro, o que contribuiu para que o inverno de 2015 possa ser considerado o mais favorável à dispersão de poluentes dos últimos dez anos (CETESB, 2014; CETESB, 2015).

O método utilizado pela CETESB para medição do parâmetro PM10 pelas redes automáticas de monitoramento foi radiação beta, processando no próprio local e em tempo real na forma de médias horárias as amostragens realizadas em intervalos de cinco segundos. Posteriormente estas médias são transmitidas para a central de telemetria e armazenadas em servidor de banco de dados, em que passam por processo de validação técnica periódica e são

disponibilizadas de hora em hora no endereço eletrônico da CETESB. Para garantir a representatividade temporal dos dados, a CETESB adota como critério para média horária $3/4$ das medidas válidas na hora (CETESB, 2015).

Toda a análise subsequente foi desenvolvida por meio de linguagem de programação R; (R-Team, 2015).

5.1 ANÁLISE EXPLORATÓRIA DOS DADOS

As séries são compostas por dados mensurados em médias horárias do poluente PM10 em ($\mu g/m^3$), observados durante os meses de janeiro e agosto de 2015, para as estações de monitoramento Osasco (A), Santa Gertrudes (B) e Cubatão - Vila Parisi (C). A seguir uma sucinta análise descritiva das séries analisadas.

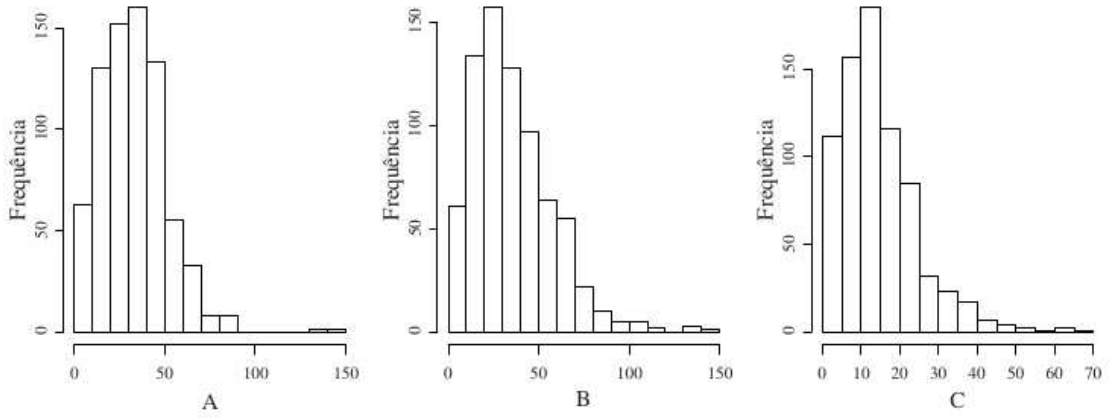
Na Tabela 1 observa-se que para o mês de janeiro de 2015 o poluente PM10 em média horária atingiu seu maior nível na estação de Cubatão ($899\mu g/m^3$). Nota-se que os dados são muito dispersos, o que pode ser observado nos respectivos desvios padrão, $21.082\mu g/m^3$, $24.582\mu g/m^3$, $64.435\mu g/m^3$. As medianas das três estações foram $31\mu g/m^3$, $37\mu g/m^3$, $70\mu g/m^3$, respectivamente. A média horária mais alta que mais foi observada (moda) foi de $86\mu g/m^3$, também para estação Cubatão, evidenciando essa estação como a mais poluída entre as analisadas. A partir dos coeficientes de variação (60.239, 57.195, 77.119), pode-se ver que as séries tem alta dispersão e os dados são bastante heterogêneos. Das assimetrias (1.425, 2.107, 3.897) constata-se assimetrias fortes e das curtoses (2.795, 7.554, 36.284) observa-se que as séries são caracterizadas por curvas platicúrticas, ou seja, uma curva mais aberta que a da distribuição normal.

Tabela 1 - Estatísticas descritivas para a PM10 em ($\mu g/m^3$) no mês de janeiro

Estatísticas	Osasco	Santa Gertrudes	Cubatão
Média	34.997	42.98	83.553
Mediana	31	37	70
Moda	27	24	86
Desvio padrão	21.082	24.582	64.435
Coeficiente de variação	60.239	57.195	77.119
Assimetria	1.425	2.107	3.897
Curtose	2.795	7.554	36.284
Máximo	141	208	899
Mínimo	2	11	2
n	744	744	744

A seguir na Figura 2 tem-se os histogramas dos dados confirmando as características constatadas pela Tabela 1, ou seja, dados heterogêneos, assimetria forte (à direita) e distribuição aberta.

Figura 2 - Histogramas das séries PM10 em ($\mu g/m^3$) no mês de janeiro



Note que a distribuição BS pode ser usada para modelar esses dados devido a sua natureza assimétrica, o seu nível de curtose, em adição aos seus argumentos teóricos para modelar este tipo de dados; (Leiva et al., 2015).

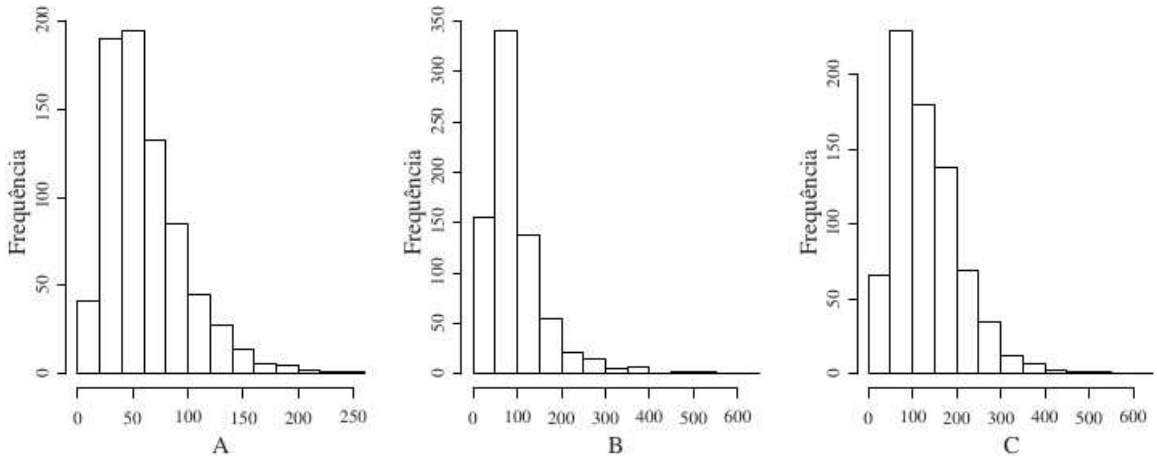
Na Tabela 2 observa-se que para o mês de agosto de 2015 o poluente PM10 em média horária atingiu seu maior nível na estação Cubatão (646$\mu g/m^3$). Nota-se que os dados são muito dispersos, o que pode ser observado nos respectivos desvios padrão, 36.245 $\mu g/m^3$, 75.073$\mu g/m^3$, 80.416$\mu g/m^3$. As medianas das três estações foram 52.5$\mu g/m^3$, 78$\mu g/m^3$, 120$\mu g/m^3$, respectivamente, relembrando que, a concentração máxima permitida do PM10 segundo a legislação do estado de São Paulo é de 120$\mu g/m^3$. A média horária mais alta que mais foi observada (moda) foi de 91$\mu g/m^3$, para estação Cubatão. A partir dos coeficientes de variação (58.005, 76.251, 58.842), vê-se que as séries tem alta dispersão e os dados são bastante heterogêneos. Das assimetrias (1.329, 2.664, 1.518) constata-se assimetrias fortes e das curtoses (2.357, 10.256, 4.322) observa-se que as séries são caracterizadas por curvas platicúrticas, com destaque para estação de Santa Gertrudes.

Tabela 2 - Estatísticas descritivas para a PM10 em ($\mu g/m^3$) no mês de agosto

Estatísticas	Osasco	Santa Gertrudes	Cubatão
Média	62.487	98.454	136.665
Mediana	52.5	78	120
Moda	45	54	91
Desvio padrão	36.245	75.073	80.416
Coeficiente de variação	58.005	76.251	58.842
Assimetria	1.329	2.664	1.518
Curtose	2.357	10.256	4.322
Máximo	247	612	646
Mínimo	3	2	15
n	744	744	744

A seguir na Figura 3 tem-se os histogramas dos dados confirmando as características constatadas pela Tabela 2, ou seja, dados heterogêneos, assimetrias forte (à direita) e distribuição aberta.

Figura 3 - Histogramas das séries PM10 em ($\mu g/m^3$) no mês de agosto



Para o mês de agosto também nota-se que a distribuição BS pode ser usada para modelar esses dados, devido a sua natureza assimétrica e o seu nível de curtose. Observa-se também que o inverno tem níveis de poluição mais altos que o verão, em razão de que o inverno é mais desfavorável à dispersão do poluente PM10.

Os boxplots dos dados de PM10 em $\mu g/m^3$ para as estações em análise se encontram nas Figuras 4 e 5. Eles evidenciam um razoável número de observações possivelmente atípicas nas caudas a direita. No entanto, quando

os dados de concentração de contaminantes têm uma distribuição assimétrica, o boxplot habitual pode resultar em observações que estão sendo classificadas como atípicas, apesar de não serem atípicas. Com o auxílio do boxplot ajustado observa-se que a quantidade de outliers em relação ao boxplot usual diminui. Estes gráficos mostram como as observações que podem ser consideradas atípicas no boxplot habitual, tornam-se não atípicas pelo boxplot ajustado (Marchant et al., 2013).

Figura 4 - Boxplot usual e ajustado para a concentração de PM10 em (μg/m ^{3}) nas estações observadas no mês de janeiro

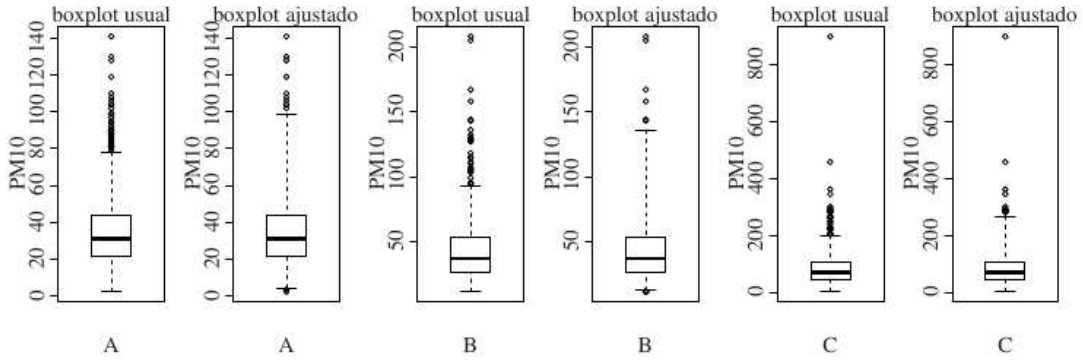
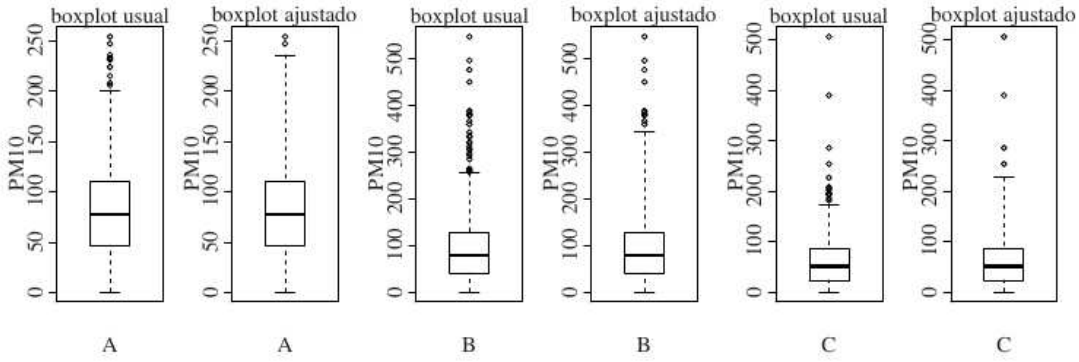
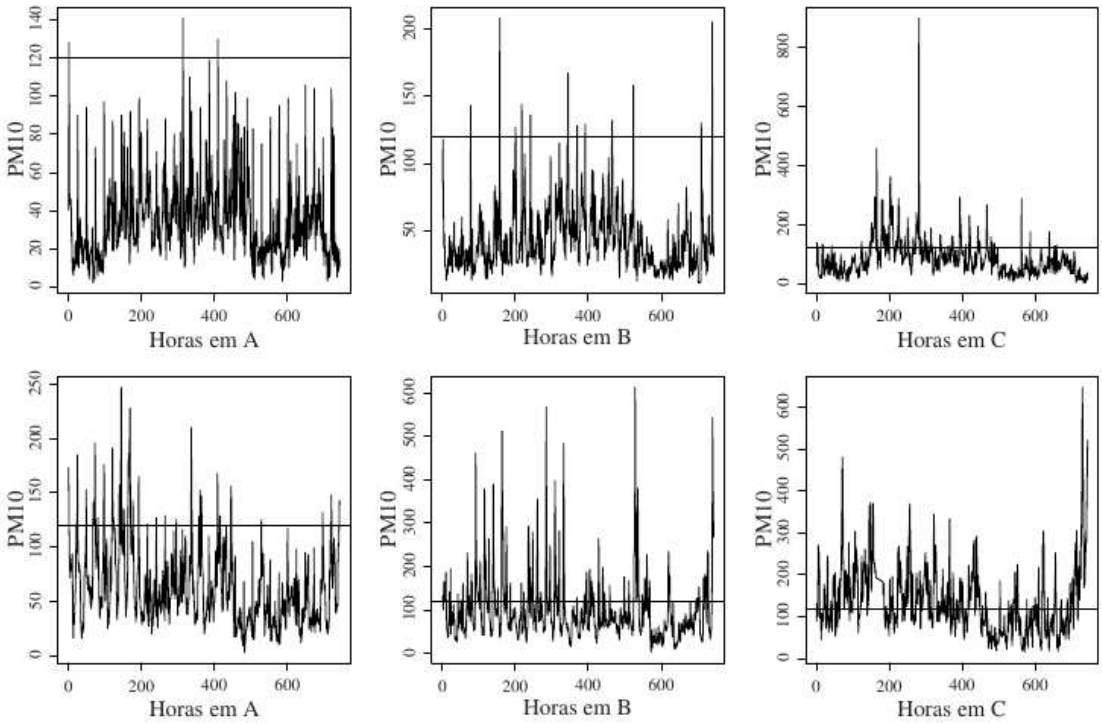


Figura 5 - Boxplot usual e ajustado para a concentração de PM10 em (μg/m ^{3}) nas estações observadas no mês de agosto



Para conhecer melhor as séries PM10 nas estações analisadas, foram feitos os gráficos das séries temporais; ver Figura 6. A partir desta Figura observa-se que não parece haver uma tendência explicita no decorrer dos dias de janeiro (acima) e agosto (abaixo), porém pode se observar vários *"picos"* ultrapassando a concentração máxima permitida de 120 $\mu g/m^3$ (linha horizontal), segundo o decreto estadual estabelecido. Todas as estações ultrapassaram o padrão nos dois períodos analisados. Santa Gertrudes e Cubatão registraram níveis para estado de emergência com observações acima de 500 $\mu g/m^3$ no inverno.

Figura 6 - Séries temporais do PM10 em ($\mu g/m^3$) nos meses de janeiro (acima) e agosto (abaixo)



5.2 CARTAS DE *NP* PARA ESTAÇÕES DO ESTADO DE SÃO PAULO

Para testar a auto correlação das séries utilizou-se o teste Ljung-Box que apresentou valor p menor que $2.2e^{-16}$ para todas as estações nos períodos analisados, rejeitando assim a hipótese de independência entre as observações; veja mais sobre Ljung-Box em (Bueno, 2011).

Tondolo, 2016 propõe o ajuste de modelos autorregressivos integrados e de médias móveis (ARIMA) para modelar os dados e eliminar o efeito da autocorrelação e então controlar estatisticamente a parte aleatória do processo. Nesses casos, o monitoramento é feito por meio do gráfico de controle de resíduos, assumindo que as observações do processo é uma série temporal. No entanto essa modelagem aplica transformações nos dados que torna a interpretação inviável, alternativa que não é interessante, visto que, nossa metodologia objetiva a comparação dos dados com o padrão estabelecido pela legislação oficial. Outra alternativa é diminuir a frequência de amostragem dos dados para atenuar essa questão, porém, essa alternativa subestima os dados e pode levar ao não conhecimento do processo.

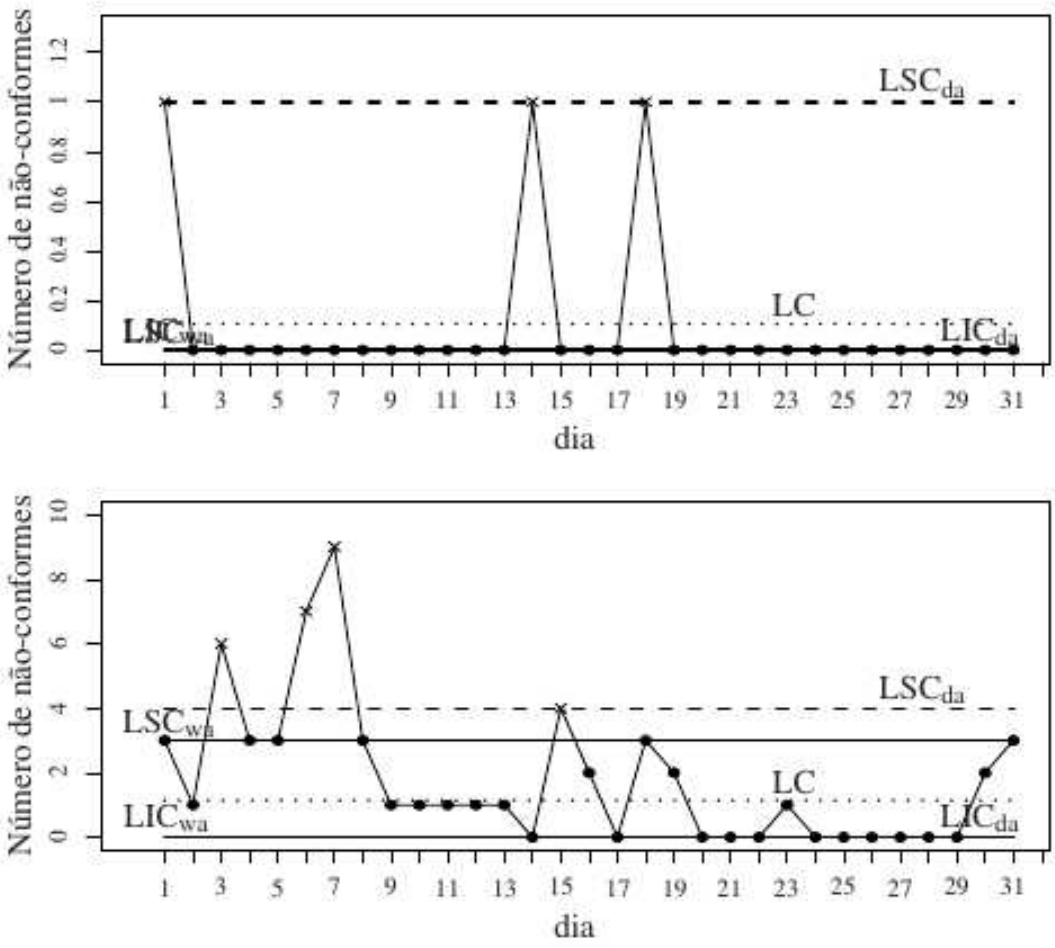
Marchant et al., 2013 defende que na realização de uma análise de autocorrelação o componente espaço-tempo pode ser ou não descartado e a metodologia a ser utilizada ou considera este componente de dependência ou realiza a análise com os dados como uma amostra aleatória. Neste estudo

assume-se que os dados de concentrações de poluentes atmosféricos não estão correlacionados e são independentes, sustentado por (Saghir & Zhengyan, 2015; Marchant et al., 2013; Vilca et al., 2011; Gokhale & Khare, 2007 e Horowitz, 1980).

Para construção das cartas de controle para os dados em questão seguiu-se o algoritmo descrito no Seção 4. Para o passo 1 considerou-se N=31 e n=24. Para o 2 os dados coletados são as médias horárias do poluente PM10 em $\mu g/m^3$, em janeiro e agosto de 2015, para as estações Osasco, Santa Gertrudes e Cubatão, os mesmos da análise descritiva. Para o passo 3 identificou-se autocorrelação, porém, como descrito nos parágrafos anteriores serão desconsideradas. Para o 4 a média alvo (μ_0) é estimada a partir da distribuição BS, $t_0 = 120$ oriundo do decreto estadual de São Paulo, a constante a calculada e $k = 2$, para nível de alerta, e $k = 3$, para nível de perigo. Por conseguinte efetuou-se os cálculos dos passos 5, 6 e construiu-se as cartas que se encontram nas Figuras 7, 8 e 9, para fins de comparação as cartas foram colocadas duas a duas, ou seja, verão e inverno para a mesma estação.

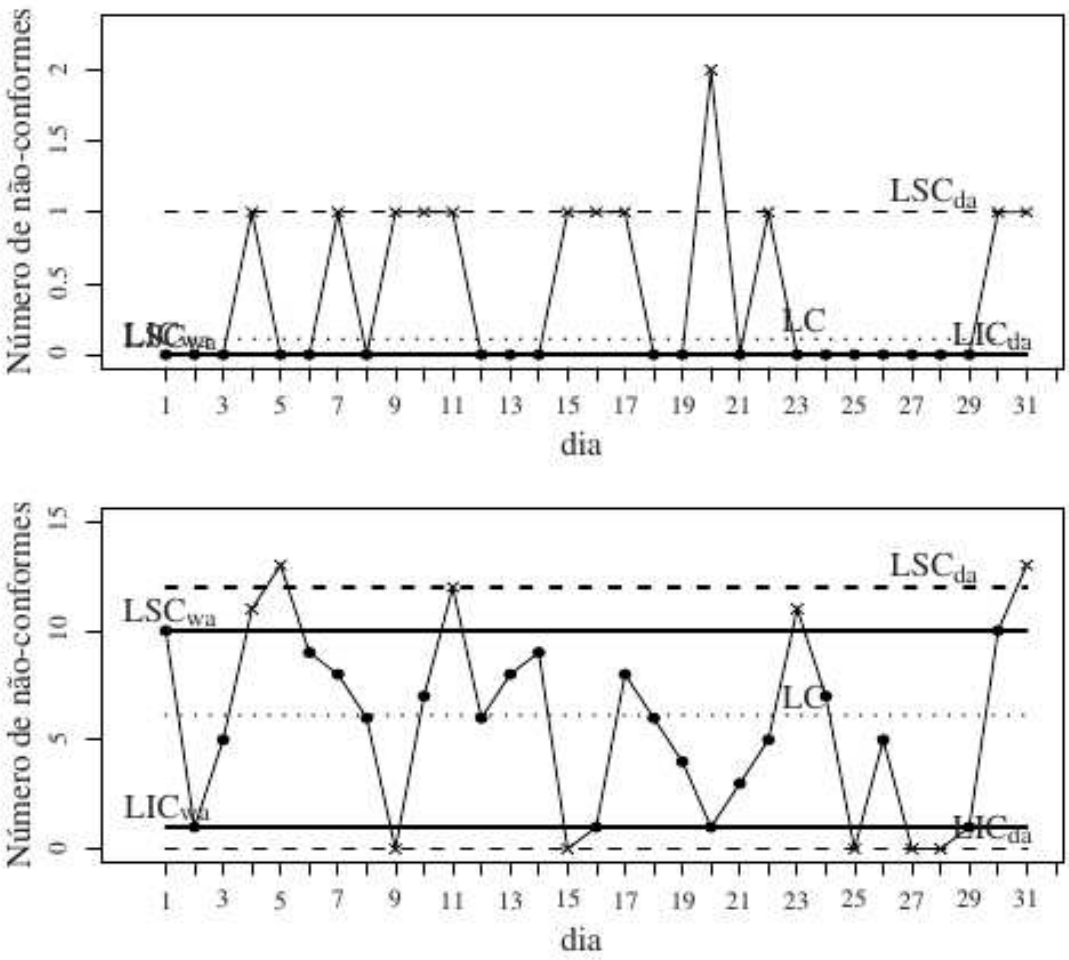
Na Figura 7 as cartas de controle para estação Osasco podem ser observadas. Verifica-se que no mês de janeiro foram registrados três episódios (nos dias 1, 14 e 18) no qual ultrapassaram uma vez cada o padrão ambiental de 120 $\mu g/m^3$ do poluente PM10, atingindo assim o nível de perigo (linha tracejada). Em agosto constata-se que houve um número maior de não conformidades do que em janeiro, foram registrados vinte episódios de ultrapassagem do padrão, em que no dia 7 houve o maior registro de não conformidade (9) observado. Embora ambas cartas tenham ultrapassado o padrão ambiental, apenas a carta de agosto se mostrou fora de controle estatístico. Ainda, não foi observado a existência de comportamentos que evidenciam algum padrão não aleatório.

Figura 7 - Carta de np para BS com $t_0 = 120$, onde "wa" indica nível de alerta ($k = 2$, linha sólida) e "da" indica nível de perigo ($k = 3$, linha tracejada) situações para a concentração de PM10 (em μg/m ^{3}) na estação Osasco nos meses de janeiro (acima) e agosto (abaixo)



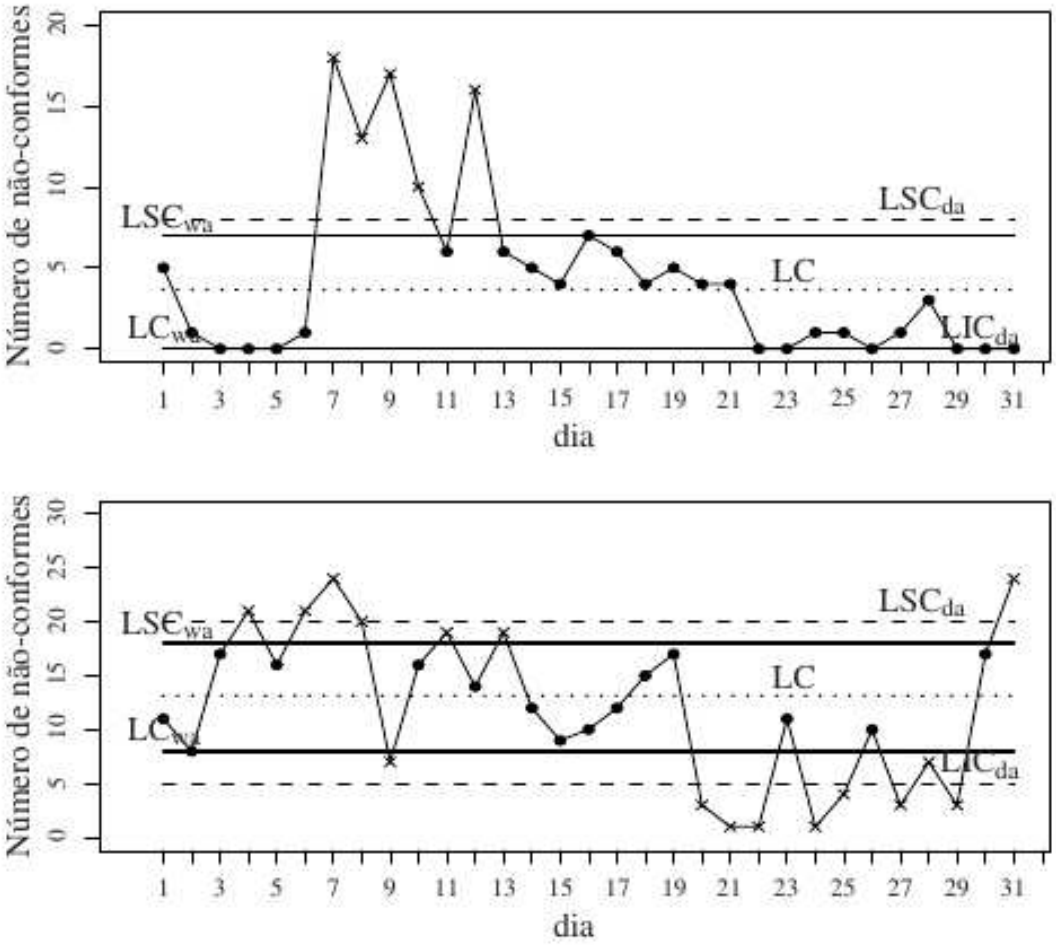
Na Figura 8 são apresentadas as cartas de controle para estação Santa Gertrudes. Verifica-se que no verão foram registrados doze episódios no qual ultrapassaram uma vez cada, exceto dia 18 que ultrapassou duas vezes o padrão ambiental de 120 $\mu g/m^3$ do poluente PM10, transpondo assim o nível de perigo (linha tracejada). No inverno constata-se que houve um número bem maior de não conformidades do que no verão. Foram registrados vinte e cinco episódios de ultrapassagem do padrão, em que nos dias 5 e 31 houveram os maiores registros de não conformidade (13) observados e apenas esses dois dias cruzaram a linha de perigo. Ambas as cartas se mostraram fora de controle estatístico.

Figura 8 - Carta de np para BS com $t_0 = 120$, onde "wa" indica nível de alerta ($k = 2$, linha sólida) e "da" indica nível de perigo ($k = 3$, linha tracejada) situações para a concentração de PM10 (em μg/m ^{3}) na estação Santa Gertrudes nos meses de janeiro (acima) e agosto (abaixo)



Na Figura 9 observa-se as cartas de controle para estação Cubatão-Vila Parisi. Verifica-se que no verão foram registrados vinte e dois episódios no qual ultrapassaram o máximo permitido várias vezes, com destaque para o dia 7 que ultrapassou dezoito vezes o padrão ambiental de 120 μg/m^3 do poluente PM10, transpondo assim o nível de perigo (linha tracejada). Ainda no verão observa-se que o mês começou com a qualidade do ar razoável se elevando a níveis altíssimos de não conformidades na segunda semana e recuperando a qualidade do ar na terceira semana até o fim do mês. Em agosto constata-se que houve um número bem maior de não conformidades do que em janeiro, além disso não houve nenhum dia em que não se registrou episódios de ultrapassagem do padrão, sendo que nos dias 7 e 31 houveram os maiores registros de não conformidade (24) observados. Ambas cartas se mostraram fora de controle estatístico.

Figura 9 - Carta de np para BS com $t_0 = 120$, onde "wa" indica nível de alerta ($k = 2$, linha sólida) e "da" indica nível de perigo ($k = 3$, linha tracejada) situações para a concentração de PM10 (em μg/m ^{3}) na estação Cubatão - Vila Parisi nos meses de janeiro (acima) e agosto (abaixo)



Nesta análise observa-se que o inverno de São Paulo em 2015, nas estações observadas, apresentou um número maior de não conformidade do poluente PM10, isso ocorre devido ao fato que o período de maio a setembro é, geralmente, o mais desfavorável para a dispersão de poluentes primários. Cada região do Estado tem topografia e condições meteorológicas específicas e constata-se que a RMSP (Osasco) apresentou menor número de não conformidade do PM10 do que as outras regiões analisadas, embora tenha apresentado episódios de perigo provavelmente devido ao grande número de emissões veiculares. O interior paulista (Santa Gertrudes) também exibiu níveis de perigo presumivelmente pelas atividades do polo industrial de piso cerâmico em Santa Gertrudes. E na baixada santista (Cubatão - Vila Parisi) constatou-se um proeminente número de não conformidades do poluente devido a característica de intensa atividade industrial de Cubatão. Apesar de não ter sido realizadas inferências este estudo indica que o estado de São Paulo se encontra fora dos padrões ambientais estabelecidos. Nota-se em nossos resultados a

coerência entre o nosso critério e as informações oficiais fornecidas pela CETESB, 2015 que identificou episódios de ultrapassagem em Santa Gertrudes, estabeleceu alertas ambientais durante os dias 7 e 31 de agosto de 2015 para a estação Cubatão - Vila Parisi, em que o relato fornece um aclaramento especial para os altos níveis de poluição dessa região e em Osasco o relatório aponta que o padrão anual de 40 μg/m^3 foi atingido, porém, não apresenta ultrapassagem do padrão diário de 120 μg/m^3.

6. CONCLUSÃO

Neste estudo foi proposto a carta de controle por atributos de np sob o molde da BS, devido ao ajuste das propriedades desse modelo para dados de concentração de poluentes, baseados no algoritmo estabelecido por Leiva et al., 2015. Aplicou-se a metodologia proposta a dados reais recolhidos no sistema da CETESB para três estações em diferentes áreas do estado de São Paulo. Realizou-se uma descrição das séries afim de traçar o panorama ambiental em questão e construiu-se as referidas cartas do número de não conformidade do PM10. Observou-se os altos níveis de poluição do PM10 para as regiões em estudo de São Paulo, os números de não conformidades que vão na contra mão da qualidade do ar e o processo de mensuração do poluente fora de controle estatístico. Notou-se haver um grau maior de poluição no inverno e a proeminente poluição na área industrial. As cartas de controle se mostraram eficazes, no sentido que constatou-se essa afirmação na congruência da análise com o relatório realizado pela CETESB. As cartas de controle podem ser ferramentas eficientes e muito importantes para detectar as ultrapassagens do padrão ambiental oficial, auxiliando nas tomadas de decisões e em implementações de políticas que visem a qualidade ar, como, por exemplo, propor modificações na frota de veículos, alterações no tráfego, mudanças de combustível, alterações no parque industrial e implantação de tecnologias mais limpas.

Como trabalhos futuros podemos considerar alguns problemas específicos explanados a seguir. Primeiro, a dependência serial e/ou a sazonalidade podem ser modeladas usando algum modelo de séries temporais. Segundo, modelos multivariados podem ser considerados de maneira a levar em conta diferentes concentrações de poluentes. Terceiro, técnicas para criação de monitoração simultânea podem também ser consideradas. Os trabalhos sobre esses três problemas estão atualmente em andamento, e esperamos reportar os resultados em trabalhos futuros.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Aslam, M., Jun, C.-H., e Ahmad, M. (2011). New acceptance sampling plans based on life tests for Birnbaum-Saunders distributions. Journal of Statistical Computation and Simulation, 81:461–470.

Barlow, R. e Proschan, F. (1965). Mathematical Theory of Reliability. Wiley, New York, US.

Bartlett, M. e Kendall, D. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. Journal of The Royal Statistical Society B, 8:128–138.

Birnbaum, Z. e Saunders, S. (1969). Estimation for a family of life distributions with applications to fatigue. Journal of Applied Probability, 6:328–347.

Bueno, R. L. S. (2011). Econometria de Séries Temporais. Cengage Learning, São Paulo, BR.

CETESB (2014). Qualidade do ar no estado de São Paulo. Access date: 26 nov. 2015.

CETESB (2015). Qualidade do ar no estado de São Paulo. Access date: 10 out. 2016.

Chhikara, R. e Folks, J. (1977). The inverse Gaussian distribution as a lifetime model. Technometrics, 19:461–468.

Costa, A. F. B., Epprecht, E. K., and Carpinetti, L. C. R. (2005). Controle Estatístico de Qualidade. Atlas, São Paulo, BR.

Desmond, A. (1985). Stochastic models of failure in random environments. Canadian Journal of Statistics, 13:171–183.

Gokhale, S. e Khare, M. (2007). Statistical behavior of carbon monoxide from vehicular exhausts in urban environments. Environmental Modelling & Software, 22:526–535.

Gumbel, E. (1958). Statistics of Extremes. Columbia University Press, New York, US.

Horowitz, J. (1980). Extreme values from a nonstationary stochastic process: an application to air quality analysis. Technometrics, 22:469–478.

Kalbfleish, J. e Prentice, R. (1980). The Statistical Analysis of Failure Time Data. Wiley, New York, US.

Kundu, D., Kannan, N., e Balakrishnan, N. (2008). On the hazard function of Birnbaum-Saunders distribution and associated inference. Computational Statistics and Data Analysis, 52:2692–2702.

Leiva, V. (2015). The Birnbaum-Saunders distribution. Academic Press, New York, US.

Leiva, V., Marchant, C., Ruggeri, F., e Saulo, H. (2015). A criterion for environmental assessment using Birnbaum-Saunders attribute control charts. Environmetrics, 26:463–476.

Leiva, V., Marchant, C., Saulo, H., Aslam, M., e Rojas, F. (2014). Capability indices for Birnbaum-Saunders processes applied to electronic and food industries. Journal of Applied Statistics, 41:1881–1902.

Lio, Y., Tsai, T., e Wu, S. (2010). Acceptance sampling plans from truncated life tests based on the Birnbaum-Saunders distribution for percentiles. Communications in Statistics: Simulation and Computation, 39:119–136.

Marchant, C., Leiva, V., Cavieres, M., and Sanhueza, A. (2013). Air contaminant statistical distributions with application to PM10 in Santiago, Chile. Reviews of Environmental Contamination and Toxicology, 223:1–31.

Marques, R. e Dos Santos, E. (2012). Redes de monitoramneto de material particulado inalável, legislação e os riscos à saúde. Revista Brasileira de Geografia Médica e da Saúde, 8:115–128.

Montgomery, D. C. (2005). Introduction to Statistical Quality Control. Wiley, New York, US.

Nelson, W. e Hahn, G. (1972). Linear estimation of a regression relationships from censored data, part I-simple methods and their applications (with discussion). Technometrics, 14:247–276.

Ott, W. (1990). A physical explanation of the lognormality of pollution concentrations. Journal of the Air and Waste Management Association, 40:1378–1383.

Pinto, W. P. (2013). O uso da metodologia de dados faltantes em séries temporais com aplicação a dados de concentração (PM10) observados na região da grande Vitória. PhD thesis, Universidade Federal do Espírito Santo, Brasil, Vitória, BR.

R-Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Saghir, A. e Zhengyan, L. (2015). Control Charts for Dispersed Count Data: An Overview. Quality and Reliability Engineering, in press:725–739.

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V. e Ahmed, S.E. (2015). On new parameterizations of the Birnbaum-Saunders distribution. Pakistan Journal of Statistics, 28:1–26.

Saulo, H., Leiva, V., and Ruggeri, F. (2015). Monitoring environmental risk by a methodology based on control charts. In Kitsos, C., Oliveira, T., e Rigas, A.and Gulati, S., editors, Theory and Practice of Risk Assessment, pages 177–197. Springer, Switzerland.

Tondolo, C. M. (2016). Gráficos de controle para dados do tipo taxas e proporções autocorrelacionados. Master's thesis, Universidade Federal de Santa Maria, Brasil, Santa Maria, RS, BR.

Vilca, F., Santana, L., Leiva, V., e Balakrishnan, N. (2011). Estimation of extreme percentiles in Birnbaum-Saunders distributions. Computational Statistics and Data Analysis, 55:1665–1678.

Revista Brasileira de Estatística, V.77, N. 242, p.33-51, julho/dezembro 2019

DISTRIBUIÇÃO SIMPLEX: AVALIAÇÃO E ESTIMAÇÃO DOS PARÂMETROS

Fidel Henrique Fernandes

fidelhenrique@hotmail.com

Universidade Federal do Rio Grande do Norte - UFRN

Renato Santos da Silva

renatoifpi@gmail.com

Universidade de São Paulo - USP

Resumo: Este trabalho objetiva analisar e discutir o desempenho dos estimadores μ e λ da distribuição Simplex. A partir de amostras geradas por essa distribuição, calculamos estimativas pontuais dos parâmetros utilizando o método de máxima verossimilhança, a sua versão corrigida por *bootstrap*, o método de momentos e método bayesiano. As Simulações de Monte Carlo foram utilizados para o desenvolvimento do trabalho em diversos cenários, destacando-se o tamanho amostral e verificando algumas propriedades desses estimadores, como a média, variância, viés e erro quadrático médio. O estimador corrigido por *bootstrap* apresentou melhores resultados do que os demais estimadores. Por fim, aplicamos a metodologia estudada para um conjunto de dados reais referente a proporção da população ateísta em 137 países.
Palavras-chave: *Bootstrap*, Distribuição Simplex, Máxima Verossimilhança, Método Bayesiano, Método de Momentos, Monte Carlo.

Abstract: This paper aims to analyze and discuss the performance of the estimators μ e λ of Simplex distribution. From samples generated by this distribution, we calculate punctual perform estimations of the parameters using the maximum likelihood method, your version corrected by *bootstrap*, method of moments and method bayesian. The Monte Carlo simulations were used for the development of paper in different scenarios, specially the sample size used and checking some properties these estimators, as mean, variance, bias and mean squared error. The estimators corrected by *bootstrap* showed better results compared to the uncorrected estimators than the other estimators. Finally, we apply the methodology studied for a set of real data concerning the atheist proportion in 137 countries.
Keywords: *Bootstrap*, Maximum Likelihood, Method Bayesian, Method of Moments, Monte Carlo, Simplex Distribution.

1.INTRODUÇÃO

A distribuição Simplex foi inicialmente proposta por Barndorff-Nielsen & Jørgensen (1991) para a modelagem de dados restritos no intervalo contínuo (0,1). Em muitas situações práticas, podemos estar diante de variáveis restritas ao intervalo (0,1) tais como: taxas, proporções e índices. Vale frisar, que a distribuição simplex faz parte dos modelos de dispersão propostos por Jørgensen (1997).

Por se tratar de uma distribuição que comporta valores no intervalo (0,1), essa distribuição tem sido utilizada como uma alternativa à distribuição beta para modelar proporções, ver Silva (2015b). Nesta instância, Song & Tan (2000) propuseram um modelo de regressão simplex com dispersão constante para modelar dados contínuos de proporção longitudinais, sob enfoque de equações de estimações generalizadas. Além disso, Miyashiro (2008) propôs o modelo de regressão simplex com dispersão constante, apresentando expressões analíticas para o vetor escore e para a matriz de informação de Fisher correspondente ao vetor de parâmetros, além de medidas de diagnósticos (Silva, 2015a). Os estimadores usuais, contudo, podem não apresentar desempenho satisfatório em um cenário de tamanho amostral pequeno. Logo, torna-se necessária a obtenção de estimadores que sejam menos tendenciosos em amostras pequenas.

Neste contexto, o presente trabalho apresenta como foco principal avaliar o desempenho dos estimadores pontuais dos parâmetros da distribuição Simplex comportada em dois parâmetros. Mediante o exposto, é pertinente destacar que realizamos simulações de Monte Carlo utilizando o *software R* Team (2014) e o pacote *VGAM*, desenvolvido por Yee et al. (2010) na obtenção de amostras aleatórias da distribuição Simplex. Os estimadores pontuais serão estimados pelo método da máxima verossimilhança e sua versão corrigida por *bootstrap*, pelo estimador de momentos e por estimação bayesiana (utilizando distribuições à *priori* não informativa, e estimativas da média *posteriori* para cada parâmetro), foram considerados diferentes tamanhos amostrais, e valores diferentes para cada parâmetro da distribuição Simplex.

Mediante exposição introdutória feita nesta seção, apresentamos e discutimos na Seção 2 as principais características da distribuição Simplex. Na Seção 3 apresentamos e discutimos cada estimador proposto. Na Seção 4, retratamos a metodologia juntamente com os resultados e discussões. Na Seção 5, aplicamos a metodologia estudada para um conjunto de dados reais e por fim, algumas conclusões dos resultados encontrados.

2. REFERENCIAL TEÓRICO

2.1 DISTRIBUIÇÃO SIMPLEX

Dizemos que uma variável aleatória Y possui distribuição pertencente aos modelos de dispersão se sua função densidade de probabilidade pode ser escrita na forma:

$$f(y|\mu,\sigma^2) = a(y,\sigma^2)e^{\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}},\ y \in \Theta. \quad (1)$$

em que Θ é o suporte da distribuição, $\mu \in \Omega$, em que Ω é o espaço amostral, ou seja, $\mu \in (0,1)$ é o parâmetro de locação, $\sigma^2 > 0$ é o parâmetro de dispersão e $a(y,\sigma^2)$ é uma constante normalizada, independente de μ. A função $d(y;\mu)$ é conhecida como componente *deviance* definida em $(y,\mu) \in (\Theta,\Omega)$ e deve satisfazer as seguintes propriedades

$$d(y,y) = 0,\ \forall y = \mu \in \Omega \quad \text{e} \quad d(y,\mu) > 0,\ \forall y \neq \mu. \quad (2)$$

A função de variância para os modelos de dispersão é definida como

$$V(\mu) = \frac{2}{\frac{\partial^2 d(y;\mu)}{\partial \mu^2}|_{y=\mu}}, \quad (3)$$

desde que $d(y;\mu)$ seja contínua e duas vezes diferenciável com respeito a μ e satisfaça

$$\frac{\partial^2 d(y;\mu)}{\partial \mu^2}|_{y=\mu} > 0, \forall \mu \in (0,1). \quad (4)$$

Para mais detalhes, ver Silva (2015a). A partir da definição da expressão (1), a variável aleatória Y segue uma distribuição Simplex indexada por dois tipos de parâmetros locação $\mu \in (0,1)$ e dispersão $\lambda > 0$, em que $\lambda = \frac{1}{\sigma^2}$, a expressão (1) é dada por:

$$f(y|\mu,\lambda) = \left\{\frac{\lambda}{2\pi[y(1-y)]^3}\right\}^{1/2} e^{\left\{-\frac{\lambda}{2}d(y;\mu)\right\}},\ 0 < y < 1, \quad (5)$$

em que

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}, \quad (6)$$

é chamada unidade *deviance*.

A média e a variância de Y são dadas, respectivamente por

$$E(Y) = \mu\,, \tag{7}$$

$$V(Y) = \mu(1-\mu) - \sqrt{\frac{\lambda}{2}}\, e^{\left\{\frac{\lambda}{\mu^2(1-\mu)^2}\right\}} \Gamma\left\{\frac{1}{2}, \frac{\lambda}{2\mu^2(1-\mu)^2}\right\}, \tag{8}$$

em que $\Gamma\{a,b\}$ e a função gama incompleta superior definida como

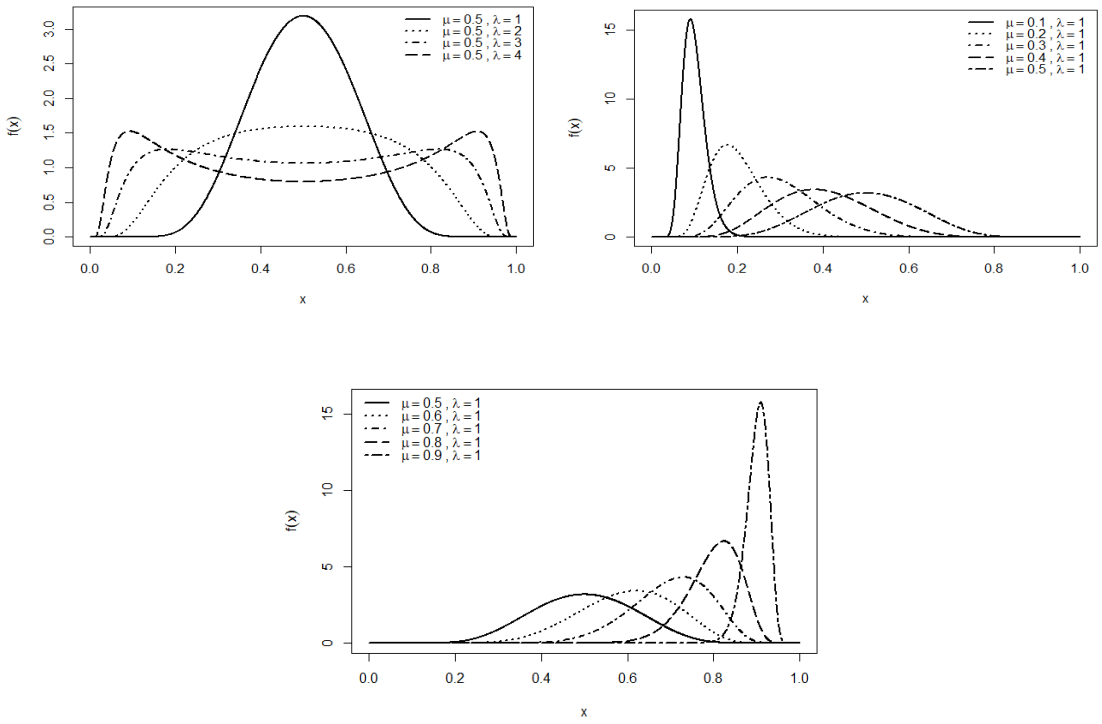
$$\Gamma\{a,b\} = \int_b^{\infty} u^{a-1}\, e^{-u} du. \tag{9}$$

Outro resultado bastante interessante desta distribuição é o fato de que além de assemelhar-se com a densidade da distribuição normal, segundo Jorgensen (1997) a teoria assintótica de pequena dispersão quando o parâmetro de dispersão $\sigma^2 \rightarrow 0$, diz que

$$\frac{Y-\mu}{\sigma\sqrt{V(\mu)}} \rightarrow N(0,1). \tag{10}$$

Ilustramos abaixo três gráficos da densidade variando os valores de μ e λ. A Figura 1 apresenta diferentes densidades simplex correspondentes a alguns valores dos parâmetros (μ, λ). A densidade Simplex é assimétrica quando $\mu \neq 0.5$ e simétrica quando $\mu = 0.5$. Fixando o $\lambda = 1$ e para valores de $0.1 \leq \mu < 0.5$, a distribuição apresenta uma assimetria à direita e quando $0.5 < \mu \leq 0.9$, apresenta uma assimetria à esquerda.

Figura 1: Densidade da distribuição Simplex para diferentes valores de μ e λ



Fonte – Autoria própria

3. ESTIMAÇÃO DOS PARÂMETROS

Uma das etapas de extrema importância na análise de ajuste de um modelo é a estimação de seus parâmetros. Existem diversos métodos na literatura para a realização deste procedimento. Mediante o exposto, salientamos que o objetivo é apresentar alguns métodos que foram utilizados da estimação dos parâmetros da distribuição Simplex. Dessa forma, os métodos aqui utilizados foram máxima verossimilhança, método dos momentos, estimação bayesiana e o método *bootstrap*. Para a distribuição simplex definimos o vetor de parâmetros $\theta \in (\mu, \lambda)$.

3.1 MÉTODO DE MÁXIMA VEROSSIMILHANÇA

O método de máxima verossimilhança trata do problema de estimação baseado nos resultados obtidos pela amostra. Dessa forma, é intuitivo pensar que a melhor estimativa para o parâmetro, baseado na amostra, é o valor do parâmetro que tem maior probabilidade de gerar o resultado visto na amostra, portanto, é o valor que maximiza a função de máxima verossimilhança.

Sejam $y_1, \ldots, y_n$ uma amostra aleatória de tamanho n, da distribuição Simplex com parâmetros μ e λ, com densidade dada em (5). Dessa forma, a sua função de máxima verossimilhança é dada por

$$L(\theta, y) = \prod_{i=1}^{n} f(y|\mu,\lambda) = \prod_{i=1}^{n} \left\{\frac{\lambda}{2\pi\{y_i(1-y_i)\}^3}\right\}^{1/2} e^{\left\{-\frac{\lambda}{2}d(y_i,\mu)\right\}}$$

$$= \left\{\frac{\lambda}{2\pi\prod_{i=1}^{n}\{y_i(1-y_i)\}^3}\right\}^{n/2} e^{\sum_{i=1}^{n}\left\{-\frac{\lambda}{2}d(y_i,\mu)\right\}} \quad (11)$$

A função de log-verossimilhança é apresentada da seguinte forma

$$l(\theta, y) = \log L(\theta|y) = \frac{n}{2}\left(\log\lambda - log(2\pi\prod_{i=1}^{n}\{y_i(1-y_i)\}^3)\right) - \frac{\lambda}{2}\sum_{i=1}^{n} d(y_i,\mu) \quad (12)$$

Assim, os estimadores de máxima verossimilhança de μ e λ são obtidos através da maximização do logaritmo da função de verossimilhança. Logo, fazendo a derivada parcial de $l(\theta, y)$ em relação à μ e em relação à λ respectivamente, obtemos os estimadores de máxima verossimilhança resolvendo as seguintes equações

$$\frac{\partial}{\partial\mu} = -\frac{\lambda}{2}\sum_{i=1}^{n} d(y_i,\mu), \quad (13)$$

$$\frac{\partial}{\partial\lambda} = \frac{n}{2\lambda} - \frac{1}{2}\sum_{i=1}^{n} d(y_i,\mu). \quad (14)$$

em que

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}. \quad (15)$$

É notório que os estimadores de máxima verossimilhança de μ e λ não podem ser resolvidos explicitamente. Neste caso, necessitaremos de métodos numéricos de maximizações de funções. No presente trabalho, utilizou-se função **maxLik** para a maximização, sendo uma ferramenta para a estimativa máxima e não-linear de otimização.

3.2 MÉTODO DOS MOMENTOS

Uma outra forma de encontrar estimadores é através do método dos momentos. Este método é baseado nos momentos teóricos e amostrais das variáveis aleatórias envolvidas. Seja

$$m_k = \frac{1}{n}\sum_{i=1}^{n} y_i^k, \ \ k \geq 1, \quad (16)$$

o k - ésimo momento amostral, e $\mu_k = E(Y^k)$, o k - ésimo momento populacional.

Se existem k parâmetros para serem estimados, o método consiste em expressar os primeiros k momentos da população em termos dos k parâmetros através de k equações. A solução do sistema formado por essas equações leva a determinação dos estimadores. O procedimento é igualar os momentos teóricos pelos respectivos momentos amostrais, ou seja, resolvendo $m_k = \mu_k$. O valor esperado da distribuição simplex é dado por $E(Y) = \mu$ assim, igualando-a com a média amostral, temos que

$$\tilde{\mu} = \bar{x}, \quad (17)$$

em que

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (18)$$

Como a variância da distribuição Simplex depende da função gama e não há forma analítica, por essa razão não foi exposto o estimador de momentos para o λ.

3.3 ESTIMAÇÃO BAYESIANA

Uma outra abordagem para a estimação dos parâmetros é o método Bayesiano. Quando se estuda características da população, que normalmente são descritas pela curva de densidade e pelos seus parâmetros, é possível que o pesquisador desse estudo tenha algum conhecimento sobre o comportamento destes parâmetros. Na abordagem Bayesiana é assumido que os parâmetros de interesse possuem uma distribuição de probabilidade. Esta distribuição inicial pode ser incorporada em uma densidade $p(\theta)$, que representa distribuição do parâmetro θ chamada de distribuição à *priori*.

A inferência Bayesiana descreve incertezas sobre quantidades incertas de forma probabilística. Dessa forma, para estimar o comportamento dos parâmetros necessitamos de dois componentes, a função de verossimilhança $l(\theta|y)$ e a distribuição à *priori* $p(\theta)$. A inferência sobre θ é baseada na sua função de densidade de probabilidade após os dados serem observados. Assim, esta função é denominada função densidade de probabilidade à *posteriori*, que é a distribuição do vetor de parâmetros, após inserirmos a informação contida nos dados observados, além de que é representada por $\pi(\theta) = f(\theta|y)$, sendo obtida por meio do teorema de Bayes representado por:

$$\pi(\theta) = f(\theta|y) = \frac{l(\theta|y)p(\theta)}{\int l(\theta|y)p(\theta)d\theta} \propto l(\theta|y)p(\theta), \tag{19}$$

em que $\int l(\theta|y)p(\theta)d\theta$ não depende do parâmetro θ.

Para encontrarmos a distribuição à *posteriori* $\pi(\theta)$ não precisamos ter o conhecimento de toda a sua densidade, entretanto, necessitamos do núcleo da distribuição que representa a parte que depende de θ. Devido a isso, poderemos encontrar problemas para obter uma forma fechada da distribuição à *posteriori*, em que muitas vezes não resulta em uma distribuição conhecida. Precisamos de um método que facilite encontrar esta distribuição, chamado de método de Monte Carlo via cadeia de Markov (MCMC), pode ser visto com mais detalhes em Gamerman & Lopes (2006). Os dois principais métodos MCMC são, o amostrador de Gibbs e o algoritmo de Metropolis - Hastings. Nesse aspecto, salientamos que para o desenvolvimento do presente trabalho foi utilizado o algoritmo de Metropolis-Hastings para estimar as distribuições posteriores marginais para cada parâmetro da distribuição Simplex, e este algoritmo pode ser visto em Chib & Greenberg (1995).

Considerando a função de densidade da distribuição Simplex em (5) e a função de verossimilhança de μ e λ em (11), levando em conta que $\mu \in (0,1)$ e $\lambda > 0$, consideramos as distribuições à *priori* da seguinte forma:

- $\mu \sim Beta(a_1, b_1)$, a_1 e b_1 são conhecidos.
- $\lambda \sim Gamma(a_2, b_2)$, a_2 e b_2 são conhecidos.

Dessa forma, obtemos a distribuição à *priori* para os parâmetros μ e λ dado por:

$$p(\mu, \lambda) = f_B(\mu|a_1, b_1) f_G(\lambda|a_2, b_2) \quad (20)$$

$$= \frac{1}{B(a_1,b_1)} \mu^{a_1-1}(1-\mu)^{b_1-1} * \frac{{b_2}^{a_2}}{\Gamma(a_2)} \lambda^{a_2-1} e^{-b_2\lambda} \quad (21)$$

$$\propto \mu^{a_1-1}(1-\mu)^{b_1-1}\lambda^{a_2-1}e^{-b_2\lambda} \quad (22)$$

em que $B(a_1, b_1) = \int_0^1 t^{a_1-1}(1-t)^{b_1-1}dt$ e $\Gamma(a_2) = \int_0^1 t^{a_2-1} e^{-t} dt$ (23)

Assumindo independência entre os parâmetros μ e λ. A distribuição conjunta à *posteriori* é

$$\pi(\theta) \propto l(\theta|y)p(\theta) \quad (24)$$

$$\propto \left[\frac{n}{2}\log(\lambda) - \frac{\lambda}{2}\sum_{i=1}^{n} d(y_i;\mu)\right] \mu^{a_1-1}(1-\mu)^{b_1-1}\lambda^{a_2-1}e^{-b_2\lambda}. \quad (25)$$

Os valores escolhidos para a_1 e b_1 foram 0.001 e para a_2 e b_2 foram de 0.01, a fim de obter uma variância grande por se tratar de uma *priori* não informativa, dado que não temos conhecimento sobre a distribuição dos parâmetros.

3.4 MÉTODO BOOTSTRAP

Muitas vezes na estatística necessitamos de métodos para simular um experimento diversas vezes, para obter resultados que a teoria matemática não pode fornecer de maneira rápida. O método *bootstrap* é um método computacional introduzido por Efron (1992), que consiste em um método de reamostragem, baseando-se em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra, substituindo cálculos analíticos por esforço computacional, e além disso é bastante útil quando se deseja avaliar, para um certo estimador, o seu viés, ou ainda a distribuição de probabilidade. A enorme capacidade de cálculo dos computadores da atualidade facilita consideravelmente a aplicabilidade deste método tão custoso computacionalmente.

Considerando uma amostra aleatória $y = (y_1, \dots, y_n)$ de uma determinada distribuição F desconhecida e seja $\beta(y, F)$ uma certa quantidade de interesse. Normalmente, estamos interessados em estimar alguma característica probabilística $\beta(y, F)$. O *bootstrap* é uma técnica apropriada para este procedimento. O método pode ser descrito em três passos:

- Primeiro fazemos uma troca da distribuição F por uma conhecida $\hat{F}$;
- Obter F amostras $y^{*1}, \dots, y^{*B}$ de $\hat{F}$ e calcular $T^i = \beta(y^{*i}, \hat{F})$ para $i = 1, \dots, B$;
- Calcular características de interesse, como média, viés, a partir da distribuição amostral dos valores $T^1, \dots, T^B$.

Existem duas variações do método *bootstrap*, são eles o *bootstrap* paramétrico e o *bootstrap* não paramétrico. No *bootstrap* paramétrico, fazemos suposições da distribuição, ou seja, F é de alguma família de distribuições paramétricas e reamostramos observações da distribuição postulada, mas usando os valores das estimativas, por exemplo, os estimadores de máxima verossimilhança dos parâmetros no processo de geração das pseudo-amostras. No *bootstrap* não paramétrico o processo de reamostragem se dar a partir da distribuição empírica dos dados. Nesse sentido, salientamos que a utilização do *bootstrap* paramétrico se deu ao fato de termos conhecimento da distribuição de F.

3.4.1 CORREÇÃO DE VIÉS POR BOOTSTRAP

A correção de viés tem sido um assunto bastante estudado e discutido na literatura estatística. Com intuito de obter estimadores corrigidos, ou seja, estimadores com seu menor viés, visto que tais estimadores são, tipicamente, viesados para os verdadeiros valores dos parâmetros quando o tamanho amostral é pequeno ou a informação de Fisher é reduzida. Lopes (2007) apresenta outras formas de correções de viés. A principal ideia é usar reamostragens dos dados com o objetivo de estimar a função de viés.

Podemos utilizar o método bootstrap para estimarmos o viés de um determinado estimador. Suponha que $Y = (Y_1, Y_2, \dots, Y_n)$ um conjunto de variáveis aleatórias independentes e identicamente distribuídas de uma função de distribuição populacional $F = F_\theta(y)$, em que $\theta = t(F)$ é o parâmetro e seja $\hat{\theta} = s(Y)$ o estimador de θ. O viés do $\hat{\theta}$ é definido como

$$B_F(\hat{\theta}, \theta) = E_F[s(Y)] - t(F). \quad (26)$$

em que o subscrito F indica que a esperança matemática é calculada com base em F. Dessa forma, as estimativas de viés *bootstrap* paramétrico são dados por

$$B_F(\hat{\theta}) = E_{\hat{F}}[s(Y)] - t(\hat{F})$$

$$= \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^* - \hat{\theta}. \quad (27)$$

Gerando B amostras *bootstrap* independentes (B grande), $Y = (Y^{*1}, Y^{*2}, \dots, Y^{*B})$ de Y, calculando as suas respectivas réplicas *bootstrap* $(\hat{\theta}^{*1}, \hat{\theta}^{*2}, \dots, \hat{\theta}^{*B})$, em que $\hat{\theta}^{*i} = s(Y^{*i})$ para $i = (1, \dots B)$, logo podemos aproximar $E_{\hat{F}}[s(Y)]$ por $\bar{\hat{\theta}}_i^*$ na qual

$$\bar{\hat{\theta}}_i^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*. \quad (28)$$

Obtemos assim, as estimativas *bootstrap* de viés por

$$\hat{B}_{\hat{F}}(\hat{\theta}) = \bar{\hat{\theta}}_i^* - s(Y). \quad (29)$$

Observe que a diferença entre as estimativas está na forma de geração das subamostras. Por isso, podemos a partir das *bootstrap* estimativas de viés, definir estimadores corrigidos (Mackinnon & Smith, 1998) por *bootstrap* dado por

$$\hat{\theta}^{cc} = 2\hat{\theta} - \bar{\hat{\theta}}^* \quad (30)$$

4. ESTUDOS DE SIMULAÇÃO

A fim de avaliar os desempenhos dos estimadores propostos, realizamos simulações de Monte Carlo com amostras de tamanho finito e de diferentes valores para o vetor paramétrico. Todo o processo de simulação foi desenvolvido utilizando a linguagem de programação R Team (2014). A geração de números aleatórias da distribuição Simplex foi através do pacote VGAM desenvolvido por Yee et al. (2010).

Os tamanhos amostrais considerados foram $n = 25{,}50{,}75\ e\ 100$ e os valores considerados para o parâmetro λ foram $3.0, 5.0\ e\ 7.0$, e para o parâmetro μ foram $0.3{,}0.4\ e\ 0.7$. Consideramos $R = 5000$ (número de réplicas de Monte Carlo) e $B = 500$ (número de réplicas *bootstrap*). Para cada réplica de Monte Carlo, ou seja, para cada estimativa de $\hat{\theta} = (\hat{\mu}, \hat{\lambda})$, foi gerado B réplicas *bootstrap* de forma não-paramétrica, isto é, geramos

$$x^{*1}, \lambda, x^{*B} \quad (31)$$

em que

$$x^{*i} = (x_1^*, \dots, x_n^*), i = 1, \dots, B \quad (32)$$

Com essas réplicas *bootstrap* determina-se as estimativas dos vieses de $\hat{\theta} = (\hat{\mu}, \hat{\lambda})$ e calcula-se, assim, as estimativas corrigidas por *bootstrap* $\hat{\theta}^* = (\hat{\mu}^*, \hat{\lambda}^*)$ de acordo com (30). Para a análise dos resultados da estimação pontual calculamos a média, o viés, a variância (Var) e o erro quadrático médio (EQM) dos estimadores.

4.1 RESULTADOS E DISCUSSÕES

A avaliação dos desempenhos dos estimadores dos parâmetros da distribuição Simplex através de métodos empíricos é a motivação do estudo apresentado nesta seção.

Os resultados das simulações encontram-se nas Tabelas 1, 2, 3 e 4. Nelas, obtemos as estimativas de μ e λ, considerando os tamanhos amostrais 25,50,75 e 100. Em que $\hat{\theta}$ e $\hat{\theta}^*$ são os estimadores de máxima verossimilhança e sua correção por *bootstrap*, respectivamente, $\tilde{\theta}$ o estimador de momentos e $\bar{\theta}$ o bayesino. Vale salientar que, para o estimador de momentos de λ não foi obtido, uma vez que variância da distribuição Simplex não apresenta forma analítica.

A Tabela 1 apresenta as estimativas para o caso em que $\mu = 0.7$ e $\lambda = 3$. Neste cenário, notamos que os estimadores $\hat{\mu}^*$ apresenta, em módulo, um viés menor do que os demais estimadores em todos os diferentes tamanhos amostrais. Por exemplo, para $n = 25$ o viés de $\hat{\mu}^*$ é dado, por 0.0003 enquanto o viés de $\tilde{\mu}$ é 0.0004. Nos tamanhos amostrais de 75 e 100, sobressaiu além do $\hat{\mu}^*$ o estimador de momentos $\hat{\mu}^*$ apresentaram um viés menor do que os outros estimadores. Por exemplo, para n = 100 o viés de $\hat{\mu}^*$ resultou, em módulo, de 0.00006 enquanto que o viés de $\tilde{\mu}$ é 0.0001.

Em termos do EQM é interessante frisar que para o parâmetro μ, o estimador que apresentou um EQM menor foi o $\bar{\mu}$, junto com $\hat{\mu}$ e $\hat{\mu}^*$, e como as estimativas foram muito próximas alguns estimadores apresentaram EQM iguais, isso vale também em termos de variabilidade. Nota-se também que, à medida que o tamanho da amostra cresce o EQM dos quatro estimadores diminuem.

Tabela 1: Estimativas dos parâmetros $\mu = 0.7$ e $\lambda = 3.0$

		Estimativas de μ				Estimativas de λ			
n	Estimador	Média	Viés	Var	EQM	Média	Viés	Var	EQM
25	$\hat{\theta}$	0.7009	0.0009	0.0015	0.0015	2.9046	−0.0953	0.1756	0.1847
	$\hat{\theta}^*$	0.6996	−0.0003	0.0015	0.0015	2.9924	−0.0075	0.1868	0.1869
	$\bar{\theta}$	0.6995	−0.0004	0.0017	0.0017	-	-	-	-
	$\bar{\bar{\theta}}$	0.6961	−0.0038	0.0015	0.0015	3.2870	0.2870	0.2676	0.3500
50	$\hat{\theta}$	0.7009	0.0009	0.00074	0.0007	2.9572	−0.0427	0.0907	0.0926
	$\hat{\theta}^*$	0.7002	0.0002	0.0007	0.0007	3.0017	0.0017	0.0937	0.0937
	$\bar{\theta}$	0.7004	0.0004	0.0008	0.0008	-	-	-	-
	$\bar{\bar{\theta}}$	0.6987	−0.0012	0.0007	0.0007	3.1350	0.1350	0.1097	0.1279
75	$\hat{\theta}$	0.7007	0.0007	0.0005	0.0005	2.9694	−0.0305	0.0580	0.0589
	$\hat{\theta}^*$	0.7002	0.0002	0.0005	0.0005	2.9993	−0.0006	0.0591	0.0591
	$\bar{\theta}$	0.7001	0.0001	0.0006	0.0006	-	-	-	-
	$\bar{\bar{\theta}}$	0.6992	−0.0007	0.0005	0.0005	3.0844	0.0844	0.0659	0.0731
100	$\hat{\theta}$	0.7003	0.0003	0.0003	0.0003	2.9772	−0.0227	0.0452	0.0457
	$\hat{\theta}^*$	0.6999	−0.00006	0.0003	0.0003	2.9995	−0.0004	0.0459	0.0459
	$\bar{\theta}$	0.7001	0.0001	0.0004	0.0004	-	-	-	-
	$\bar{\bar{\theta}}$	0.6992	−0.0007	0.0003	0.0004	3.0626	0.0626	0.0501	0.0540

Fonte – Autoria própria.

Em relação ao parâmetro λ nota-se que o estimador $\hat{\lambda}^*$ apresentou menor viés em todos os cenários apresentados. Por exemplo, para n=100 o $\hat{\lambda}^*$ o viés é, em módulo, de 0.0004 enquanto que o $\hat{\lambda}$ o viés é 0.027 e $\bar{\lambda}$ o viés é de 0.0626. Vale notar também que o estimador $\bar{\lambda}$ é o que apresenta o maior viés em todos os cenários. Já em relação ao EQM o estimador que apresentou o menor EQM foi $\hat{\lambda}$ em relação aos demais. E ainda para o $\hat{\lambda}^*$ apresentou um viés menor comparado a sua versão sem correção, porém o seu EQM é maior. Nota-se também que à medida que o tamanho amostral aumenta os EQM diminuem.

Na Tabela 2 é exposto as estimativas para o caso em que $\mu = 0.7$ e $\lambda = 5$. Mostrando as mesmas interpretações da Tabela 1, o estimador de máxima verossimilhança corrigido $\hat{\mu}^*$ apresentou melhores resultados com menor viés em relação as demais estimadores em todos os tamanhos amostrais, evidenciando o êxito das correções de viés por *bootstrap*.

Já em relação ao EQM, o estimador de momentos foi o que apresentou um EQM maior em todos os cenários mostrados para a estimativa de μ, evidenciando o mesmo ocorrido na Tabela 1 para o μ. Para a estimativa do λ o estimador corrigido $\hat{\lambda}^*$ apresentou um viés bem menor significativamente em relação a todos os outros. Com relação a estimativa do EQM o $\hat{\lambda}$ apresentou um menor EQM com respeito aos demais estimadores.

Tabela 2: Estimativas dos parâmetros $\mu = 0.7$ e $\lambda = 5$.

		Estimativas de μ				Estimativas de λ			
n	Estimador	Média	Viés	Var	EQM	Média	Viés	Var	EQM
25	$\tilde{\theta}$	0.7018	0.0018	0.0022	0.0023	4.8534	−0.1465	0.4903	0.5118
	$\hat{\theta}^*$	0.6990	−0.00097	0.0022	0.0022	5.0022	0.0022	0.5218	0.5218
	$\bar{\theta}$	0.6995	−0.0004	0.0029	0.0029	-	-	-	-
	$\bar{\bar{\theta}}$	0.6962	−0.0037	0.0022	0.0022	5.4571	0.4571	0.8709	1.0799
50	$\tilde{\theta}$	0.7013	0.0013	0.0011	0.0011	4.9271	−0.0728	0.2497	0.2551
	$\hat{\theta}^*$	0.6998	−0.0001	0.0011	0.0011	5.0019	0.0019	0.2582	0.2582
	$\bar{\theta}$	0.6996	−0.0003	0.0015	0.0015	-	-	-	-
	$\bar{\bar{\theta}}$	0.6986	−0.0013	0.0011	0.0011	5.2210	0.2211	0.3239	0.3728
75	$\tilde{\theta}$	0.7010	0.0010	0.0007	0.0007	4.9532	−0.0467	0.1673	0.1695
	$\hat{\theta}^*$	0.6999	−0.00001	0.0007	0.0007	5.0035	0.0035	0.1710	0.1710
	$\bar{\theta}$	0.6996	−0.0003	0.0009	0.0009	-	-	-	-
	$\bar{\bar{\theta}}$	0.6992	−0.0007	0.0006	0.0007	5.1430	0.1430	0.2034	0.2239
100	$\tilde{\theta}$	0.7008	0.0008	0.0005	0.0005	4.9657	−0.0342	0.1239	0.1251
	$\hat{\theta}^*$	0.7001	0.0001	0.0005	0.0005	5.0029	0.0029	0.1259	0.1259
	$\bar{\theta}$	0.7000	0.0000	0.0007	0.0007	-	-	-	-
	$\bar{\bar{\theta}}$	0.6995	−0.0004	0.0005	0.0005	5.1085	0.1085	0.1440	0.1558

Fonte – Autoria própria.

A Tabela 3, apresenta as estimativas para o caso em que $\mu = 0.3$ e $\lambda = 7$. Notamos que o estimador de máxima verossimilhança corrigido mostrou-se eficaz em termos de viés para alguns tamanhos amostrais, por exemplo, para n=100, o viés de $\hat{\mu}^*$ é de 0.00048 enquanto que o $\hat{\mu}$ o viés é de 0.00052, em módulo. Em relação ao EQM o estimador que apresentou um EQM maior foi o estimador de momentos $\tilde{\mu}$, e os menores foram o $\hat{\mu}^*$ e $\bar{\mu}$.

Em relação ao parâmetro λ nota-se que o estimador $\hat{\lambda}^*$ apresentou um menor viés em todos os cenários, evidenciando o que foi mostrado nas Tabelas 1 e 2. Por exemplo, para n=50, o viés do $\hat{\lambda}^*$ é 0.0049, em módulo, já para o $\hat{\lambda}$ foi de 0.1093, em módulo, e $\bar{\lambda}$ o viés foi de 0.2929. Em relação ao EQM, o estimador de $\bar{\lambda}$ foi o que apresentou um EQM maior do que os demais. Nota-se também que, à medida que o tamanho da amostra cresce o viês vai diminuindo assim como o EQM.

Tabela 3: Estimativas dos parâmetros $\mu = 0.3$ e $\lambda = 7$

		Estimativas de μ				Estimativas de λ			
n	Estimador	Média	Viés	Var	EQM	Média	Viés	Var	EQM
25	$\hat{\theta}$	0.2965	−0.0034	0.0027	0.0027	6.7681	−0.2318	0.9233	0.9771
	$\hat{\theta}^*$	0.3005	0.0005	0.0026	0.0026	6.9764	-0.0235	0.9870	0.9876
	$\tilde{\theta}$	0.2998	-0.0001	0.0040	0.0040	-	-	-	-
	$\bar{\theta}$	0.3024	0.0024	0.0026	0.0026	7.5373	0.5373	1.6479	1.9366
50	$\hat{\theta}$	0.2970	−0.0029	0.0012	0.0012	6.8906	−0.1093	0.4766	0.4885
	$\hat{\theta}^*$	0.2990	−0.0009	0.0012	0.0012	6.9950	−0.0049	0.4926	0.4926
	$\tilde{\theta}$	0.2991	−0.0008	0.0019	0.0019	-	-	-	-
	$\bar{\theta}$	0.2999	−0.0001	0.0012	0.0012	7.2929	0.2929	0.6797	0.7655
75	$\hat{\theta}$	0.2982	−0.0017	0.0008	0.0008	6.9168	−0.0832	0.31820	0.3251
	$\hat{\theta}^*$	0.2995	−0.0004	0.0008	0.0008	6.9859	−0.0141	0.3244	0.3246
	$\tilde{\theta}$	0.2988	−0.0011	0.0012	0.0012	-	-	-	-
	$\bar{\theta}$	0.3001	0.0001	0.0007	0.0007	7.1797	0.1797	0.4296	0.4619
100	$\hat{\theta}$	0.2994	−0.00052	0.0006	0.0006	6.9517	−0.0482	0.2410	0.2433
	$\hat{\theta}^*$	0.3005	0.00048	0.0006	0.0006	7.0039	0.0039	0.2445	0.2446
	$\tilde{\theta}$	0.3007	0.0007	0.0009	0.0009	-	-	-	-
	$\bar{\theta}$	0.3008	0.0008	0.0006	0.0006	7.1457	0.1457	0.2961	0.3174

Fonte – Autoria própria.

Na Tabela 4 é apresentada as estimativas para o caso de $\mu = 0.4$ e $\lambda = 7$. Nesta situação notamos que $\hat{\mu}^*$ e $\bar{\mu}$ são os que apresentam viés menores. Contudo, há situações em que um dos dois estimadores é melhor. Por exemplo, para o tamanho amostral n=25 o viés de $\hat{\mu}^*$ é dado, por 0.0002 em módulo, enquanto que o viés de $\bar{\mu}$ é dado 0.0006 e para n=75 o viés de $\bar{\mu}$ é dado por 0.0002 em módulo, enquanto que o viés de $\hat{\mu}^*$ foi de 0.0004, apenas para n=100 o estimador de máxima verossimilhança é melhor em termos de viés do que os demais estimadores. Em termos de EQM é interessante observar que para o parâmetro μ o estimador que apresentou um EQM maior foi o $\tilde{\mu}$, em relação a todos os estimadores expostos, e à medida que o tamanho da amostra cresce, o EQM dos quatro estimadores diminui.

Em relação ao parâmetro λ nota-se que o estimador $\hat{\lambda}^*$ apresentou um viés menor em relação a todos os estimadores, por exemplo, para n=25, o $\hat{\lambda}^*$ o viés é 0.0103, enquanto que o $\hat{\lambda}$ é de 0.1978, em módulo, e o $\bar{\lambda}$ o viés é de 0.5822. Vale destacar também que o estimador que obteve o pior resultado em termos de viés foi o estimador $\bar{\lambda}$ e consequentemente foi o que apresentou maior EQM.

Tabela 4: Estimativas dos parâmetros $\mu = 0.4$ e $\lambda = 7$.

		Estimativas de μ				Estimativas de λ			
n	Estimador	Média	Viés	Var	EQM	Média	Viés	Var	EQM
25	$\hat{\theta}$	0.3976	−0.0023	0.0033	0.00337	6.8022	−0.1978	1.0066	1.0457
	$\hat{\theta}^*$	0.3998	−0.0002	0.0032	0.00324	7.0103	0.0103	1.07059	1.0707
	$\tilde{\theta}$	0.4014	0.0014	0.0050	0.0050	-	-	-	-
	$\bar{\theta}$	0.4006	0.0006	0.0031	0.00318	7.5822	0.5822	1.8039	2.1429
50	$\hat{\theta}$	0.3983	−0.00172	0.0015	0.0015	6.8902	−0.1098	0.4903	0.5024
	$\hat{\theta}^*$	0.3993	−0.0007	0.0015	0.0015	6.9946	−0.0054	0.5047	0.5048
	$\tilde{\theta}$	0.3994	−0.0006	0.0025	0.0025	-	-	-	-
	$\bar{\theta}$	0.3997	−0.0003	0.0015	0.0015	7.3041	0.3041	0.7163	0.8088
75	$\hat{\theta}$	0.3988	−0.0011	0.0010	0.0010	6.8986	−0.1013	0.3240	0.3343
	$\hat{\theta}^*$	0.3995	−0.0004	0.0010	0.0010	6.9678	−0.0321	0.3321	0.3331
	$\tilde{\theta}$	0.3996	−0.0003	0.0017	0.0017	-	-	-	-
	$\bar{\theta}$	0.3998	−0.0002	0.0010	0.0010	7.1586	0.1586	0.4331	0.4582
100	$\hat{\theta}$	0.4001	0.0001	0.0007	0.0007	6.9451	−0.0549	0.2439	0.2469
	$\hat{\theta}^*$	0.4006	0.0006	0.0007	0.0007	6.9974	−0.0026	0.2484	0.2484
	$\tilde{\theta}$	0.4004	0.0004	0.0012	0.0012	-	-	-	-
	$\bar{\theta}$	0.4008	0.0008	0.0007	0.0007	7.1439	0.1439	0.3004	0.3212

Fonte – Autoria própria.

5. APLICAÇÃO

Nesta seção, apresentamos uma aplicação a um conjunto de dados reais e compara-se a eficiência dos estimadores em termos de AIC e BIC. A base de dados pode ser encontrada em Lynn, Harvey & Nyborg (2009). Os dados são provenientes da proporção da população ateísta em 137 países.

Tabela 5: Estatísticas descritivas sobre a proporção da população ateísta

Dados	1º Quartil	Mediana	Média	3º Quartil	Variância
137	0.0050	0.0200	0.1009	0.1400	0.0250

Fonte – Autoria própria

Destaca-se na Tabela 5 que a proporção média da população ateísta nos 137 países é de aproximadamente 10%. Obtida as estimativas dos parâmetros para os estimadores precisamos identificar se os dados são de fato provenientes de uma distribuição simplex. Para isso usamos o teste de Kolmogorov - Smirnov, no software R (Team, 2014) utilizamos a função *ks.test*, no entanto, baseando na distribuição acumulada empírica provenientes dos pontos gerados pela distribuição simplex. Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida para os dados, no caso a Simplex, e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância. Considere uma amostra aleatória simples $Y_1, Y_2, \cdots, Y_n$ de uma população com função de distribuição acumulada contínua F_Y desconhecida. A estatística utilizada para o teste é:

$$D_n = sup_y |F(y) - F_n(y)| \,. \tag{33}$$

Esta função corresponde a distância máxima vertical entre os gráficos de $F(y)$ e $F_n(y)$ sobre a amplitude dos possíveis valores de y. Em D_n temos que $F(y)$ representa a função de distribuição acumulada assumida para os dados e $F_n(y)$ representa a função de distribuição acumulada empírica dos dados. Neste caso, queremos testar a hipótese $F_Y = F$ contra a hipótese alternativa $F_Y \neq F$. Para isto, tomamos $Y_{(1)}, Y_{(2)}, \cdots, Y_{(n)}$ as observações aleatórias ordenadas de forma crescente da população com função de distribuição contínua F_Y.

Segue a Tabela abaixo juntamente com o AIC e BIC.

Tabela 6: Parâmetros Estimados, AIC, BIC e *ks.test*

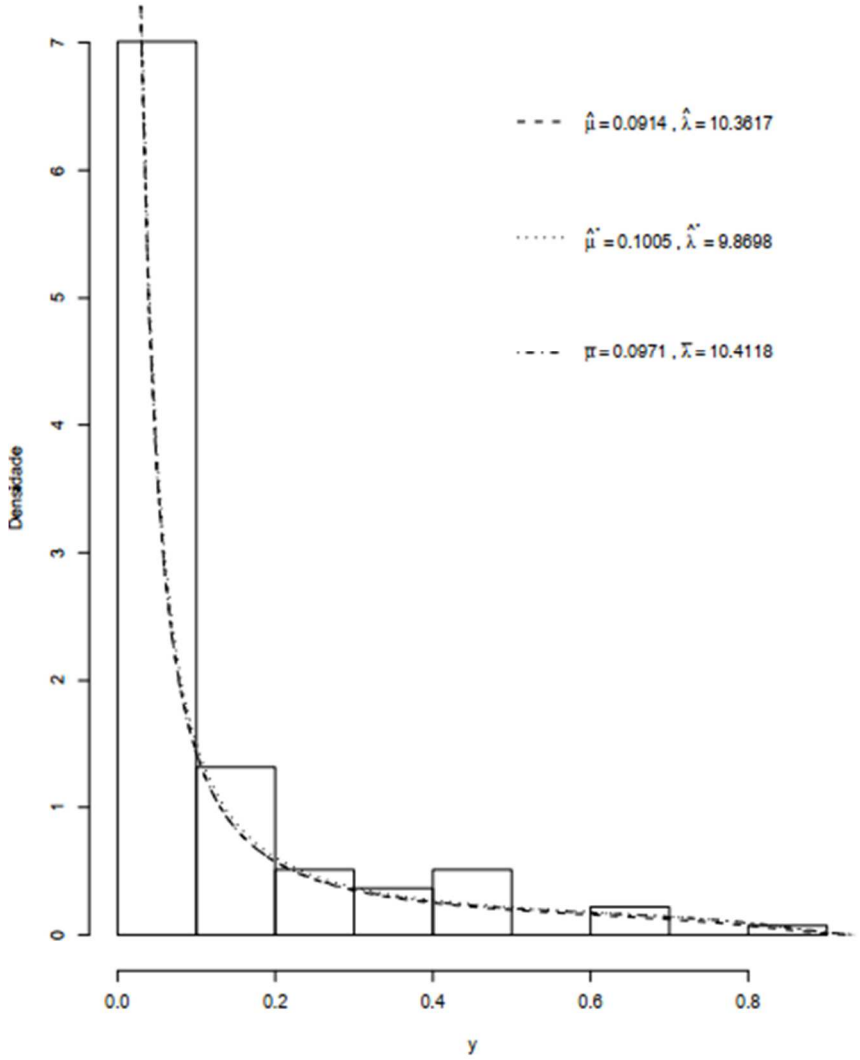
Estimador	Estimativas	AIC	BIC	ks.test (p-$valor$)
$\hat{\theta}$	$\mu = 0.0914$ $\lambda = 10.3617$	500.8684	506.7084	0.1543
$\tilde{\theta}^*$	$\mu = 0.1005$ $\lambda = 9.8698$	499.7785	505.6185	0.2912
$\tilde{\theta}$	$\mu = 0.0971$ $\lambda = 10.4118$	501.0387	506.8786	0.2392

Fonte – Autoria própria

Na Tabela 6 encontram-se as estimativas, AIC, BIC e os *p-valores* dos parâmetros do estimador de Máxima verossimilhança, a sua versão corrigida e o bayesiano. Vale ressaltar, que não foi apresentado a estimativa de momentos do λ, visto que a variância da distribuição simplex depende da função gama e não há forma analítica. Podemos notar que os *p-valores* obtidos não são significativos ao nível de significância de 10%, ou seja, não temos evidências para rejeitar a hipótese dos dados serem provenientes de uma distribuição simplex.

Assim, de forma geral o estimador de máxima verossimilhança atinge melhores resultados em termos de AIC e BIC neste conjunto de dados, pois os valores são menores em relação aos outros estimadores. Para analisarmos a adequabilidade da distribuição simplex aos dados graficamente, apresentamos um histograma juntamente com as curvas de densidades dos parâmetros estimados para avaliar a qualidade de ajuste. De acordo com a Figura 2 nota-se que as curvas de densidades dos parâmetros estimados se aproximam do histograma dos dados, evidenciando a qualidade do ajuste. Nota-se também que as três linhas do gráfico ficam próximas, justificando ao fato de que os três métodos de estimação são análogos.

Figura 2: Histograma da proporção da população ateísta em 137 países juntamente com as suas curvas de densidades dos parâmetros estimados



Fonte – Autoria própria

6.CONCLUSÕES

Para o desenvolvimento deste trabalho retratamos algumas características da distribuição Simplex e propriedades. Foram avaliados a partir de simulações de Monte Carlo os comportamentos das estimativas dessa distribuição por diferentes métodos: máxima verossimilhança, a sua respectiva versão corrigida por *bootstrap*, momentos e o estimador bayesiano.

Os resultados obtidos mostraram que a correção de viés por *bootstrap* para o método de máxima verossimilhança apresentou melhor desempenho, em termos de viés e EQM, para os dois parâmetros μ e λ. Para o parâmetro μ, o estimador de máxima verossimilhança, a sua versão corrigida por *bootstrap* e bayeseano apresentaram menor EQM em todos os cenários, e para o parâmetro λ o estimador de máxima verossimilhança apresentou menor EQM em todos os

cenários. É possível ver que, sistematicamente os vieses e EQM's do estimador bayesiano para o λ são bem maiores que os outros estimadores.

Dessa forma, para estimar os parâmetros μ e λ da distribuição Simplex, recomendamos o estimador corrigido de viés por *bootstrap*, visto que esse estimador apresentou desempenho satisfatório em termos de viés e EQM. Ao realizarmos uma aplicação a dados reais comparando os estimadores, o estimador de máxima verossimilhança constatamos o melhor desempenho em relação a sua versão corrigida e o bayesiano, baseando-se no critério do AIC e BIC.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Barndorff-Nielsen; Jørgensen, B. Some parametric models on the simplex. *Journal of Multivariate Analysis*, Elsevier, v. 39, n. 1, p. 106-116,1991.

Chib; Greenberg, E. Understanding the metropolis-hastings algorithm. *The american statistician*, Taylor and Francis Group, v. 49, n. 4, p. 327-335, 1995.

Efron, B. *Bootstrap methods: another look at the jackknife*. [S.l.]: Springer, 1992.

Gamerman; Lopes, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* [S.l.]: CRC Press, 2006.

Jørgensen, B. *The theory of dispersion models*. [S.l.]: CRC Press, 1997.

Lopes, E. A. C. *Correção de viés no modelo de regressão normal assimétrico*. Tese (Doutorado) - Universidade Federal de Pernambuco, 2007.

Lynn, R.; Harvey, J.; Nyborg, H. Average intelligence predicts atheism rates across 137 nations. *Intelligence*, Elsevier, v. 37, n. 1, p. 11-15, 2009.

Mackinnon, J. G.; Smith, A. A. Approximate bias correction in econometrics. *Journal of Econometrics*, Elsevier, v. 85, n. 2, p. 205-230, 1998.

Miyashiro, E. S. *Modelos de regressão beta e simplex para análise de proporções*. Tese (Doutorado) - Universidade de São Paulo, 2008.

Silva, A. de O. *Regressão Simplex não Linear: Inferência e Diagnóstico*. Dissertação (Mestrado) - Universidade Federal de Pernambuco, 2015.

Silva, L. C. M. da. *Coeficientes de Predição para os modelos de Regressão Beta e Simplex*. Dissertação (Mestrado) - Universidade Federal de Pernambuco, 2015.

Song; Tan, M. Marginal models for longitudinal continuous proportional data. *Biometrics*, Wiley Online Library, v. 56, n. 2, p. 496-502, 2000.

Team, R. C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*. [S.l.]: ISBN 3-900051-07-0,2014.

Yee, T. W. et al. The vgam package for categorical data analysis. *Journal of Statistical Software*, v. 32, n. 10, p. 1-34, 2010.

Revista Brasileira de Estatística, V. 77, N. 242, p.52-78, julho-dezembro 2019

UMA AVALIAÇÃO DO IMPACTO ELEITORAL DO PROGRAMA BOLSA FAMÍLIA NAS ELEIÇÕES PRESIDENCIAIS DOS ANOS 2006, 2010 E 2014

Camila Cristina Lopes

ccl1@de.ufpe.br

Universidade Federal de Pernambuco - UFPE

Francisco Cribari-Neto

cribari@de.ufpe.br

Universidade Federal de Pernambuco - UFPE

Tatiene Correia de Souza

tatiene@de.ufpb.br

Universidade Federal da Paraíba - UFPB

Resumo: Mediante a utilização de um modelo de regressão beta, o interesse deste artigo reside em avaliar e comparar as magnitudes dos impactos que os gastos com programas assistenciais (especialmente o Bolsa Família) exerceram sobre as proporções de votos válidos recebidos pelo candidato governista nos segundos turnos das eleições presidenciais de 2006, 2010 e 2014. Também, objetiva-se obter uma previsão da quantidade de votos perdidos caso o gasto *per capita* com o Bolsa Família houvesse permanecido no nível em que se encontrava na eleição anterior. A comparação entre os impactos estimados revela atenuação do impacto eleitoral com o passar do tempo. Quanto às perdas estimadas de votos caso os gastos tivessem permanecido no nível em que estavam 4 anos antes, constatou-se a situação em que o resultado da eleição não sofreria alteração, apesar da perda de votos ser apreciável, como também verificou-se a situação em que o resultado teria sido alterado.
Palavras-chave: Eleições presidenciais, Modelo de regressão beta, Programa Bolsa Família.

Abstract: Using a beta regression analysis, the chief goal of this paper is to measure impact of assistance programs spending (especially the *Bolsa Família* Program) on the proportion of valid votes in the second rounds of the 2006, 2010 and 2014 presidential elections. We obtain an estimate of the number of lost votes had per capita spending on *Bolsa Família* remained at the previous election level. The comparison of the estimated impacts of the spending on assistance programs on the proportion of valid votes for the elected presidents in the three elections show they weakened over time. In what concerns the predicted amount of votes that would have been lost had per capita spending remained at the level it was 4 years before the election, there was the case where the estimated loss of votes was reasonable, but the result of the election would not be reverted and also a situation in which such a result would be reversed.
Keywords: Presidential elections, Beta regression model, *Bolsa Família* Program.

1. INTRODUÇÃO

Sancionado pela Lei nº 10.836, o Programa Bolsa Família (PBF) surgiu em 2004 com a proposta de unificar os Programas de Transferência de Renda brasileiros. É gerenciado pelo Ministério do Desenvolvimento Social e Combate à Fome (MDS) e atua em prol da superação da pobreza.

Segundo a determinação legal, para garantir o benefício, as famílias selecionadas devem cumprir determinadas condicionalidades, fixadas nas áreas de Saúde e Educação, assim como devem estar inscritas no Cadastramento Único para Programas Sociais do Governo Federal (CadÚnico). O processo de seleção para a concessão do auxílio é automatizado e a preferência recai sobre as famílias que apresentam menor renda.

O Bolsa Família oferta quatro tipos de benefício. Em conformidade com o MDS[1], o Benefício Básico, no valor de R$ 85, é destinado às famílias consideradas em situação de extrema pobreza, isto é, com renda mensal *per capita* de até R$ 85; o Benefício Variável, no valor de R$ 39, é concedido às famílias em situação de pobreza (renda mensal *per capita* de até R$ 170) e extrema pobreza constituídas por gestantes, nutrizes e/ou crianças e adolescentes com até 15 anos de idade, havendo o limitante de 5 benefícios por unidade familiar; o Benefício Variável Vinculado ao Adolescente (BVJ), no valor de R$ 46, é remetido às famílias com jovens de 16 e 17 anos frequentadores da escola, limitado a dois auxílios por unidade familiar; por fim, o Benefício para a Superação da Extrema Pobreza (BSP) é reservado às famílias que não alcançam o limiar de R$ 85 *per capita* após o recebimento dos recursos do PBF que lhes são de direito, de modo que ao valor do auxílio é acrescentada a quantia referente ao alcance desse patamar.

Aos beneficiários é permitido o acúmulo de mais de um tipo de benefício, conforme a sua composição familiar. As referidas importâncias foram ajustadas pelo Decreto nº 8.794, de 2016, corrigindo os antigos valores e são as que vigoram atualmente. Maiores detalhes sobre o PBF podem ser obtidos em Brasil (2015).

Após cumprir dois mandatos, Lula ficou impossibilitado de concorrer a uma nova eleição e, em 2010, o PT lançou a candidatura de Dilma Vana Rousseff ao cargo máximo do Poder Executivo Federal. Ela concorreu com 8 candidatos e venceu José Serra, do PSDB, em segundo turno, tendo obtido 56,05% dos votos válidos, cerca de 55,7 milhões de votos, tornando-se a

[1] BRASIL. Ministério do Desenvolvimento Social e Combate à Fome – MDS. **Bolsa Família**: Transferência de Renda e Apoio à Família no Acesso à Saúde, à Educação e à Assistência Social. Brasília, 2015.

primeira mulher Presidente da República Federativa do Brasil. Em 2014, candidatou-se à reeleição e disputou o segundo turno com Aécio Neves da Cunha, filiado ao PSDB, derrotando-o com 51,64% dos votos válidos, aproximadamente 54,5 milhões de votos.

Após a reeleição do ex-presidente Lula, levantou-se uma polêmica sobre a influência que o investimento em programas sociais exerceu sobre o resultado da votação, destaque para o Bolsa Família. Marques et al. (2009) apontaram que o PBF foi determinante para o resultado positivo do ex-presidente na reeleição ao passo que Shikida et al. (2009) entenderam que o programa não foi decisivo.

Souza e Cribari-Neto (2013) avaliaram os impactos dos programas assistenciais e do crescimento da economia sobre o resultado da eleição presidencial de 2006, chegando à conclusão que se os gastos com programas assistenciais tivessem sido mantidos nos níveis de 2002 haveria uma redução de cerca de 7,116 milhões de votos recebidos pelo presidente e na ausência de crescimento econômico no período a redução seria de 2 milhões de votos. O resultado da eleição não seria revertido.

Trabalhos anteriores, por exemplo Zucco (2008), indicaram que houve uma mudança na base eleitoral do presidente Lula entre a eleição de 2002 e a de 2006. Na primeira eleição, o candidato obteve maior votação nas regiões brasileiras mais desenvolvidas, ao passo que na reeleição seu melhor desempenho se deu nas áreas mais pobres do país. No mesmo estudo, o autor assinalou que essa ocorrência pode estar ligada à implantação do Bolsa Família. Zucco (2010) reiterou a mutação do caráter eleitoral em Lula, como também a criação do PBF como fator de importância e acrescentou a informação que o padrão de votação observado em 2006 manteve-se em 2010 na eleição da presidente Dilma Rousseff.

Pereira et al. (2015) levantaram duas hipóteses para a similaridade no cenário de votação entre a reeleição de Lula e a eleição de Dilma: (i) os fatores econômicos e sociais que garantiram a vitória dele em regiões desfavorecidas configuraram-se como condições relevantes na eleição dela e (ii) o apoio de Lula à candidata Dilma a amparou na votação em geral, acarretando o pressuposto que o novo governo estenderia as ações de seu precursor, o que pode ter preservado a lealdade do eleitorado petista.

Outros trabalhos investigaram e avaliaram os efeitos eleitorais que o assistencialismo exerce sobre as eleições presidenciais no Brasil, tais como os de Zucco (2013), no qual são estimados os efeitos dos programas de transferência de renda nas eleições de 2002, 2006 e 2010 a partir da análise

de resultados eleitorais a nível municipal; Zucco (2015), que estendeu o trabalho anterior examinando os impactos eleitorais desses programas na eleição de 2014 e comparando-os aos resultados das três eleições anteriores; Canêdo-Pinheiro (2015), que investigou o papel do PBF e do desempenho econômico como determinantes dos resultados da eleição presidencial de 2006, principalmente na questão da migração da base eleitoral de Lula para as regiões menos desenvolvidas; e Corrêa (2015), o qual reavaliou os impactos do Bolsa Família sobre a eleição de 2006, demonstrando que à medida que a cobertura do Programa cresce, o desempenho eleitoral de Lula entre 2002 e 2006 melhora e à medida que o tamanho da classe alta cresce, pior é o desempenho.

Diante do exposto, torna-se cabível investigar, a nível nacional, a influência que os investimentos na continuidade dos programas sociais tiveram sobre a votação favorável de Luiz Inácio Lula da Silva e Dilma Rousseff no segundo turno das últimas três eleições presidenciais – 2006, 2010 e 2014 – e avaliar se os resultados das eleições sofreriam alteração caso essas ações não viessem a ser concretizadas.

Para cumprir tal objetivo, a modelagem foi feita por meio da utilização do modelo de regressão beta, indicado para dados de taxas e proporções, de forma que para cada ano eleitoral considerado um modelo para explicar a proporção de votos válidos no segundo turno da eleição foi ajustado. Para avaliar e comparar as magnitudes dos impactos que os gastos com programas assistenciais exerceram sobre as proporções de votos válidos, foram considerados os valores de gasto *per capita* anual com assistencialismo por município brasileiro. Uma análise sobre a votação final de cada eleição foi feita com base na previsão do modelo para a perda de votos caso o gasto *per capita* com o Programa Bolsa Família houvesse permanecido no nível em que se encontrava na eleição anterior.

Mais cinco seções compõem o artigo, além desta introdução. Na Seção 2 é definido o modelo de regressão beta. Os dados são descritos na Seção 3 e a especificação de cada um dos modelos de regressão ajustados é feita na Seção 4. A Seção 5 apresenta os resultados obtidos com o cálculo dos impactos juntamente com uma análise sobre o que eles representam no contexto em que estão inseridos. Finalmente, na Seção 6 constam algumas considerações gerais a respeito do trabalho.

2. MODELO DE REGRESSÃO BETA

O modelo de regressão beta surgiu a partir de uma reparametrização da densidade beta, proposta por Ferrari e Cribari-Neto (2004), de forma que fosse possível modelar a média (μ) da resposta juntamente com um parâmetro de precisão (ϕ). Assim, a função densidade é dada por

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad ,$$

em que $0 < y < 1$, $0 < \mu < 1$ e $\phi > 0$. A média e a variância de y são $\mathbb{E}(y) = \mu$ e $\mathrm{Var}(y) = \frac{V(\mu)}{1+\phi}$, respectivamente, onde $V(\mu) = \mu(1-\mu)$ é a função de variância.

Tal como feito por Souza (2011), neste trabalho admitiu-se um parâmetro de dispersão (σ) na modelagem ao invés do parâmetro de precisão. Desse modo, sendo $\sigma^2 = \frac{1}{1+\phi}$, ou $\phi = \frac{1-\sigma^2}{\sigma^2}$, a densidade passa a ser

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\mu\frac{1-\sigma^2}{\sigma^2}\right)\Gamma\left((1-\mu)\frac{1-\sigma^2}{\sigma^2}\right)} y^{\mu\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}(1-y)^{(1-\mu)\left(\frac{1-\sigma^2}{\sigma^2}\right)-1} \quad , \qquad (1)$$

em que $0 < y < 1$, $0 < \mu < 1$ e $0 < \sigma < 1$.

Ademais, admitiu-se também que o parâmetro de dispersão é variável. Logo, torna-se necessário modelar σ por meio de uma estrutura de regressão envolvendo covariáveis, parâmetros desconhecidos e uma função de ligação, similarmente ao modo que é feito para a média da resposta. Esta extensão do modelo original, denominada modelo de regressão beta com dispersão variável, foi apresentada por Simas, Barreto-Souza e Rocha (2010).

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, tal que cada y_t, $t = 1, \ldots, n$, segue a densidade em (1) com média μ_t e dispersão σ_t, desconhecidas. As estruturas de regressão para a média e dispersão de y_t são dadas, respectivamente, por

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t \qquad (2)$$

e $h(\sigma_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \upsilon_t$, (3)

em que $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}} \in \mathbb{R}^k$ e $\gamma = (\gamma_1, \ldots, \gamma_q)^{\mathrm{T}} \in \mathbb{R}^q$ são vetores de parâmetros desconhecidos, $x_{t1}, \ldots, x_{tk}$ e $z_{t1}, \ldots, z_{tq}$ são observações de k e q covariáveis

$(k + q < n)$, respectivamente, assumidas fixas e conhecidas, $\eta_t = \sum_{i=1}^{k} x_{ti}\beta_i$ e $\upsilon_t = \sum_{j=1}^{q} z_{tj}\gamma_j$ são os preditores lineares e $g(\cdot)$ e $h(\cdot)$ são funções de ligação estritamente monótonas e duas vezes diferenciáveis, em que $g: (0,1) \to \mathbb{R}$ e $h: (0,1) \to \mathbb{R}$. Por fim, tem-se que $\mu_t = g^{-1}(\eta_t)$, $\sigma_t = h^{-1}(\upsilon_t)$ e $\mathrm{Var}(y) = \mu_t(1 - \mu_t)\sigma_t^2$.

Os estimadores dos parâmetros do modelo de regressão beta são obtidos por meio da maximização do logaritmo da função de verossimilhança, todavia, os estimadores de β e γ não possuem forma fechada e, portanto, devem ser obtidos numericamente via algum algoritmo de otimização não-linear. Para grandes tamanhos de amostra e sob certas condições de regularidade, a distribuição conjunta de $\hat{\beta}$ e $\hat{\gamma}$ é aproximadamente normal, em que $\hat{\beta}$ e $\hat{\gamma}$ são os estimadores de máxima verossimilhança de β e γ, respectivamente.

3. DESCRIÇÃO DOS DADOS

As variáveis utilizadas neste estudo bem como suas respectivas descrições encontram-se na Tabela 1. O conjunto de regressores escolhidos para serem testados em cada modelo de interesse foi definido com base nos trabalhos de Zucco (2008, 2010, 2013, 2015) e Souza e Cribari-Neto (2013), que também investigaram os efeitos que o assistencialismo exerce sobre as eleições presidenciais no Brasil.

Os dados referentes ao modelo ajustado para a eleição do ano 2006 foram obtidos de Souza (2011) e os demais dados foram obtidos a partir dos bancos de dados disponibilizados pelo Tribunal Superior Eleitoral (TSE), Ministério do Desenvolvimento Social e Combate à Fome (MDS), Atlas do Desenvolvimento Humano no Brasil e Instituto Brasileiro de Geografia e Estatística (IBGE).

Visto que o PBF foi uma ação implantada no governo petista e que sua criação é um fator que pode ter influenciado o resultado das eleições presidenciais em análise, especialmente em regiões menos favorecidas, as variáveis *dummies govpt2010* e *dnord* foram incluídas no modelo. Do total de municípios considerados, 11,82% deles eram governados pelo PT no ano 2010 e 32,25% são pertencentes à região Nordeste.

Cabe salientar que no processo de seleção de um modelo adequado aos dados para o ano 2010, observou-se uma grande quantidade de pontos discrepantes, sendo eles, em sua maioria, ligados ao estado da Paraíba, de forma que todos os modelos testados sugeriam um ajuste não satisfatório.

Diante de tal fato, a variável *dummy dbp* foi acrescentada no modelo e resolveu o problema, permitindo a obtenção de ajuste adequado aos dados.

Tabela 1 - Descrição das variáveis consideradas no ajuste dos modelos

Variável	Descrição
eleicoes2014	Proporção de votos válidos de Dilma no segundo turno da eleição de 2014
gastopc2014	Gasto *per capita* com o PBF em 2014
eleicoes2010	Proporção de votos válidos de Dilma no segundo turno da eleição de 2010
gastopc2010	Gasto *per capita* com o PBF em 2010
rur2010	População rural do município em 2010
panalf2010	Proporção de analfabetos entre pessoas com 15 anos ou mais em 2010
ppent2010	Proporção da população evangélica de origem pentecostal em 2010
rendapc2010	Renda *per capita* do município em 2010
govpt2010	Variável *dummy*: 1 se o governador era do PT em 2010, 0 caso contrário
Dnord	Variável *dummy*: 1 se o município é do nordeste, 0 caso contrário
Dpb	Variável *dummy*: 1 se o município é da Paraíba, 0 caso contrário
lula2006	Proporção de votos válidos de Lula no segundo turno da eleição de 2006
gastopc2006	Gasto *per capita* com os programas assistenciais em 2006
Analf	Proporção de analfabetos entre pessoas acima de 15 anos em 2000
Renda	Renda *per capita* do município em 2000
DU	Variável *dummy*: 1 se a população do município em 2006 era maior que 50.000 habitantes, 0 caso contrário

A fim de caracterizar as variáveis em questão, na Tabela 2 podem ser visualizadas as correspondentes estatísticas descritivas calculadas tomando-se 5.560 municípios brasileiros para os anos 2010 e 2014 e 4.864 municípios para o ano 2006. É possível verificar que a renda *per capita* e o valor médio do gasto *per capita* com o PBF aumentaram com o tempo, a saber, a renda média em 2000 era de R$ 173,70 e em 2010 passou a R$ 493,10, representando um aumento de quase três vezes. Quanto ao gasto, em 2006 o valor registrado foi R$ 64,08, passando a R$ 111,40 em 2010 e R$ 201,50 em 2014, o que representou um aumento médio de R$ 137,42 *per capita* em um período de 8 anos. Observa-se ainda que em 75% dos municípios brasileiros a renda era menor ou igual a R$ 235,00 em 2000, passando a R$ 650,40 em 2010 e que o gasto com programas assistenciais que era de R$ 95,06 em 2006 atingiu a marca de R$ 311,90 no ano 2014.

Em contrapartida, constatou-se que o percentual de analfabetismo entre pessoas com 15 anos ou mais diminuiu com o tempo, de modo que a taxa média registrada na variável *analf* foi de 21,38% enquanto que em *panalf2010* o percentual baixou para 16,17%.

Tabela 2 - Estatísticas descritivas das variáveis consideradas no ajuste dos modelos

Variável	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
eleicoes2014	0,1186	0,4376	0,5692	0,5775	0,7316	0,9393
gastopc2014	1,58	79,24	162,20	201,50	311,90	832,90
eleicoes2010	0,1967	0,4757	0,5818	0,5949	0,7143	0,9650
gastopc2010	0,25	49,42	99,64	111,40	166,90	3.850,00
rur2010	0	1.600	3.234	5.348	6.768	125.300
panalf2010	0,00950	0,08088	0,13120	0,16170	0,24320	0,44400
ppent2010	0,0005	0,0595	0,0979	0,1099	0,1480	0,5211
rendapc2010	96,25	281,00	467,40	493,10	650,40	2044,00
lula2006	0,1960	0,4861	0,6143	0,6178	0,7574	0,9626
gastopc2006	2,40	31,24	54,22	64,08	95,06	384,20
analf	0,00907	0,11490	0,17570	0,21380	0,31500	0,60660
renda	28,38	88,78	163,90	173,70	235,00	954,60

Cabe salientar que os maiores percentuais de votos válidos recebidos pelos ex-presidentes Lula e Dilma nos segundos turnos das três eleições estudadas ocorreram primordialmente em municípios pertencentes à região Nordeste, ao passo que as menores votações concentraram-se nas regiões Sul e Sudeste. Comportamento similar foi observado com os valores do gasto *per capita* com o Bolsa Família.

As regiões Norte e Nordeste apareceram com as maiores taxas de analfabetismo, tendo sido registrada no Nordeste a menor renda, ao passo que a região Sul apareceu com o menor percentual de analfabetos e a região Sudeste com a maior renda *per capita* do país. Este cenário sugere que as regiões contempladas com os maiores recursos provenientes dos programas assistenciais, que apresentaram menor renda *per capita* e maior taxa de analfabetismo tiveram influência direta sobre o resultado final das eleições presidenciais em estudo.

4. ESPECIFICAÇÃO DOS MODELOS

Esta seção objetiva definir cada um dos modelos ajustados aos dados das eleições presidenciais em estudo. Todas as regressões foram estimadas utilizando o software estatístico R[2], por meio do pacote gamlss[3]. O processo de seleção das covariáveis ocorreu mediante o uso da função *stepGAIC* e também minuciosa investigação manual. A ferramenta *stepGAIC* emprega o método *stepwise* e utiliza o Critério de Informação de Akaike Generalizado (*Generalized Akaike Information Criterion* – GAIC) para a seleção de variáveis.

A variável de interesse (*y*) é definida como a proporção de votos válidos recebidos pelos ex-presidentes Lula e Dilma nos segundos turnos das eleições presidenciais de 2006, 2010 e 2014. As funções de ligação para cada modelo, para a média e dispersão, foram escolhidas de acordo com o ajuste mais satisfatório.

4.1 ESPECIFICAÇÃO DO MODELO PARA A ELEIÇÃO DE 2006

As covariáveis selecionadas para o ano 2006 foram *analf*, *renda*, *DU* e *gastopc2006*, o modelo escolhido sendo

$$\begin{aligned}&\text{logit}(\mu_t) = \beta_0 + \beta_1 analf_t + \beta_2 renda_t + \beta_3 DU_t + \beta_4 gastopc2006_t + \\ &\beta_5 gastopc2006_t \times analf_t + \beta_6 gastopc2006_t \times renda_t \ , \\ &\text{probit}(\sigma_t) = \gamma_0 + \gamma_1 analf_t + \gamma_2 renda_t \ , \qquad (4)\end{aligned}$$

para $t = 1, 2, \ldots, 4.864$.

Os coeficientes estimados das covariáveis do modelo e os correspondentes *p*-valores podem ser consultados na Tabela 3, na qual é possível averiguar que o efeito positivo sobre a proporção de votos válidos do presidente eleito no segundo turno da eleição de 2006 é proveniente da variável independente *DU*, ou seja, municípios com população maior que 50.000 habitantes registraram maior proporção de votos válidos em Lula na eleição de 2006. Todavia, diante da presença de interações no modelo, o mesmo não pode ser afirmado em relação às covariáveis *analf*, *renda* e *gastopc2006*, visto que o efeito de interação mascara o efeito principal de cada um desses regressores.

[2] Para maiores informações sobre o software R, consultar Venables e Smith (2020).
[3] Maiores detalhes relacionados à classe de modelos de regressão GAMLSS podem ser obtidos em Stasinopoulos, Rigby e Akantziliotou (2008).

Tabela 3 - Estimativas pontuais dos parâmetros do modelo 2006 e *p*-valores correspondentes

	$\hat{\beta}$	*p*-valor	$\hat{\gamma}$	*p*-valor
Intercepto	–0,202500	$3{,}53 \times 10^{-3}$	–0,548100	$< 2 \times 10^{-16}$
Analf	2,067000	$< 2 \times 10^{-16}$	–0,334000	$1{,}05 \times 10^{-4}$
Renda	–0,000997	$7{,}00 \times 10^{-9}$	–0,000656	$7{,}11 \times 10^{-11}$
DU	0,470400	$< 2 \times 10^{-16}$	---	---
gastopc2006	0,015530	$< 2 \times 10^{-16}$	---	---
gastopc2006 × *analf*	–0,017150	$2{,}97 \times 10^{-10}$	---	---
gastopc2006 × *renda*	–0,000035	$1{,}86 \times 10^{-15}$	---	---

A modelagem do parâmetro de dispersão indica que maiores valores de renda *per capita* e taxas de analfabetismo mais elevadas estão associadas com menor dispersão na proporção de votos válidos do presidente eleito no ano considerado. Ademais, todas as variáveis mostram-se significativas ao nível de significância de 5%.

4.2 ESPECIFICAÇÃO DO MODELO PARA A ELEIÇÃO DE 2010

As covariáveis selecionadas para o ano 2010 foram *rendapc2010*, *govpt2010*, *ppent2010*, *gastopc2010*, *dnord* e *dpb*. O modelo escolhido foi

$$\begin{aligned}\text{cloglog}(\mu_t) = {} & \beta_0 + \beta_1 rendapc2010_t + \beta_2 govpt2010_t + \beta_3 ppent2010_t + \\ & \beta_4 gastopc2010_t + \beta_5 dnord_t + \beta_6 dpb_t + \beta_7 gastopc2010_t \times rendapc2010_t + \\ & \beta_8 gastopc2010_t \times dnord_t \ , \\ \text{probit}(\sigma_t) = {} & \gamma_0 + \gamma_1 rendapc2010_t + \gamma_2 ppent2010_t + \gamma_3 dnord_t \ , \end{aligned} \tag{5}$$

para $t = 1, 2, \ldots, 5.560$.

Os coeficientes estimados das covariáveis do modelo e os correspondentes *p*-valores podem ser consultados na Tabela 4. Verificou-se que as covariáveis *govpt2010*, *ppent2010* e *dpb* operam negativamente sobre a proporção de votos válidos da presidente eleita no segundo turno da eleição de 2010, enquanto o efeito de interação entre *rendapc2010* e *dnord* com *gastopc2010* não permite qualquer afirmativa nesse sentido, dado que os efeitos de *rendapc2010* e *dnord* sobre a proporção média de votos válidos obtidos por Dilma dependem do nível de *gastopc2010* e vice-versa.

Tabela 4 - Estimativas pontuais dos parâmetros do modelo 2010 e *p*-valores correspondentes

	$\hat{\beta}$	*p*-valor	$\hat{\gamma}$	*p*-valor
intercepto	–0,347500	$< 2 \times 10^{-16}$	–0,765800	$< 2 \times 10^{-16}$
rendapc2010	–0,000128	$1{,}42 \times 10^{-4}$	–0,000230	$6{,}35 \times 10^{-14}$
govpt2010	–0,183900	$< 2 \times 10^{-16}$	---	---
ppent2010	–0,634600	$< 2 \times 10^{-16}$	0,597400	$4{,}36 \times 10^{-8}$
gastopc2010	0,004887	$< 2 \times 10^{-16}$	---	---
Dnord	0,756900	$< 2 \times 10^{-16}$	0,072370	$6{,}83 \times 10^{-5}$
Dpb	–0,257000	$< 2 \times 10^{-16}$	---	---
gastopc2010 × *rendapc2010*	–0,000004	$< 2 \times 10^{-16}$	---	---
gastopc2010 × *dnord*	–0,003695	$< 2 \times 10^{-16}$	---	---

Tomando as estimativas resultantes do ajuste feito para o parâmetro de dispersão, notou-se a influência positiva das covariáveis *ppent2010* e *dnord* sobre a dispersão de *y*, diferentemente do que ocorre com *rendapc2010* que mantém relação inversa com a dispersão. A análise dos *p*-valores revela que todas as variáveis são significativas ao nível nominal de 5%.

4.3 ESPECIFICAÇÃO DO MODELO PARA A ELEIÇÃO DE 2014

As covariáveis selecionadas para o ano 2014 foram *rur2010*, *panalf2010*, *ppent2010*, *dnord*, *gastopc2014* e *rendapc2010*. O modelo selecionado, por sua vez, foi

$$\begin{aligned}\text{logit}(\mu_t) = {} & \beta_0 + \beta_1 rur2010_t + \beta_2 panalf2010_t + \beta_3 ppent2010_t + \beta_4 dnord_t \\ & + \beta_5 gastopc2014_t + \beta_6 rendapc2010_t \\ & + \beta_7 panalf2010_t \times ppent2010_t \\ & + \beta_8 panalf2010_t \times gastopc2014_t \ , \\ \text{probit}(\sigma_t) = {} & \gamma_0 + \gamma_1 ppent2010_t + \gamma_2 gastopc2014_t \ , \end{aligned} \tag{6}$$

para $t = 1, 2, \ldots, 5.560$.

Os coeficientes estimados das covariáveis do modelo e os correspondentes *p*-valores estão apresentados na Tabela 5. De acordo com os resultados, nos municípios com maior população rural e pertencentes à região Nordeste a proporção de votos válidos em Dilma tende a ser maior, contrariamente ao que ocorre com *rendapc2010* que exerce efeito negativo sobre *y*. As covariáveis *panalf2010*, *ppent2010* e *gastopc2014* estão sob o

efeito de interação, o que significa que os efeitos de *ppent2010* e *gastopc2014* sobre a resposta média dependem do nível de *panalf2010*, bem como o efeito de *panalf2010* depende do nível de *ppent2010* e *gastopc2014*.

Em relação ao parâmetro de dispersão, um maior percentual da população evangélica de origem pentecostal acarreta resposta mais dispersa, ao passo que maiores valores de gasto *per capita* com programas assistenciais reduzem a dispersão. Conforme indicam os *p*-valores, as variáveis são significativas ao nível de significância de 5%.

Tabela 5 - Estimativas pontuais dos parâmetros do modelo 2014 e *p*-valores correspondentes

	$\hat{\beta}$	*p*-valor	$\hat{\gamma}$	*p*-valor
intercepto	–0,238500	$1{,}42 \times 10^{-4}$	–0,863100	$< 2 \times 10^{-16}$
rur2010	0,000005	$1{,}73 \times 10^{-6}$	---	---
panalf2010	2,161000	$< 2 \times 10^{-16}$	---	---
ppent2010	–0,423200	$2{,}33 \times 10^{-2}$	0,455300	$1{,}04 \times 10^{-5}$
dnord	0,374700	$< 2 \times 10^{-16}$	---	---
gastopc2014	0,003915	$< 2 \times 10^{-16}$	–0,000123	$7{,}26 \times 10^{-3}$
rendapc2010	–0,000530	$< 2 \times 10^{-16}$	---	---
panalf2010 × ppent2010	–5,123000	$7{,}13 \times 10^{-6}$	---	---
panalf2010 × gastopc2014	–0,006739	$< 2 \times 10^{-16}$	---	---

4.4 ANÁLISE DE DIAGNÓSTICO E ADEQUABILIDADE

Paralelamente ao ajuste de um modelo de regressão, é de extrema importância investigar sua adequação aos dados, ou em outras palavras, buscar eventuais afastamentos das suposições firmadas para o modelo. A fim de verificar a qualidade dos ajustes e garantir a boa representação dos dados, foi realizada uma análise de diagnóstico e adequabilidade para cada um dos modelos de regressão beta estimados.

As Figuras 1, 2 e 3 apresentam os gráficos de probabilidade meio-normal com envelopes simulados para o ano 2006, 2010 e 2014, respectivamente. É possível certificar que não há indícios de afastamento da suposição de que os modelos estimados são adequados para representar os dados, visto que, em geral, os resíduos encontram-se dentro das bandas de confiança dos envelopes nos três casos.

Todos os demais gráficos analisados atestaram a boa qualidade do ajuste para os três modelos estimados, indicando que o modelo de regressão beta assegura uma boa representação para os dados. Uma análise mais detalhada se encontra em Lopes (2017).

Figura 1 - Gráfico de probabilidade meio-normal com envelopes simulados do modelo para o ano 2006

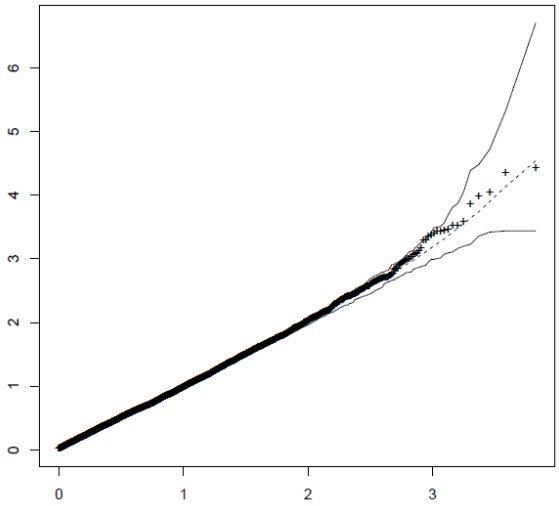


Figura 2 - Gráfico de probabilidade meio-normal com envelopes simulados do modelo para o ano 2010

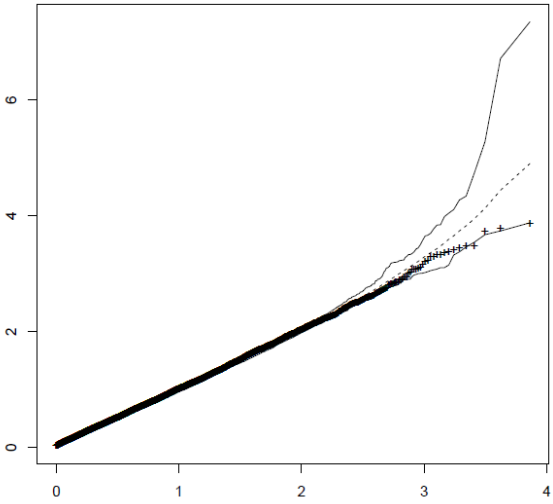
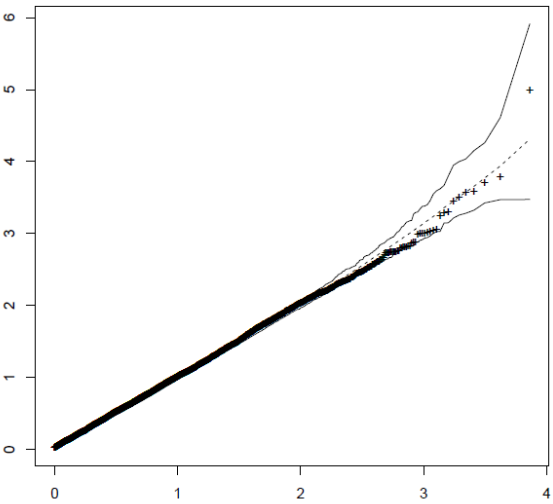


Figura 3 - Gráfico de probabilidade meio-normal com envelopes simulados do modelo para o ano 2014



A correta especificação dos modelos foi verificada por meio do teste de especificação RESET (*Regression Specification Error Test*), apresentado por Lima (2007) para modelos de regressão beta, de forma que os três modelos propostos revelaram-se corretamente especificados ao nível de significância de 5%.

Ademais, a qualidade do ajuste foi avaliada pelo cálculo do coeficiente de determinação ajustado para modelos de regressão beta (Ferrari e Cribari-Neto (2004)), conhecido como pseudo-R^2 e pelo pseudo-R^2_{RV}, apresentado por Nagelkerke (1991). Ambos representam a proporção da variabilidade dos dados que é explicada pelo modelo estimado. Dessa forma, o pseudo-R^2 obtido para o modelo ajustado para o ano 2006 foi de 0,57, o que significa que as variáveis independentes explicam cerca de 57% da variabilidade total da variável resposta, similarmente ao pseudo-R^2_{RV} que foi de 0,58. Já o modelo ajustado para o ano 2010 resultou em um pseudo-R^2 de 0,54, ou seja, cerca de 54% da variabilidade total da proporção de votos válidos da presidente Dilma no segundo turno da eleição presidencial é explicada pelas variáveis independentes, ao passo que o pseudo-R^2_{RV} foi de 0,56. Por fim, para o modelo ajustado para o ano 2014, o pseudo-R^2 foi de 0,71, isto é, a variabilidade de *y* é explicada em cerca de 71% pelas variáveis independentes e o pseudo-R^2_{RV} manteve-se em 0,72.

5. IMPACTOS ESTIMADOS

Seja $\mu = \mathbb{E}(y)$, em que *y* representa a proporção de votos válidos no segundo turno das eleições presidenciais. Define-se como medida de impacto (*I*) a derivada parcial da média da resposta (μ) em relação à covariável de interesse, neste caso, os gastos *per capita* (*gastopc*). Ou seja,

$$I_t = \frac{\partial \mathbb{E}(y_t)}{\partial gastopc_t} = \frac{\partial \mu_t}{\partial gastopc_t} \ . \tag{7}$$

Para cada impacto estimado, foi obtido o erro-padrão correspondente e os intervalos de confiança associados com nível de 95% de confiança. Esses dois processos foram realizados por meio do método bootstrap paramétrico com a utilização de 1.000 réplicas, sendo que os quantis 0,025 e 0,975 dos 1.000 impactos calculados foram usados como limites inferior e superior dos intervalos construídos.

Para obter a previsão da quantidade de votos perdidos caso o gasto *per capita* com o Programa Bolsa Família tivesse permanecido no nível em que se encontrava na eleição anterior, considerou-se o gasto *per capita* de 4 anos antes a preços da data da eleição, inflacionado pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA).

5.1 IMPACTOS ESTIMADOS DO MODELO PARA A ELEIÇÃO DE 2006

Uma vez que $\mu_t = g^{-1}(\beta_0 + \beta_1 analf_t + \beta_2 renda_t + \beta_3 DU_t + \beta_4 gastopc2006_t + \beta_5 gastopc2006_t \times analf_t + \beta_6 gastopc2006_t \times renda_t)$, o impacto exercido pela variável *gastopc2006* sobre a proporção média de votos válidos do segundo turno da eleição presidencial do ano 2006 é dado por meio da derivada

$$I_t = \frac{\partial \mu_t}{\partial gastopc2006_t}$$

$$= (\beta_4 + \beta_5 analf_t + \beta_6 renda_t) \frac{\exp(\beta_0 + \beta_1 analf_t + \beta_2 renda_t}{[1 + \exp(\beta_0 + \beta_1 analf_t + \beta_2 renda_t}$$

$$\frac{+\beta_3 DU_t + \beta_4 gastopc2006_t + \beta_5 gastopc2006_t \times analf_t + \beta_6 gastopc2006_t \times renda_t)}{+\beta_3 DU_t + \beta_4 gastopc2006_t + \beta_5 gastopc2006_t \times analf_t + \beta_6 gastopc2006_t \times renda_t)]^2}. \quad (8)$$

A Tabela 6 apresenta os impactos estimados e os respectivos erros-padrão obtidos via bootstrap. Três casos são avaliados, a saber:

- CASO 1: *renda* é fixada no primeiro quartil (Q_1); *DU*, covariável *dummy*, é fixada em 0 e 1, alternadamente; *gastopc2006* é fixado no primeiro, segundo e terceiro quartis; por fim, *analf* é fixada na mediana;
- CASO 2: *renda* é fixada no segundo quartil (Q_2) e as demais covariáveis seguem a configuração descrita no Caso 1;
- CASO 3: *renda* é fixada no terceiro quartil (Q_3) e as demais covariáveis seguem a configuração descrita no Caso 1.

Nota-se que, nos três casos estudados, o impacto de *gastopc2006* é maior quando *DU* = 0, isto é, quando se considera que as populações de todos os municípios são inferiores a 50.000 habitantes. Ademais, para todos os casos de *renda*, os maiores impactos são registrados quando *gastopc2006* é fixado no primeiro quartil. Para todos os casos, os erros-padrão bootstrap para os impactos estimados do gasto sobre a proporção média de votos válidos são valores pequenos.

Tabela 6 - Impactos estimados do gasto assistencial per capita com assistencialismo em 2006 sobre a resposta média e seus respectivos erros-padrão obtidos via bootstrap

renda	DU	Gasto PER CAPITA		
		Q_1	Q_2	Q_3
Q_1	0	0,002269 (0,000121)	0,002159 (0,000111)	0,001875 (0,000081)
	1	0,001980 (0,000113)	0,001801 (0,000097)	0,001459 (0,000064)
Q_2	0	0,001665 (0,000114)	0,001629 (0,000110)	0,001523 (0,000094)
	1	0,001504 (0,000108)	0,001422 (0,000097)	0,001257 (0,000075)
Q_3	0	0,001055 (0,000198)	0,001049 (0,000153)	0,001028 (0,000144)
	1	0,000985 (0,000145)	0,000959 (0,000138)	0,000904 (0,000120)

Os intervalos de confiança bootstrap ao nível de 95% de confiança para os impactos são apresentados na Tabela 7. É possível constatar que nenhum dos intervalos calculados contém o valor zero, o que indica que todos os impactos estimados são estatisticamente diferentes de zero ao nível nominal de 5%, isto é, tais impactos não são estatisticamente nulos.

Tabela 7 - Intervalos de confiança bootstrap ao nível de 95% de confiança para os impactos do gasto *per capita* nos programas assistenciais em 2006 sobre a resposta média

renda	DU	Gasto PER CAPITA		
		Q_1	Q_2	Q_3
Q_1	0	(0,002028; 0,002501)	(0,001937; 0,002370)	(0,001710; 0,002028)
	1	(0,001752; 0,002204)	(0,001607; 0,001987)	(0,001326; 0,001579)
Q_2	0	(0,001446; 0,001885)	(0,001417; 0,001840)	(0,001340; 0,001699)
	1	(0,001296; 0,001711)	(0,001234; 0,001607)	(0,001109; 0,001395)
Q_3	0	(0,000929; 0,001693)	(0,000764; 0,001351)	(0,000754; 0,001308)
	1	(0,000713; 0,001265)	(0,000699; 0,001219)	(0,000672; 0,001126)

A análise da Tabela 3, que apresenta as estimativas pontuais dos parâmetros do modelo 2006, revela que o coeficiente estimado da covariável *analf* é bastante elevado quando comparado às demais covariáveis. Diante disso, buscou-se verificar os comportamentos dos impactos ao fixar esta covariável no primeiro, segundo e terceiro quartis, semelhante ao modo como foi feito para *renda*. Os resultados indicaram que os impactos estimados são valores ligeiramente menores do que os obtidos quando *renda* foi fixada, contudo, o comportamento é muito similar.

Por fim, de acordo com o modelo estimado para 2006, os resultados apontam que se o gasto *per capita* com os programas assistenciais neste ano tivesse permanecido no nível em que se encontrava em 2002, considerando o *gastopc2002* corrigido pelo IPCA, o ex-presidente Lula teria obtido 52.712.657 votos, isto é, 5.582.385 votos a menos do que a quantidade conquistada por ele na eleição presidencial, cujo total foi de 58.295.042 votos. O candidato Geraldo Alckmin, que também disputou a presidência, obteve 37.543.178 votos, o que implica uma diferença de 20.751.864 votos entre os dois. Admitindo que na votação de Lula haveria tal redução e que todos os 5.582.385 votos perdidos pelo ex-presidente se *gastopc2006* = *gastopc2002* tivessem migrado para o candidato Geraldo Alckmin, este teria obtido 43.125.563 votos válidos e, portanto, segundo o modelo estimado, Lula ainda assim teria vencido a eleição presidencial ocorrida no ano 2006 mesmo se o gasto com assistencialismo tivesse permanecido no nível em que se encontrava há 4 anos.

5.2 IMPACTOS ESTIMADOS DO MODELO PARA A ELEIÇÃO DE 2010

Visto que $\mu_t = g^{-1}(\beta_0 + \beta_1 rendapc2010_t + \beta_2 govpt2010_t + \beta_3 ppent2010_t + \beta_4 gastopc2010_t + \beta_5 dnord_t + \beta_6 dpb_t + \beta_7 gastopc2010_t \times rendapc2010_t + \beta_8 gastopc2010_t \times dnord_t)$, o impacto exercido pela variável *gastopc2010* sobre a proporção média de votos válidos do segundo turno da eleição presidencial do ano 2010 é dado por meio da derivada

$$I_t = \frac{\partial \mu_t}{\partial gastopc2010_t} = \exp[-\exp(\beta_0 + \beta_1 rendapc2010_t + \beta_2 govpt2010_t + \beta_3 ppent2010_t + \beta_4 gastopc2010_t + \beta_5 dnord_t + \beta_6 dpb_t + \beta_7 gastopc2010_t \times rendapc2010_t + \beta_8 gastopc2010_t \times dnord_t)] \times \exp(\beta_0 + \beta_1 rendapc2010_t + \beta_2 govpt2010_t + \beta_3 ppent2010_t + \beta_4 gastopc2010_t + \beta_5 dnord_t + \beta_6 dpb_t + \beta_7 gastopc2010_t \times rendapc2010_t + \beta_8 gastopc2010_t \times dnord_t) \times (\beta_4 + \beta_7 rendapc2010_t + \beta_8 dnord_t)\,. \quad (9)$$

A Tabela 8 apresenta os impactos estimados e os respectivos erros-padrão obtidos via bootstrap, avaliados de forma análoga aos casos descritos anteriormente, com as três variáveis *dummy* fixadas em 0 e 1, alternadamente. Cabe ressaltar que, do total de oito cenários resultantes da combinação das três *dummies*, dois cenários foram desconsiderados por não apresentarem real sentido de ocorrência.

Nota-se que, nos três casos estudados, é evidente a diferença entre os resultados quando se considera *dnord* = 0 e *dnord* = 1, sendo que o impacto de *gastopc2010* é maior quando *dnord* = 0, isto é, quando se considera que todos os municípios não pertencem à região Nordeste. Destaca-se o fato de que, quando *dnord* = 1, os impactos têm efeito negativo para *rendapc2010* fixada no segundo e terceiro quartis. Quando *dnord* = 0, em geral, os maiores impactos são registrados quando *gastopc2010* é fixado no terceiro quartil para os três casos considerados da renda.

Entre todos os cenários considerados, o maior impacto é registrado para o caso em que todos os estados seriam governados pelo PT e todos os municípios não pertenceriam à região Nordeste nem ao estado da Paraíba, com *rendapc2010* fixada no primeiro quartil (R$ 281) e *gastopc2010* fixado no terceiro quartil (R$ 166,90), situação em que o aumento seria de 0,140720%, conforme mostra a Tabela 8. Para todos os casos, os erros-padrão bootstrap para os impactos estimados do gasto sobre a proporção média de votos válidos são valores pequenos.

Tabela 8 - Impactos estimados do gasto per capita no Programa Bolsa Família em 2010 sobre a resposta média e seus respectivos erros-padrão obtidos via bootstrap

rendapc2010	govpt2010	dnord	dpb	GASTO PER CAPITA		
				Q_1	Q_2	Q_3
Q_1	0	0	0	0,00136531 (0,00005261)	0,00140440 (0,00005690)	0,00137954 (0,00005343)
	1	0	0	0,00129364 (0,00004825)	0,00136782 (0,00003560)	0,00140720 (0,00005703)
	0	1	0	0,00004544 (0,00001879)	0,00004533 (0,00001872)	0,00004518 (0,00001858)
	1	1	0	0,00004763 (0,00001961)	0,00004758 (0,00001957)	0,00004752 (0,00001952)
	0	1	1	0,00004798 (0,00001973)	0,00004796 (0,00001972)	0,00004793 (0,00001970)
	1	1	1	0,00004773 (0,00001955)	0,00004776 (0,00001959)	0,00004781 (0,00001962)
Q_2	0	0	0	0,00109792 (0,00005104)	0,00113467 (0,00005481)	0,00114671 (0,00005434)
	1	0	0	0,00103262 (0,00004716)	0,00108961 (0,00004126)	0,00113870 (0,00005626)
	0	1	0	–0,00020331	–0,00020494 (0,00003077)	–0,00020681 (0,00003100)
	1	1	0	–0,00021030	–0,00021068 (0,00003127)	–0,00021090 (0,00003129)
	0	1	1	–0,00021091	–0,00021083 (0,00003128)	–0,00021047 (0,00003116)
	1	1	1	–0,00020766	–0,00020659 (0,00003070)	–0,00020496 (0,00003022)
Q_3	0	0	0	0,00084057 (0,00006093)	0,00086845 (0,00006477)	0,00089068 (0,00006634)
	1	0	0	0,00078517 (0,00005664)	0,00082349 (0,00005700)	0,00086440 (0,00006686)
	0	1	0	–0,00045520	–0,00046059 (0,00005522)	–0,00046455 (0,00005493)
	1	1	0	–0,00046514	–0,00046473 (0,00005435)	–0,00046141 (0,00005310)
	0	1	1	–0,00046451	–0,00046206 (0,00005401)	–0,00045624 (0,00005238)
	1	1	1	–0,00045306	–0,00044628 (0,00005192)	–0,00043533 (0,00004919)

A Tabela 9 mostra os intervalos de confiança bootstrap ao nível de 95% de confiança obtidos para os impactos. De acordo com o que é apresentado, verifica-se que todos os impactos estimados são estatisticamente diferentes de zero ao nível nominal de 5%.

Por fim, para fazer a previsão do número de votos perdidos caso o gasto *per capita* com o Programa Bolsa Família no ano 2010 tivesse permanecido no nível em que se encontrava em 2006, igualou-se o número de municípios brasileiros disponíveis para esses dois anos, de forma que 698 municípios foram desconsiderados do conjunto de dados de 2010, totalizando, assim, 4.864 municípios. Então, a partir do número estimado de votos perdidos calculados com base nessa amostra reduzida, projetou-se a proporção dos votos em relação ao ano 2006 para os 5.560 municípios disponíveis em 2010, assumindo que tal proporção se manteve constante de 2006 para 2010. Logo, os resultados apontam que, considerando o *gastopc2006* corrigido pelo IPCA, Dilma Rousseff teria obtido 52.075.946 votos, isto é, 3.676.583 votos a menos do que a quantidade conquistada por ela na eleição presidencial, cujo total foi de 55.752.529 votos. José Serra, que disputou a presidência com Dilma, obteve 43.711.388 votos, o que implica uma diferença de 12.041.141 votos entre os dois. Admitindo que na votação de Dilma haveria tal redução e se todos os 3.676.583 votos perdidos pela ex-presidente se *gastopc2010* = *gastopc2006* tivessem migrado para o candidato José Serra, este teria obtido 47.387.971 votos válidos e, portanto, segundo o modelo estimado, Dilma ainda assim teria vencido a eleição presidencial ocorrida no ano 2010 mesmo se o gasto com o Bolsa Família tivesse permanecido no nível em que se encontrava há 4 anos.

Tabela 9 - Intervalos de confiança bootstrap ao nível de 95% de confiança para os impactos do gasto *per capita* no Programa Bolsa Família em 2010 sobre a resposta média

rendapc2010	govpt2010	dnord	dpb	GASTO PER CAPITA		
				Q_1	Q_2	Q_3
Q_1	0	0	0	(0,00126533; 0,00146830)	(0,00129597; 0,00151671)	(0,00127727; 0,00148465)
	1	0	0	(0,00120284; 0,00138442)	(0,00129441; 0,00143547)	(0,00129811; 0,00152005)
	0	1	0	(0,00000760; 0,00008211)	(0,00000752; 0,00008205)	(0,00000759; 0,00008125)
	1	1	0	(0,00000796; 0,00008581)	(0,00000796; 0,00008567)	(0,00000796; 0,00008545)
	0	1	1	(0,00000803; 0,00008654)	(0,00000803; 0,00008652)	(0,00000803; 0,00008647)
	1	1	1	(0,00000800; 0,00008550)	(0,00000800; 0,00008565)	(0,00000800; 0,00008582)
Q_2	0	0	0	(0,00100023; 0,00119861)	(0,00103107; 0,00124246)	(0,00104221; 0,00125341)
	1	0	0	(0,00094111; 0,00112648)	(0,00101032; 0,00117501)	(0,00103165; 0,00124895)
	0	1	0	(–0,00025967; –0,00014305)	(–0,00026263; –0,00014366)	(–0,00026523; –0,00014487)
	1	1	0	(–0,00026834; –0,00014814)	(–0,00026883; –0,00014830)	(–0,00026898; –0,00014841)
	0	1	1	(–0,00026885; –0,00014835)	(–0,00026898; –0,00014822)	(–0,00026867; –0,00014796)
	1	1	1	(–0,00026555; –0,00014563)	(–0,00026374; –0,00014511)	(–0,00026090; –0,00014428)
Q_3	0	0	0	(0,00072618; 0,00096437)	(0,00074804; 0,00100022)	(0,00076774; 0,00102560)
	1	0	0	(0,00068103; 0,00089830)	(0,00071769; 0,00093438)	(0,00074232; 0,00099861)
	0	1	0	(–0,00055981; –0,00034473)	(–0,00056867; –0,00035002)	(–0,00057111; –0,00035011)
	1	1	0	(–0,00057065; –0,00035188)	(–0,00057172; –0,00035194)	(–0,00056562; –0,00035079)
	0	1	1	(–0,00057177; –0,00035179)	(–0,00056777; –0,00035027)	(–0,00055863; –0,00034713)
	1	1	1	(–0,00055809; –0,00034390)	(–0,00054772; –0,00034034)	(–0,00053097; –0,00033373)

5.3 IMPACTOS ESTIMADOS DO MODELO PARA A ELEIÇÃO DE 2014

Sabendo que $\mu_t = g^{-1}(\beta_0 + \beta_1 rur2010_t + \beta_2 panalf2010_t + \beta_3 ppent2010_t + \beta_4 dnord_t + \beta_5 gastopc2014_t + \beta_6 rendapc2010_t + \beta_7 panalf2010_t \times ppent2010_t + \beta_8 panalf2010_t \times gastopc2014_t)$, o impacto exercido pela variável *gastopc2014* sobre a proporção média de votos válidos do segundo turno da eleição presidencial do ano 2014 é dado por meio da derivada

$$I_t = \frac{\partial \mu_t}{\partial gastopc2014_t}$$

$$= (\beta_5 + \beta_8 panalf2010_t) \frac{\exp(\beta_0 + \beta_1 rur2010_t + \beta_2 panalf2010_t}{[1 + \exp(\beta_0 + \beta_1 rur2010_t + \beta_2 pnalf2010_t}$$

$$\frac{+\beta_3 ppent2010_t + \beta_4 dnord_t + \beta_5 gastopc2014_t + \beta_6 rendapc2010_t}{+\beta_3 ppent2010_t + \beta_4 dnord_t + \beta_5 gastopc2014_t + \beta_6 rendapc2010_t}$$

$$\frac{+\beta_7 panalf2010_t \times ppent2010_t + \beta_8 panalf2010_t \times gastopc2014_t)}{+\beta_7 panalf2010_t \times ppent2010_t + \beta_8 panalf2010_t \times gastopc2014_t)]^2} . \qquad (10)$$

A Tabela 10 apresenta os resultados para os impactos estimados e os respectivos erros-padrão obtidos via bootstrap, avaliados de forma análoga aos três casos descritos em 5.1. Nota-se que, nos três casos estudados, o impacto de *gastopc2014* é maior quando *dnord* = 0, isto é, quando se considera que todos os municípios não pertencem à região Nordeste. Ademais, quando *rendapc2010* é fixada no primeiro e segundo quartis, os maiores impactos são observados para *gastopc2014* fixado no primeiro quartil, enquanto que para *rendapc2010* fixada no terceiro quartil, o maior impacto é registrado para *gastopc2014* fixado no segundo quartil. Para todos os casos, os erros-padrão bootstrap para os impactos estimados do gasto sobre a proporção média de votos válidos são valores pequenos.

Tabela 10 - Impactos estimados do gasto *per capita* no Programa Bolsa Família em 2014 sobre a resposta média e seus respectivos erros-padrão obtidos via bootstrap

rendapc2010	dnord	Gasto PER CAPITA		
		Q_1	Q_2	Q_3
Q_1	0	0,000757 (0,000026)	0,000741 (0,000026)	0,000661 (0,000020)
	1	0,000725 (0,000027)	0,000679 (0,000025)	0,000561 (0,000018)
Q_2	0	0,000757 (0,000026)	0,000750 (0,000025)	0,000683 (0,000020)
	1	0,000739 (0,000026)	0,000699 (0,000024)	0,000589 (0,000017)
Q_3	0	0,000753 (0,000026)	0,000756 (0,000025)	0,000703 (0,000020)
	1	0,000748 (0,000026)	0,000717 (0,000024)	0,000615 (0,000017)

Na Tabela 11 estão os intervalos de confiança bootstrap ao nível de 95% de confiança obtidos para os impactos. Nenhum dos intervalos apresentados inclui o valor zero, portanto, todos os impactos estimados são estatisticamente diferentes de zero ao nível nominal de 5%.

Tabela 11 - Intervalos de confiança bootstrap ao nível de 95% de confiança para os impactos do gasto *per capita* no Programa Bolsa Família em 2014 sobre a resposta média

rendapc2010	dnord	GASTO PER CAPITA		
		Q_1	Q_2	Q_3
Q_1	0	(0,000707; 0,000806)	(0,000691; 0,000790)	(0,000621; 0,000699)
	1	(0,000673; 0,000777)	(0,000630; 0,000726)	(0,000527; 0,000595)
Q_2	0	(0,000707; 0,000805)	(0,000701; 0,000798)	(0,000644; 0,000721)
	1	(0,000687; 0,000788)	(0,000652; 0,000745)	(0,000556; 0,000621)
Q_3	0	(0,000704; 0,000801)	(0,000706; 0,000803)	(0,000663; 0,000740)
	1	(0,000698; 0,000797)	(0,000670; 0,000762)	(0,000583; 0,000647)

Tanto no modelo ajustado para o ano 2006 quanto para o ano 2014, a taxa de analfabetismo se mostrou uma variável significativa e, igualmente ao que ocorreu com o modelo 2006, o coeficiente estimado desta covariável no modelo 2014 é elevado (vide Tabela 5). Logo, o comportamento dos impactos ao fixar *panalf2010* no primeiro, segundo e terceiro quartis foi investigado e os resultados apontaram que não ocorre grandes mudanças em relação ao que é observado quando a covariável *rendapc2010* é fixada.

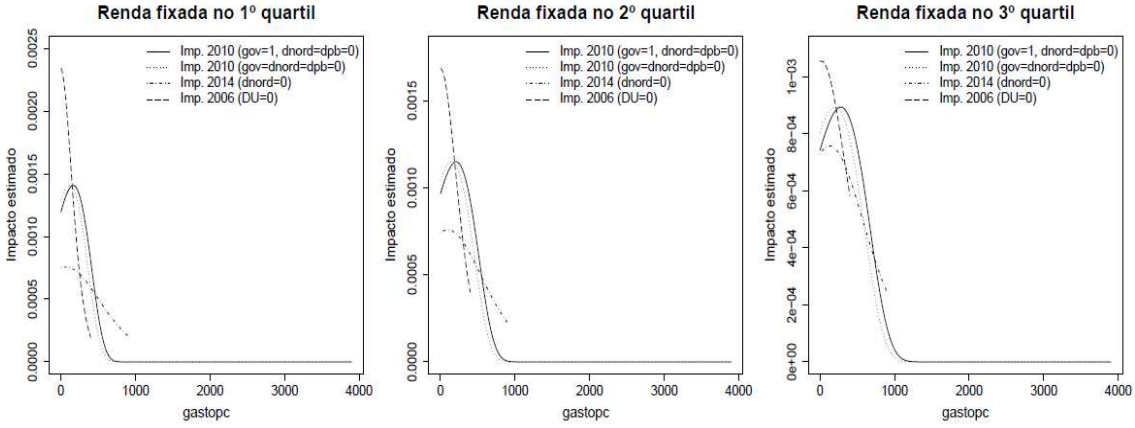
Por fim, de acordo com o modelo estimado para 2014, os resultados apontam que se o gasto *per capita* com o Programa Bolsa Família neste ano tivesse permanecido no nível em que se encontrava em 2010, com o *gastopc2010* corrigido pelo IPCA, Dilma Rousseff teria obtido 49.288.870 votos, isto é, 5.212.248 votos a menos do que a quantidade conquistada por ela na eleição presidencial, cujo total foi de 54.501.118 votos. Aécio Neves, que disputou a presidência com Dilma, obteve 51.041.155 votos, o que implica uma diferença de 3.459.963 votos entre ambos. Admitindo que na votação de Dilma haveria tal redução e se todos os 5.212.248 votos perdidos pela ex-presidente se *gastopc2014* = *gastopc2010* tivessem migrado para seu oponente, este teria obtido 56.253.403 votos válidos e, portanto, segundo o modelo estimado, o resultado final da eleição sofreria alteração e Aécio Neves teria vencido a eleição presidencial ocorrida no ano 2014. Salienta-se, também, que ainda que o total de votos perdidos por Dilma tivesse se transformado em votos brancos e nulos, Aécio Neves sairia vencedor da eleição.

Ademais, buscou-se determinar o menor percentual do total de votos perdidos por Dilma Rousseff de modo que se tais votos tivessem migrado para Aécio Neves, o resultado da eleição teria sido revertido e verificou-se que se 33,6186% dos votos válidos perdidos por Dilma tivesse migrado para seu oponente, essa quantidade bastaria para que Aécio Neves vencesse a eleição do ano em questão.

5.4 COMPARAÇÃO ENTRE AS ELEIÇÕES DE 2006, 2010 E 2014

O comportamento dos impactos estimados para cada ano estudado pode ser visualizado nas Figuras 4 e 5. A Figura 4 considera o caso em que as covariáveis *dummies*, *dnord* e *DU*, assumem o valor 0 e por meio dela verifica-se que, nos três casos considerados para a renda, o impacto dos gastos com programas assistenciais sobre a proporção de votos recebidos pelos presidentes eleitos em 2006, 2010 e 2014 diminui com o passar do tempo, isto é, a quantidade de votos que podem ser atribuídos aos benefícios do assistencialismo é menor a cada nova eleição.

Figura 4 - Impactos estimados do gasto *per capita* nos programas assistenciais sobre a resposta média quando *dnord* = 0 e *DU* = 0

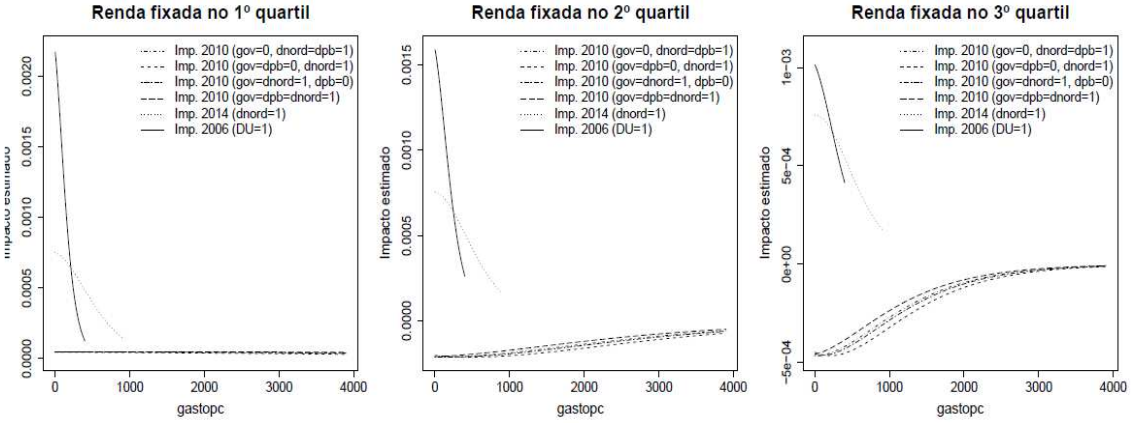


Em todos os casos avaliados, valores menores de gasto implicam impactos de maiores proporções. Contudo, conforme o valor do gasto aumenta, a proporção de votos ganhos devido ao assistencialismo diminui consideravelmente nos três anos. Isto pode ser explicado pelo fato de que o número de municípios com valores altos de gasto *per capita* é pequeno, o que implica menores proporções de votos.

Na Figura 5 observa-se o caso em que as covariáveis *dummies*, *dnord* e *DU*, assumem o valor 1. As curvas de impacto mostram que, nos três casos considerados para a renda, o impacto dos gastos com programas assistenciais sobre a proporção de votos recebidos pela presidente eleita em 2010 difere bastante do comportamento apresentado pelos anos 2006 e 2014 em dois

aspectos: os impactos estimados são relativamente menores, registrando valores negativos quando a renda é fixada no segundo e terceiro quartis e, para os mesmos quartis, o impacto é crescente à medida que o gasto *per capita* aumenta, ao passo que em todos os quartis dos outros dois anos avaliados tem-se um comportamento decrescente. Por fim, averígua-se que a diferença entre os impactos registrados nos anos 2006 e 2014 para os menores valores do gasto *per capita* diminui conforme a renda aumenta.

Figura 5 - Impactos estimados do gasto per capita nos programas assistenciais sobre a resposta média quando *dnord* = 1 e *DU* = 1



6. CONSIDERAÇÕES FINAIS

A avaliação da magnitude dos impactos exercidos pelo aumento no gasto *per capita* com programas assistenciais, especialmente o Programa Bolsa Família, sobre o resultado das três últimas eleições presidenciais ocorridas nos anos 2006, 2010 e 2014, proposta deste trabalho, foi feita por meio da utilização do modelo de regressão beta com dispersão variável, uma vez que a variável de interesse é a proporção de votos válidos recebidos pelos ex-presidentes Lula (2006) e Dilma Rousseff (2010 e 2014) no segundo turno das três eleições consideradas. Desse modo, obteve-se um submodelo para a média e outro para a dispersão da variável resposta de cada ano.

A partir do modelo ajustado e corretamente especificado, para cada eleição obteve-se uma previsão do número de votos perdidos caso o gasto *per capita* com assistencialismo não tivesse aumentado nos últimos 4 anos, considerando o gasto *per capita* no nível da eleição anterior corrigido pelo IPCA. Quando considerados os anos 2006 e 2010, a redução na votação final dos ex-presidentes não seria significativa a ponto de alterar o resultado da eleição. Contudo, ocorrendo a redução prevista na votação final da eleição do ano 2014, Dilma Rousseff não conseguiria um número suficiente de votos para

se reeleger e, portanto, Aécio Neves teria se tornado o presidente do Brasil naquele ano.

A comparação entre os impactos estimados dos gastos com programas assistenciais sobre a proporção de votos válidos recebidos pelos presidentes eleitos no segundo turno das três eleições revelou que o impacto dos gastos com programas assistenciais sobre a proporção de votos recebidos pelos presidentes eleitos em 2006, 2010 e 2014 diminui com o passar do tempo, isto é, a quantidade de votos recebida devido aos benefícios do assistencialismo é menor a cada nova eleição e, ainda, valores menores de gasto implicam impactos de maiores proporções. Ademais, notou-se que à medida que o valor do gasto aumenta, a proporção de votos ganhos devido ao assistencialismo diminui nos três anos, o que pode ser explicado pelo fato de que o número de municípios com valores elevados de gasto *per capita* é pequeno, implicando menores proporções de votos.

Nesse contexto, evidencia-se que os gastos com assistencialismo, principalmente com o Programa Bolsa Família, foco desta pesquisa, exercem influência sobre a votação dos candidatos à presidência. Em períodos mais remotos a influência era maior, isto é, acarretava maior quantidade de votos favoráveis aos presidentes, entretanto, com o passar dos anos, este cenário foi enfraquecendo, tanto que o resultado da eleição mais recente poderia ter sido alterado com base no número de votos perdidos caso tais gastos não tivessem aumentado nos últimos 4 anos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

BRASIL. Ministério do Desenvolvimento Social e Combate à Fome – MDS. Bolsa Família: Transferência de Renda e Apoio à Família no Acesso à Saúde, à Educação e à Assistência Social. Brasília, 2015.

CANÊDO-PINHEIRO, M. Bolsa Família ou desempenho da economia? Determinantes da reeleição de Lula em 2006. Economia Aplicada, v. 19, n. 1, p. 31–61, 2015.

CORRÊA, D. S. Os custos eleitorais do Bolsa Família: Reavaliando seu impacto sobre a eleição presidencial de 2006. Opinião Pública, v. 21, n. 3, p. 514–534, 2015.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. Journal of Applied Statistics, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

LIMA, L. B. de. Um teste de especificação correta em modelos de regressão beta. Dissertação (Mestrado) – Universidade Federal de Pernambuco, 2007.

LOPES, C. C. Uma avaliação do impacto eleitoral do Programa Bolsa Família. Dissertação (Mestrado) – Universidade Federal de Pernambuco, 2017.

MARQUES, R. M. et al. Discutindo o papel do Programa Bolsa Família na decisão das eleições presidenciais brasileiras de 2006. Revista de Economia Política, v. 29, n. 1, p. 114–132, 2009.

NAGELKERKE, N. J. A note on a general definition of the coeficient of determination. Biometrika, v. 78, n. 3, p. 691–692, 1991.

PEREIRA, A. E. G. et al. A eleição de Dilma em 2010 e seus determinantes: evidências empíricas do Programa Bolsa Família. Análise Econômica, v. 33, n. 64, p. 111–142, 2015.

SHIKIDA, C. D. et al. “It is the economy, companheiro!” : an empirical analysis of Lula's re-election based on municipal data. Economics Bulletin, v. 29, n. 2, p. 976–991, 2009.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. Computational Statistics & Data Analysis, v. 54, n. 2, p. 348–366, 2010.

SOUZA, T. C. de. Ensaios sobre modelos de regressão com dispersão variável. Tese (Doutorado) – Universidade Federal de Pernambuco, 2011.

SOUZA, T. C. de; CRIBARI-NETO, F. Uma estimativa do impacto eleitoral do Programa Bolsa-Família. Revista Brasileira de Biometria, v. 31, n. 1, p. 79–103, 2013.

STASINOPOULOS, D. M.; RIGBY, R. A.; AKANTZILIOTOU, C. Instructions on how to use the gamlss package in R: Second edition. [S.l.], 2008.

VENABLES, W. N.; SMITH, D. M. An introduction to R: Notes on R: A programming environment for data analysis and graphics. [S.l.], 2020.

ZUCCO, C. The president's 'new' constituency: Lula and the pragmatic vote in Brazil's 2006 presidential elections. Journal of Latin American Studies, v. 40, n. 1, p. 29–49, 2008.

ZUCCO, C. Poor voters vs. poor places: persisting patterns in presidential elections in Brazil. In: Workshop on Political Consequences of Declining Inequality in Brazil. Oxford: [s.n.], 2010.

ZUCCO, C. When payouts pay off: Conditional cash transfers and voting behavior in Brazil 2002–10. American Journal of Political Science, v. 57, n. 4, p. 810–822, 2013.

ZUCCO, C. The impacts of conditional cash transfers in four presidential elections (2002–2014). Brazilian Political Science Review, v. 9, n. 1, p. 135–149, 2015.

Revista Brasileira de estatística, V. 77, N. 242, p.79-98, julho/dezembro 2019

GEOPROCESSAMENTO NA GESTÃO URBANA: ESTUDO DA DISTRIBUIÇÃO ESPACIAL DAS ESCOLAS PÚBLICAS E VULNERABILIDADE SOCIAL NO MUNICÍPIO DE RONDONÓPOLIS- MT

Luiz Geraldo Mendes

anton_lgm@hotmail.com

Universidade Candido Mendes – UCAM

Resumo: A problemática do planejamento e a gestão pública de redes de equipamentos de ensino têm sido investigadas em muitos Municípios e Estados. Uma parte dessa atenção tem sido dirigida ao desenvolvimento de modelos de otimização destinados a apoiar os processos de tomadas de decisão relativos à localização e capacidade dos equipamentos, atendimento da demanda escolar e rendimento escolar. Neste sentido, nosso estudo visou o mapeamento e análise da distribuição espacial das escolas públicas e privadas na Zona Urbana do Município de Rondonópolis – MT, correlacionando com o estudo de vulnerabilidade social, para verificar se há influência na produção de desigualdade social, usando como indicador o Índice de Desenvolvimento da Educação Básica-IDEB das Unidades Escolares. Considerou-se nesta perspectiva, variáveis de atendimento do Ensino Fundamental – Anos Iniciais e Finais, e Ensino Médio bem como variáveis associadas aos responsáveis e condições dos domicílios dos setores censitários (renda, faixa etária, analfabetismo).
Palavras-chave: Gestão Urbana. Localização de Unidades Escolares. Geoprocessamento. Vulnerabilidade Social. Áreas prioritárias.

Abstract: The problem of planning and public management of teaching equipment networks has been investigated in many municipalities and states. Part of this attention has been directed to the development of optimization models to support the decision-making processes related to equipment location and capacity, attendance to school demand and school performance. In this sense, our study aimed at mapping and analyzing the spatial distribution of public and private schools in the urban area of Rondonópolis - MT, correlating with the study of social vulnerability, to verify if there is influence in the production of social inequality, using as an indicator the Basic Education Development Index (IDEB) of the School Units. In this perspective, variables of primary school attendance - Initial and Final Years, and High School, as well as variables associated to the heads and conditions of households in the census tracts (income, age, illiteracy) were considered.

Keywords: Urban Management. Location of School Units. Geoprocessing. Social vulnerability. Priority areas.

1. INTRODUÇÃO

O planejamento da Rede Pública Escolar de Ensino exige que sejam compreendidas todas as formas de abordagem das questões, estudos, diagnósticos, propostas, de forma a permitir intervenções estruturais e fundamentais. No nível de microplanejamento é viável um modelo de localização dos prédios escolares, avaliando o comportamento de cada uma das variáveis: capacidade de atendimento dos prédios escolares, percentual de atendimento da população total, densidade demográfica, raio de abrangência dos estabelecimentos de ensino, rendimento escolar, possibilitando o estabelecimento de novas políticas educacionais. Este estudo pode ser realizado através do levantamento de informações referentes a dados demográficos dos diferentes setores censitários, bairros e Municípios, correlacionando-os com dados do atendimento ofertado nas Unidades Escolares, considerando as modalidades de ensino que elas ofertam e a localização geográfica delas. Com a correlação das informações demográficas e escolares, utilizando-se da base de mapas dos Estados e dos Municípios, e com a elaboração de mapas temáticos, podemos apresentar um panorama do atendimento do Ensino Básico nos Estados e Municípios.

A visualização e análise das áreas onde estão disponibilizadas as Unidades Escolares por tipo de atendimento, correlacionada com a população escolarizável predominante nessas áreas, nos permite identificar onde há déficit ou superávit de atendimento e se há correlação dos dados do IDEB com a vulnerabilidade social da região de localização das mesmas. Isto nos permite identificar as áreas/Unidades Escolares que mais necessitam de investimentos na Educação.

Para auxiliar pesquisas de dados de forma georreferenciada de forma a contribuir com a reformulação de políticas públicas, alguns autores utilizam ferramentas de Geotecnologias. Dentre as ferramentas disponíveis está o Geoprocessamento que vem sendo utilizado em diversas áreas. Dias (2001) e Rodrigues (1993) definem bem o termo Geoprocessamento,

> denota a disciplina do conhecimento que utiliza técnicas matemáticas e computacionais para o tratamento da informação geográfica e que vem influenciando de maneira crescente as áreas de Cartografia, Análise de Recursos Naturais, Transportes, Comunicações, Energia e Planejamento Urbano e Regional". Câmara e Davis (2001, p.1).
> é um conjunto de tecnologias de coleta, tratamento, manipulação e apresentação de informações espaciais, voltado para um objetivo específico. (Rodrigues,1993).

Esta pesquisa teve como objetivo, verificar se há correlação entre a distribuição das escolas públicas da Zona Urbana do Município de Rondonópolis-MT e a vulnerabilidade social das regiões de localização das referidas escolas.

Utilizamos como referência a metodologia e dados apresentados nos trabalhos de Torres e Negri (2014), Dias (2013), Rezende (2016), Érnica e Batista (2012), atualizando alguns dados de acordo com os dados populacionais e de atendimento e correlacionando com o estudo de vulnerabilidade social, verificando se há influência na produção de desigualdade social, associando com o Índice de Desenvolvimento da Educação Básica-IDEB, através de mapas temáticos, usando o software livre QGis.

A construção de Índices de Vulnerabilidade Social feita com análise das informações sobre a população com a escolha de variáveis obtidas em Censos Demográficos é realizada por institutos e fundações tais como : IPECE (Instituto de Pesquisa e Estratégia Econômica do Ceará), IPVS-SEADE(Índice Paulista de Vulnerabilidade Social - Fundação Sistema Estadual de Análise de Dados), IPEA(Instituto de Pesquisa Econômica Aplicada). Alguns estão disponíveis a nível de Municípios, Estados e Brasil, como os disponibilizados pelo IPEA. Para ter uma visão mais detalhada, alguns institutos (IPECE, IPVS-SEADE), desagregam a análise a nível de setor censitário. Como a análise é feita utilizando dados do Censos Demográficos, temos que considerar o lapso temporal desses índices.

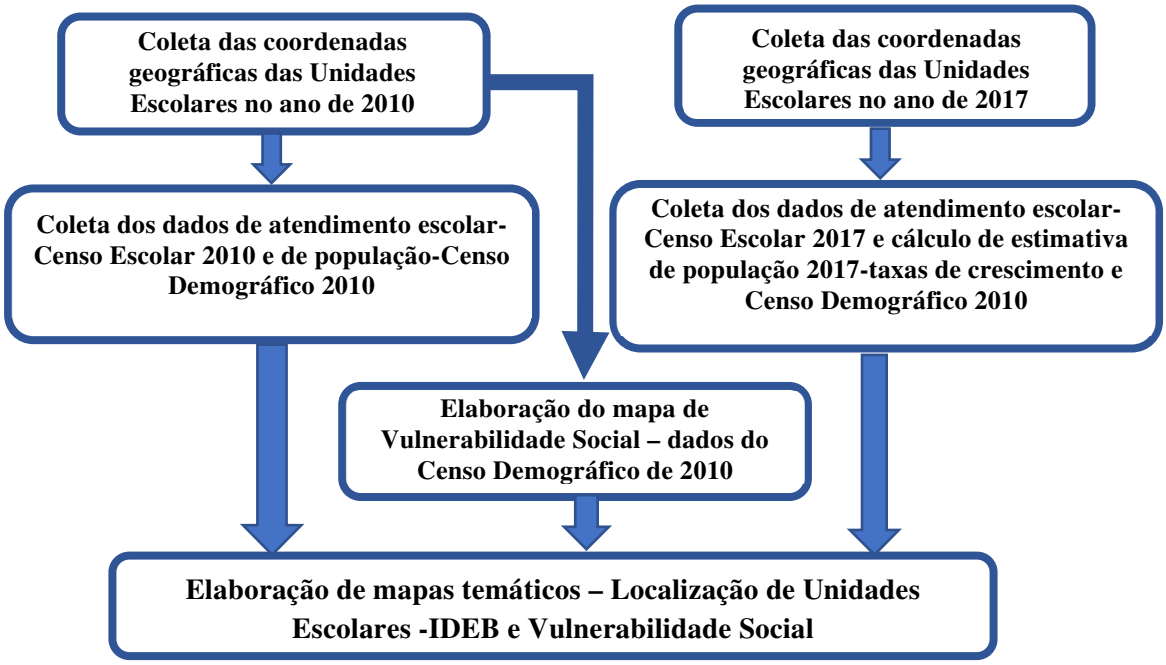
Foi elaborado um Índice de Vulnerabilidade Social – IVS, considerando as dimensões de habitação e infraestrutura, educação, renda e estrutura etária baseado na metodologia de Dias (2013).

2. DESENVOLVIMENTO

Para a viabilização desta pesquisa realizou-se inicialmente um levantamento dos dados com a coleta de coordenadas geográficas das Unidades Escolares Públicas. Na sequência, houve a construção de um banco de dados correlacionando os dados de localização das Unidades Escolares com os dados do Censo Demográfico 2010 do IBGE e finalmente a produção de mapas temáticos para os anos de 2010 e de 2017. O levantamento de dados constou de quatro etapas específicas. A primeira corresponde ao mapeamento das Unidades Escolares na área urbana de Rondonópolis – MT, nos anos de 2010 e 2017. Na segunda etapa foram coletados dados populacionais e de atendimento escolar com base nos Censo Demográfico de 2010 e Censo Escolar de 2010 e Censo Escolar de 2017. Foram calculadas as estimativas da

população na faixa escolarizável de 0 a 17 anos para os setores censitários para o ano de 2017. Na terceira etapa foi elaborado o mapa de vulnerabilidade social, com uso da metodologia de Dias (2013). Na última etapa foram elaborados mapas temáticos usando o software livre de geoprocessamento QGis, com a correlação dos dados de atendimento das Unidades Escolares e de faixas etárias de atendimento do Ensino Fundamental e Ensino Médio (Figura 1) e realizado o estudo de correlação de dados usando a ferramenta de análise de dados por correlação do EXCEL.

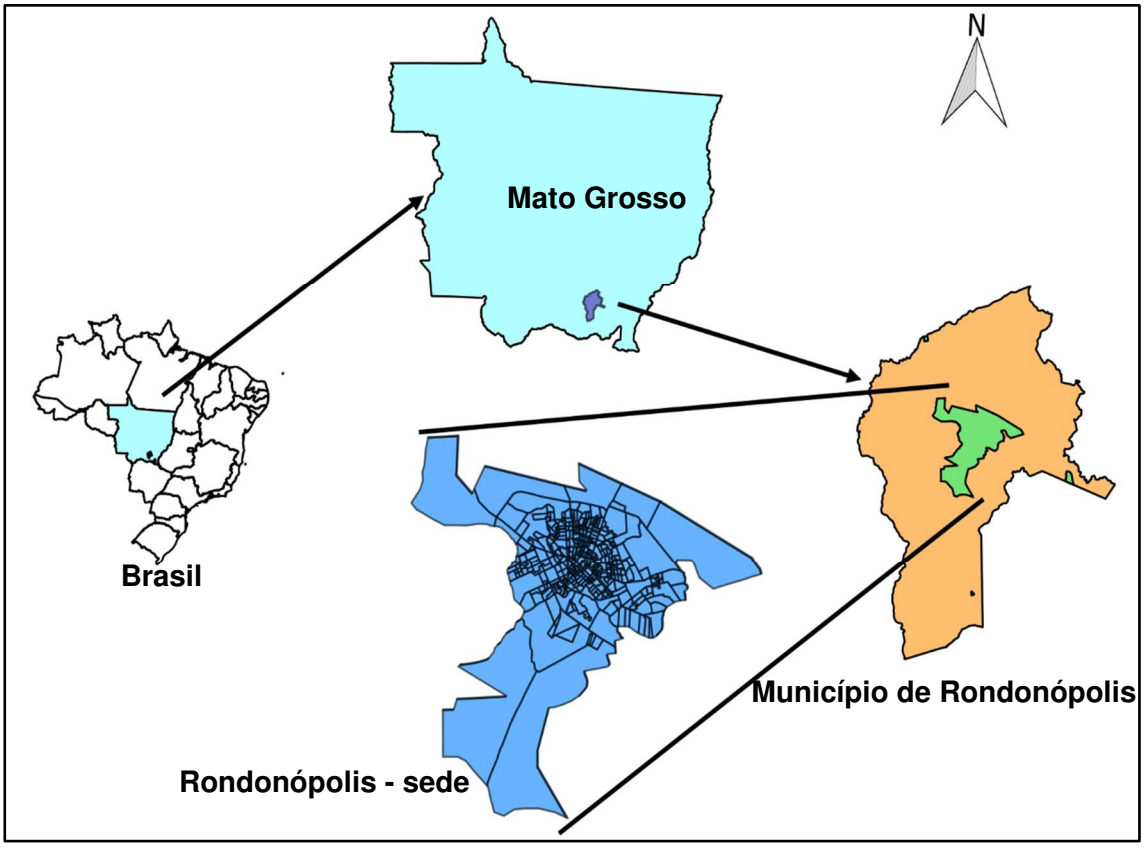
Figura 1 - Etapas metodológicas



3. RESULTADOS E DISCUSSÕES

O Município de Rondonópolis localiza-se na região Sudeste do Estado de Mato Grosso, a 228 km da capital Cuiabá, sendo o 3º Município do Estado com maior população. Possui 222.316 habitantes (2017) (Figura 2).

Figura 2 - Localização do Município de Rondonópolis/MT/Brasil



Fonte de dados: IBGE, 2010. Sem escala. Editado pelo autor (2018)

No Município de Rondonópolis - MT, nos anos de 2010 e 2017, tínhamos o número de escolas, por dependência administrativa, indicado na Tabela 1.

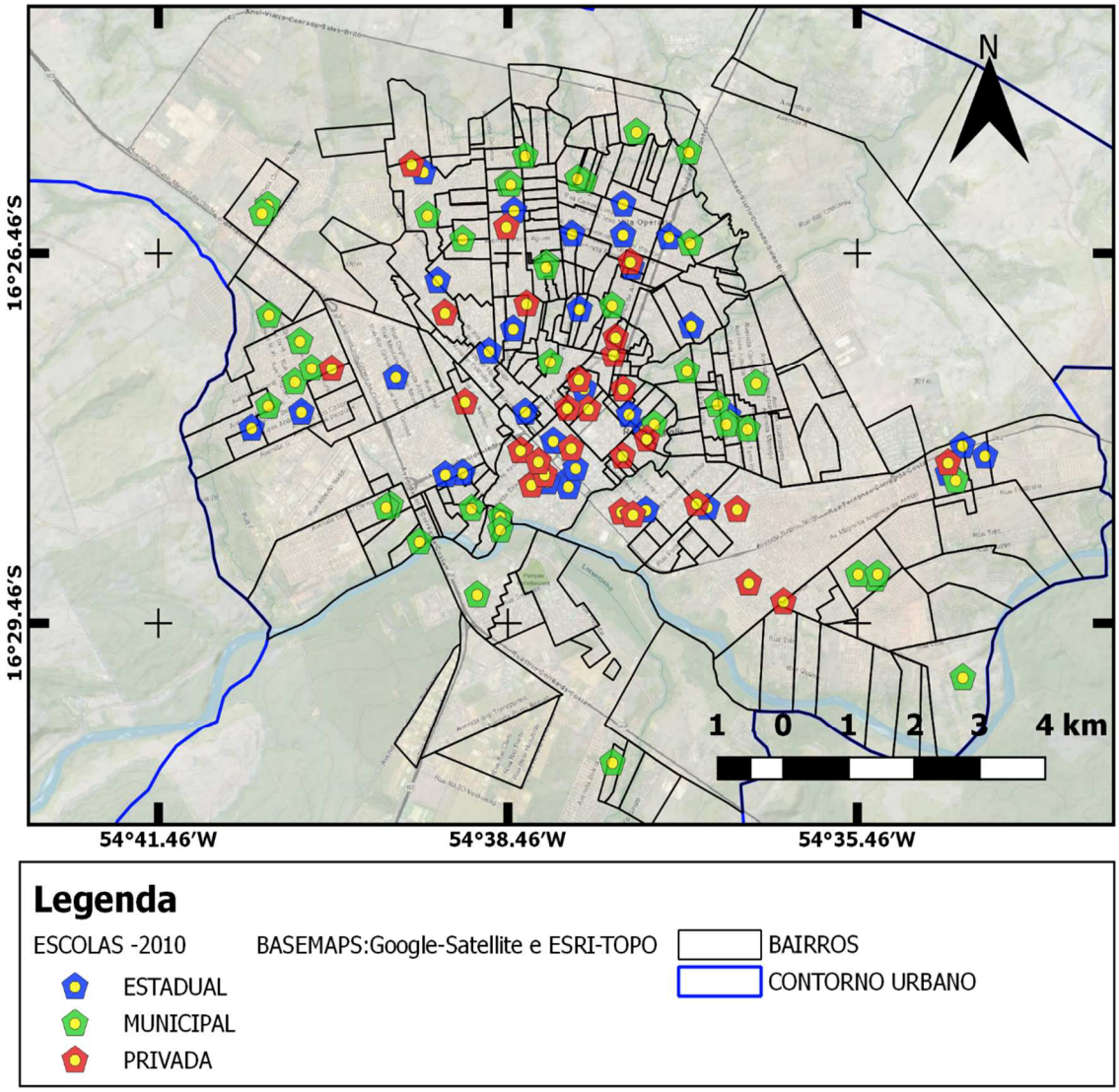
Tabela 1 - Número de escolas nos anos de 2010 e 2017, no Município de Rondonópolis-MT

Dependência Administrativa	2010			2017					
	Urbana	Rural	Total	Urbana			Rural		Total
	Ativa	Ativa	geral	Ativa	Extinta	Paralisada	Ativa	paralisada	Geral
ESTADUAL	32	2	34	34	-	-	2	-	36
FEDERAL	-	-	-	1	-	-	-	-	1
MUNICIPAL	40	12	52	52	-	-	15	1	68
PRIVADA	28	1	29	31	1	8	1	-	41
Total Geral	100	15	115	118	1	8	18	1	146

Fonte: Censo Escolar de 2010 e 2017

Na Figura 3 mostra a geocodificação das Unidades Escolares, na zona urbana do Município de Rondonópolis-MT, no ano de 2010, utilizando o software QGis.

Figura 3 - Mapa da distribuição espacial das Unidades Escolares na Zona Urbana, no Município de Rondonópolis – MT, no ano de 2010

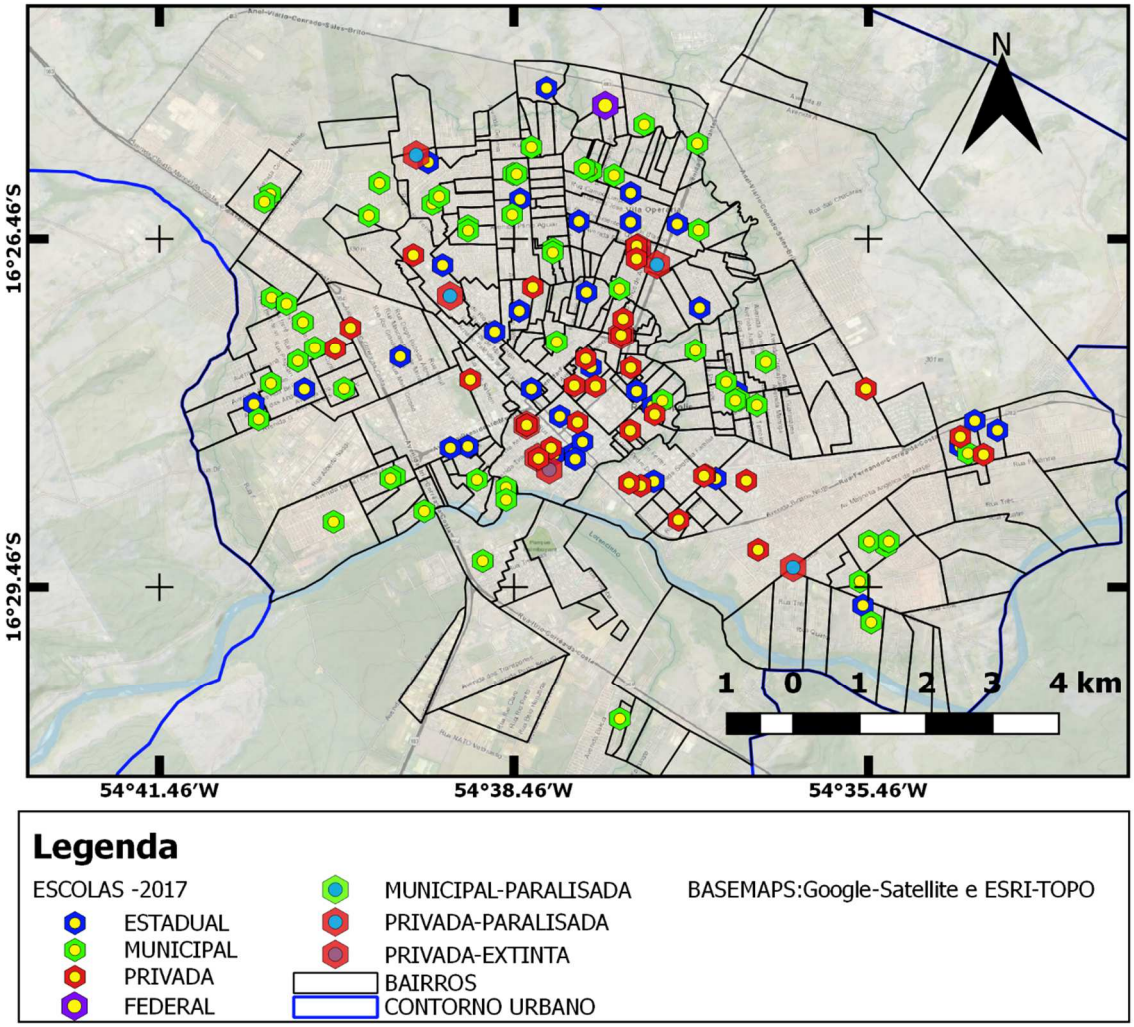


Foi observada que a distribuição espacial das Unidades Escolares na área central de Rondonópolis – MT tem raios de abrangência de atendimento sem vazios, o que é verificado na Figuras 5 e 6, mas que em outras áreas, não apresentaram centros educacionais (Figura 3).

Na Figura 4 mostra a geocodificação das Unidades Escolares, na zona urbana do Município de Rondonópolis-MT, no ano de 2017, utilizando o software QGis.

Verifica-se um aumento de Unidades Escolares, mas também a extinção/paralisação de Unidades Escolares em algumas áreas. Notamos que algumas áreas que não apresentavam centros educacionais em 2010 já começaram a ter suas demandas atendidas em 2017 (Figura 4).

Figura 4 - Mapa da distribuição espacial das Unidades Escolares na Zona Urbana, no Município de Rondonópolis – MT, no ano de 2017



Ao se considerar o raio de abrangência das Unidades Escolares por tipo de atendimento, 800m para Ensino Fundamental-Anos Iniciais e Finais (Figuras 5 e 6) e 2km para o Ensino Médio, (Figura 7), notamos que as Unidades Escolares abrangem as áreas com demandas a serem atendidas.

Figura 5 - Mapa da distribuição espacial das Unidades Escolares, na Zona Urbana, para a etapa de atendimento de Ensino Fundamental – Anos Iniciais com seus raios de abrangência e do mapa de calor da estimativa de população Ensino Fundamental – Anos Iniciais

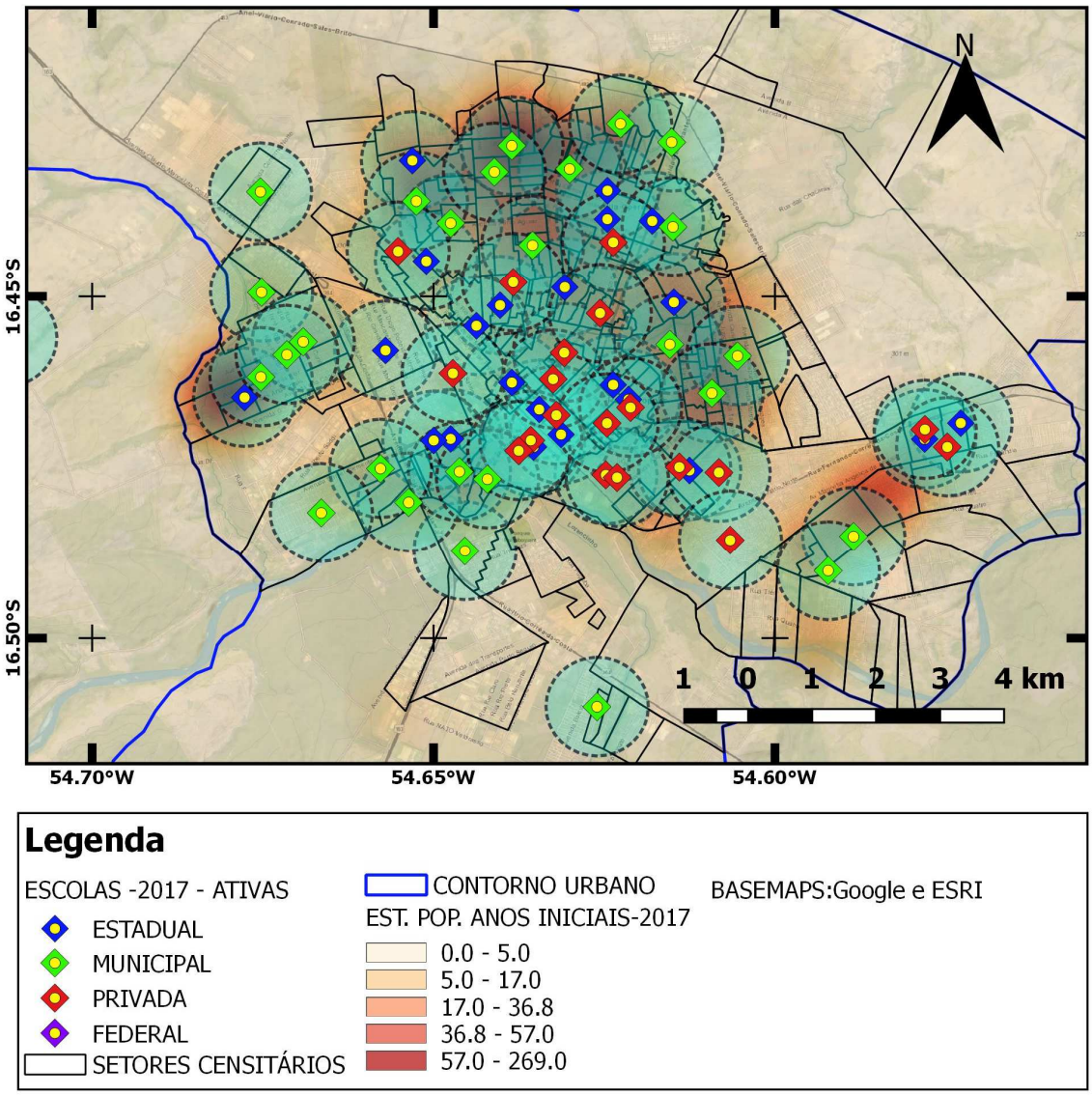
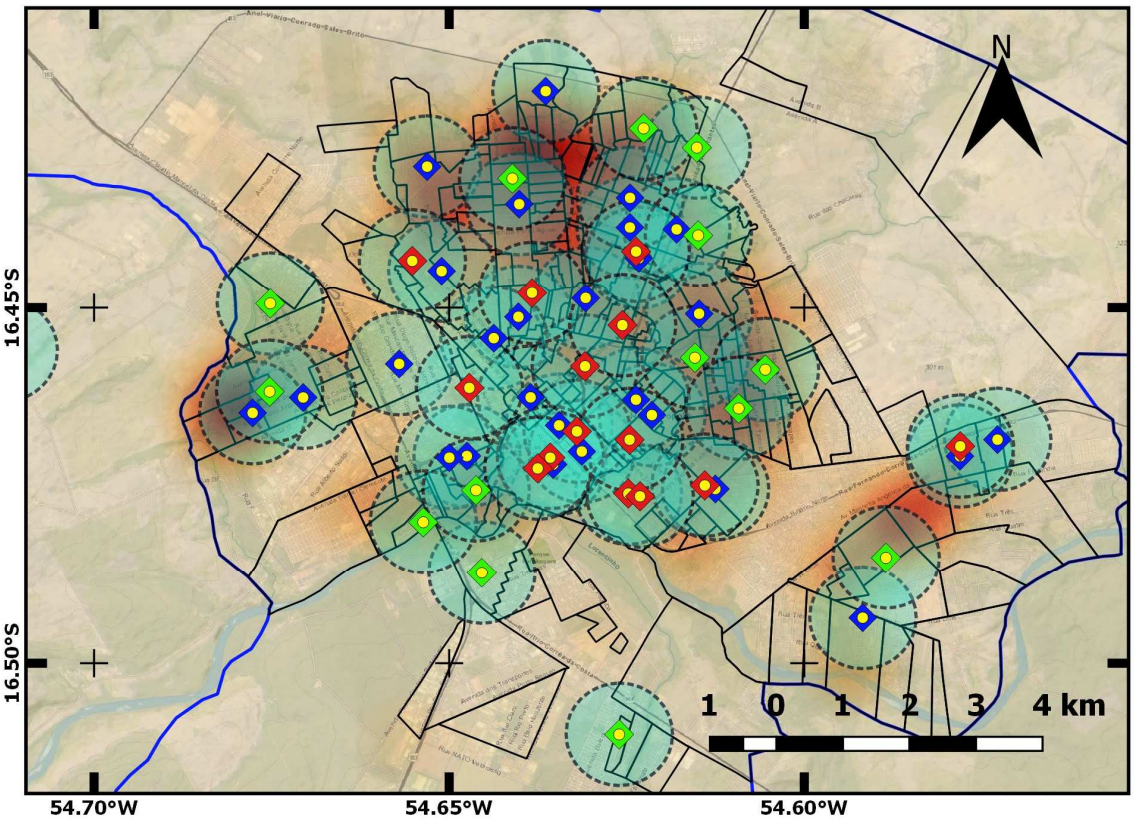


Figura 6 - Mapa da distribuição espacial das Unidades Escolares, na Zona Urbana, para a etapa de atendimento de Ensino Fundamental – Anos Finais e do mapa de calor da estimativa de população de Ensino Fundamental – Anos Finais



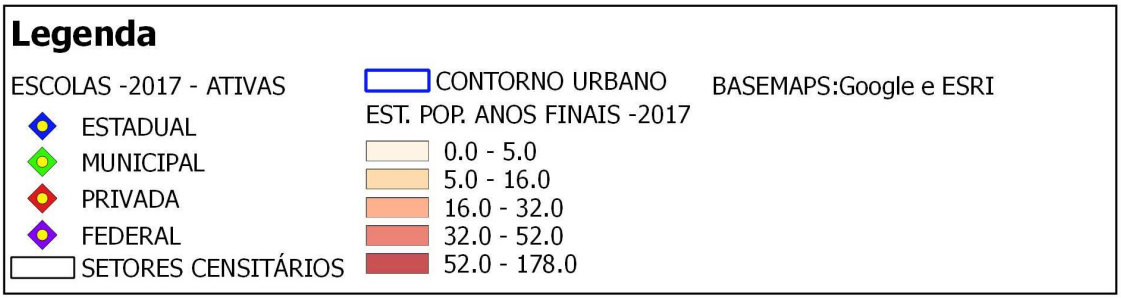
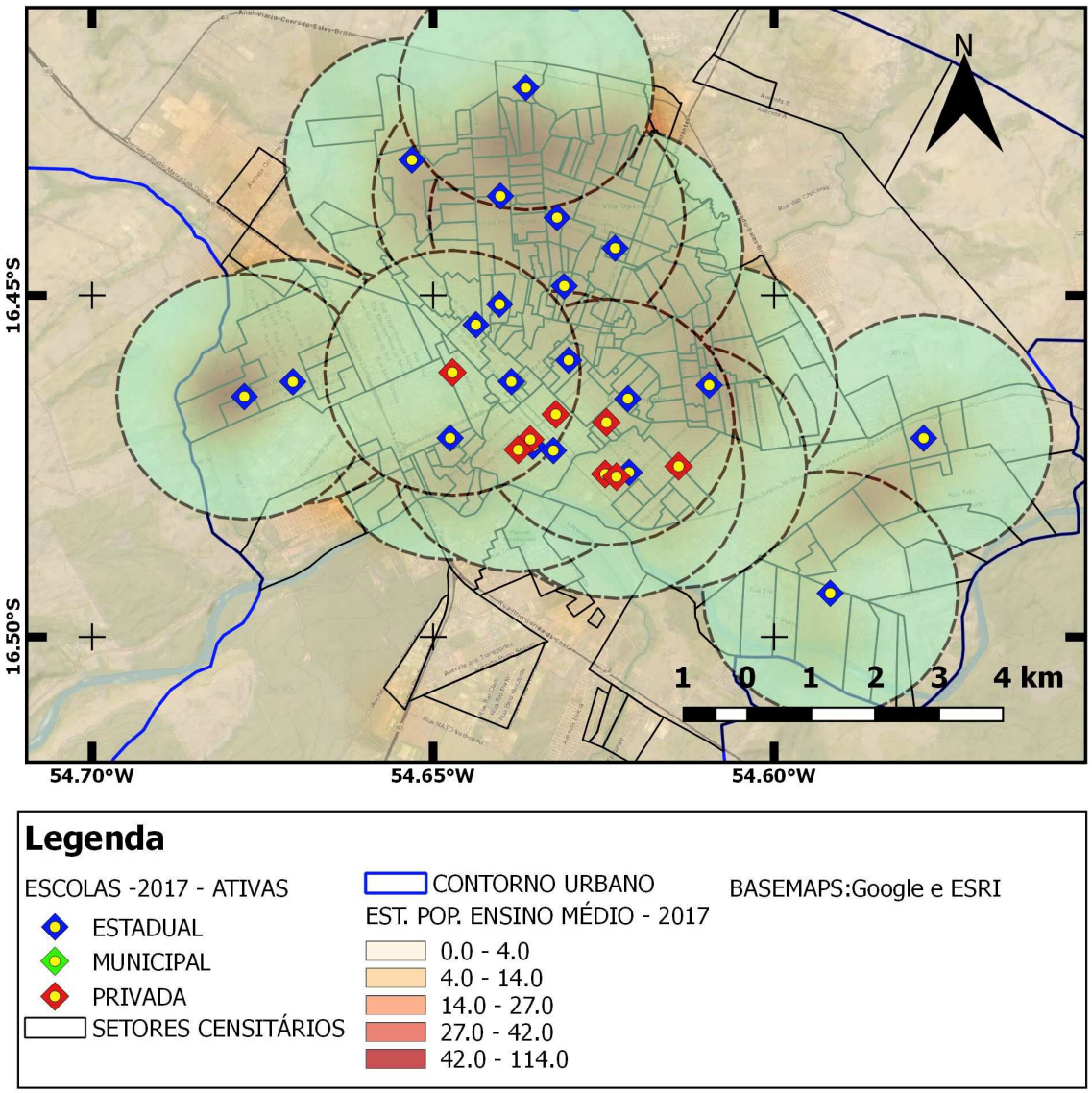


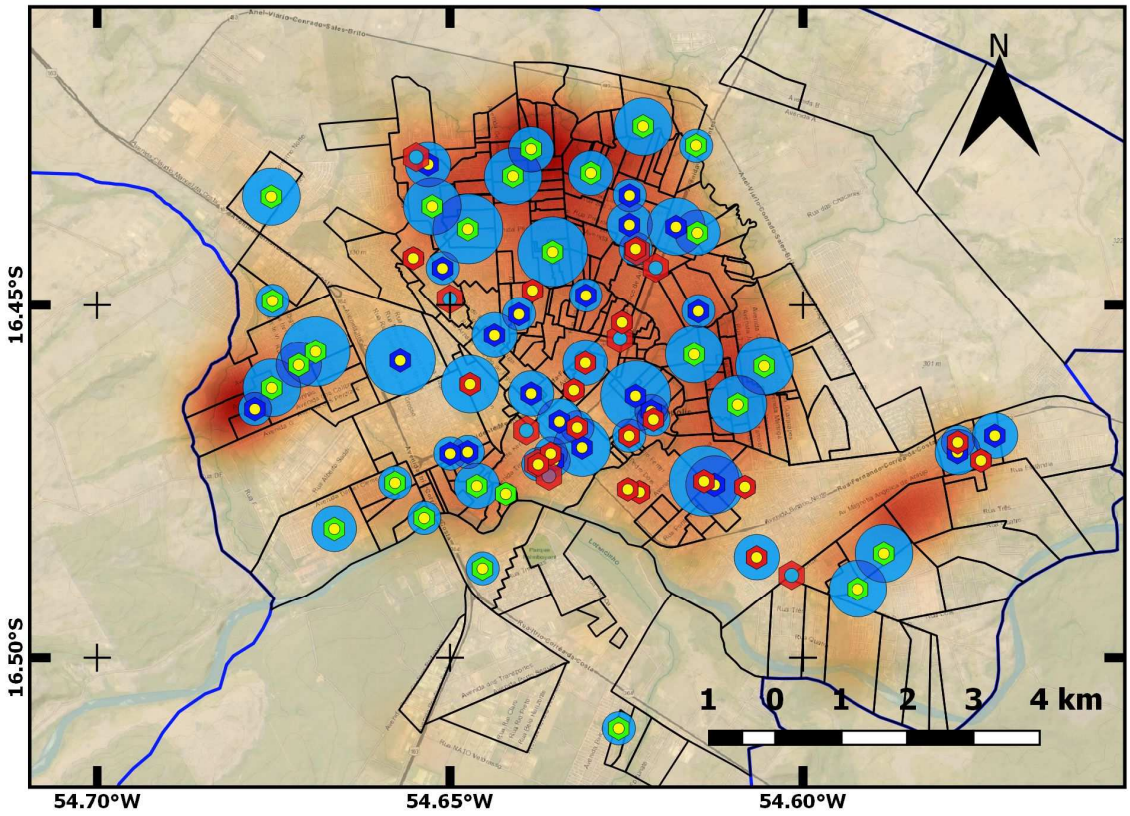
Figura 7 - Mapa da distribuição espacial das Unidades Escolares, na Zona Urbana, para a etapa de atendimento de Ensino Médio e do mapa de calor da estimativa de população para Ensino Médio



A análise dos raios de abrangência das escolas por etapa de atendimento não é suficiente para verificar a demanda de atendimento por etapa de ensino está sendo bem atendida.

Ao comparar o atendimento ofertado por etapa das Unidades Escolares com a estimativa de população por etapa dos setores censitários, para o ano de 2017, verificamos a cobertura da demanda estimada das etapas de ensino, de acordo com o raio de atendimento das escolas, na zona Urbana do Município de Rondonópolis-MT (Figuras 8 a 10).

Figura 8 - Mapa comparativo da estimativa de população de Ensino Fundamental – Anos Iniciais com o atendimento realizado pelas Unidades Escolares



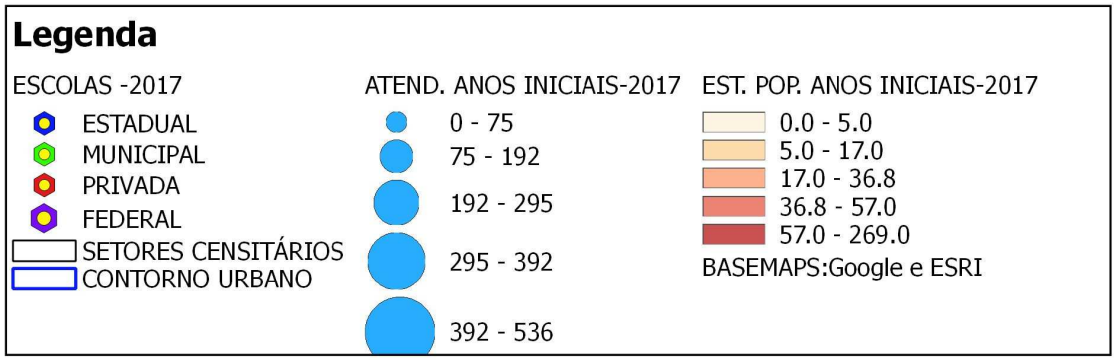


Figura 9 - Mapa comparativo da estimativa de população de Ensino Fundamental – Anos Finais com o atendimento realizado pelas Unidades Escolares

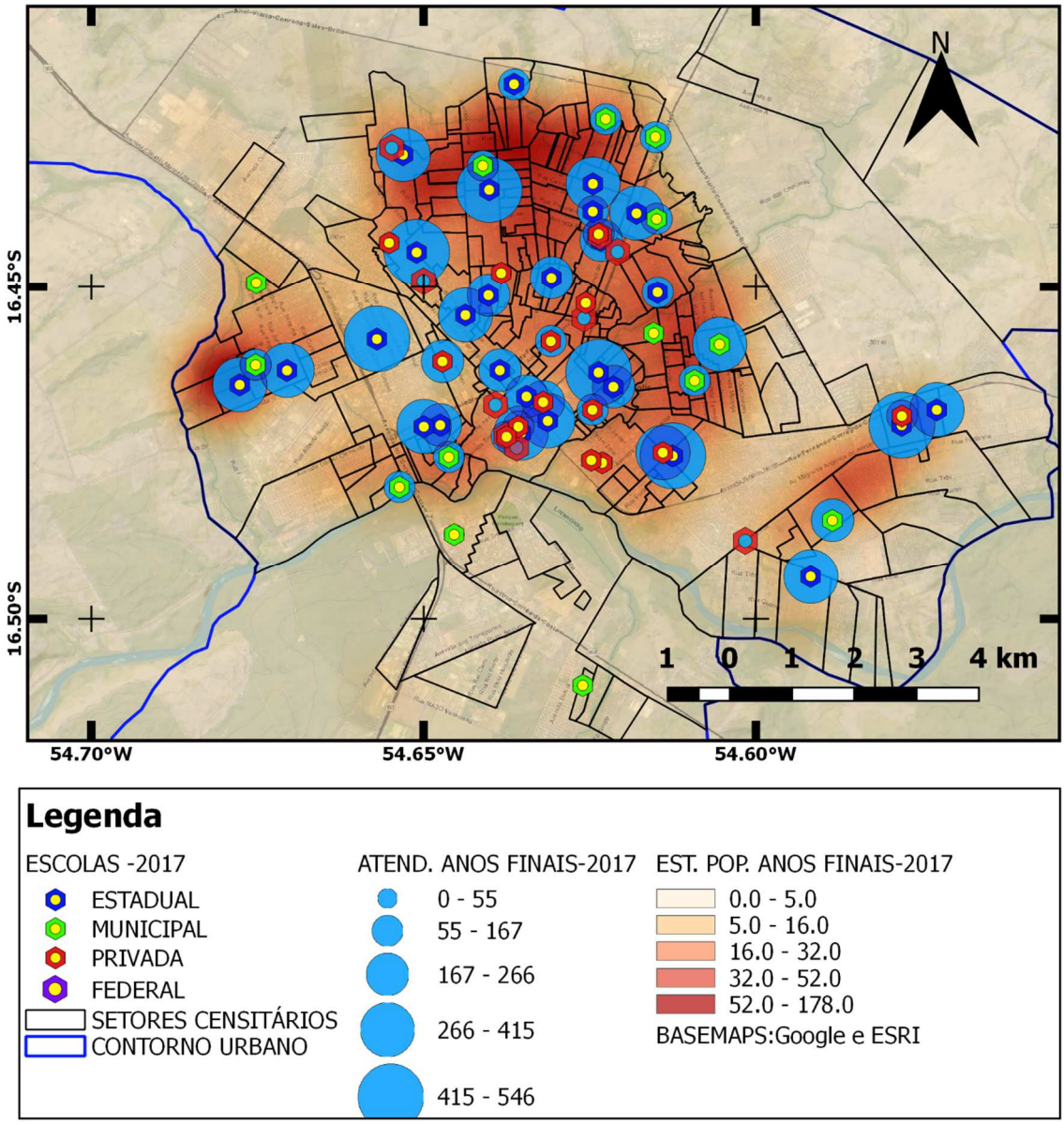
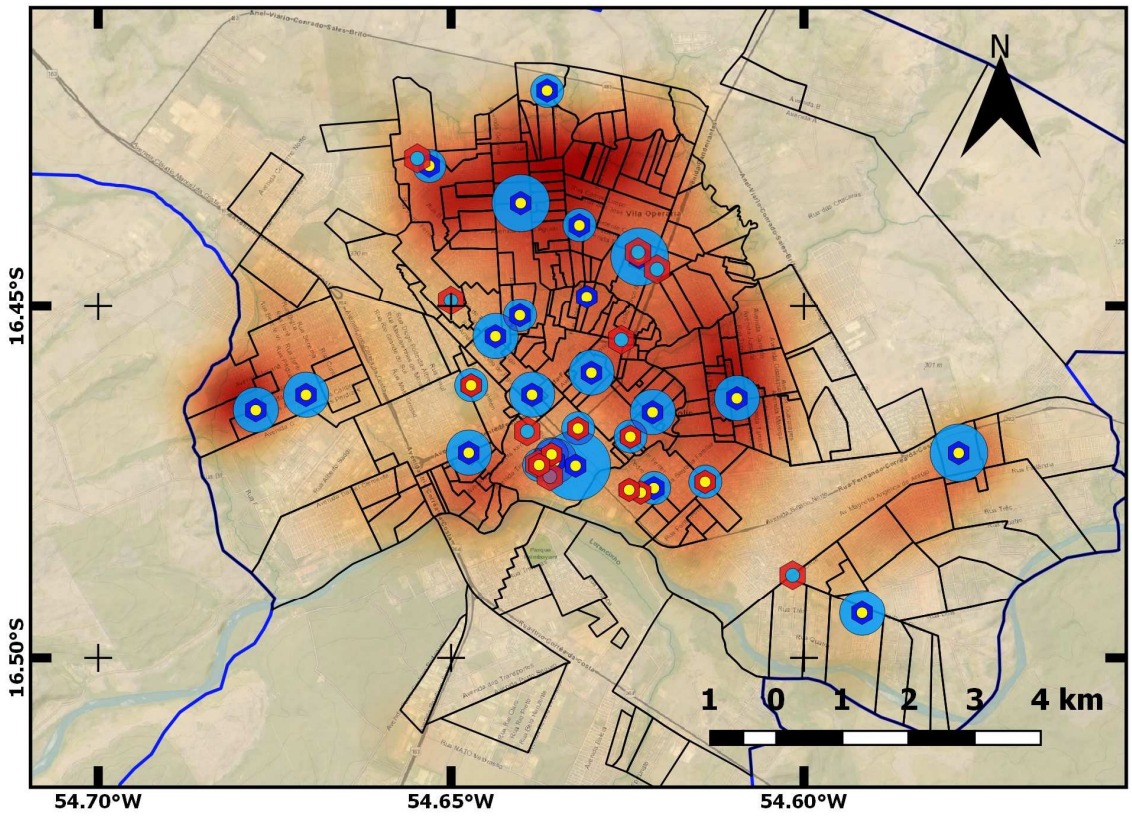
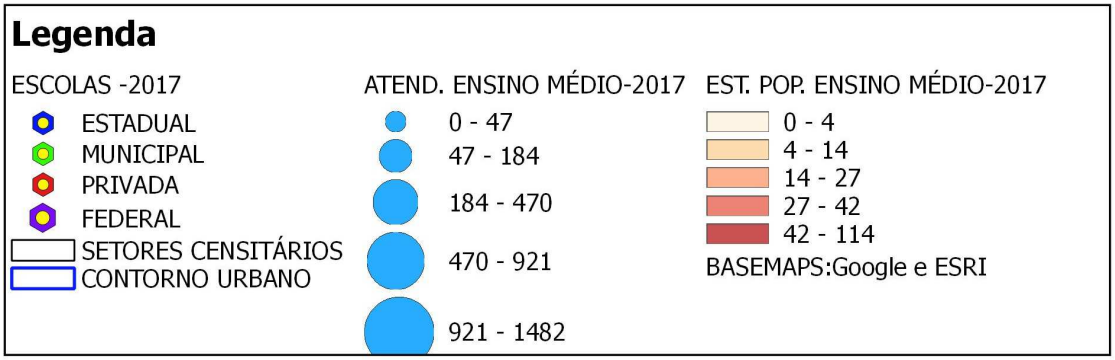


Figura 10 - Mapa comparativo da estimativa de população de Ensino Médio com o atendimento realizado pelas Unidades Escolares





Verifica-se que há um déficit no atendimento nas etapas de atendimento, o que pode ser constatado com a comparação dos dados de estimativa de população com o atendimento escolar por etapa de atendimento, indicado na Tabela 2.

Tabela 2 - Comparativo entre os dados de atendimento escolar e a estimativa de população para 2017, por etapa de atendimento, na zona Urbana do Município de Rondonópolis

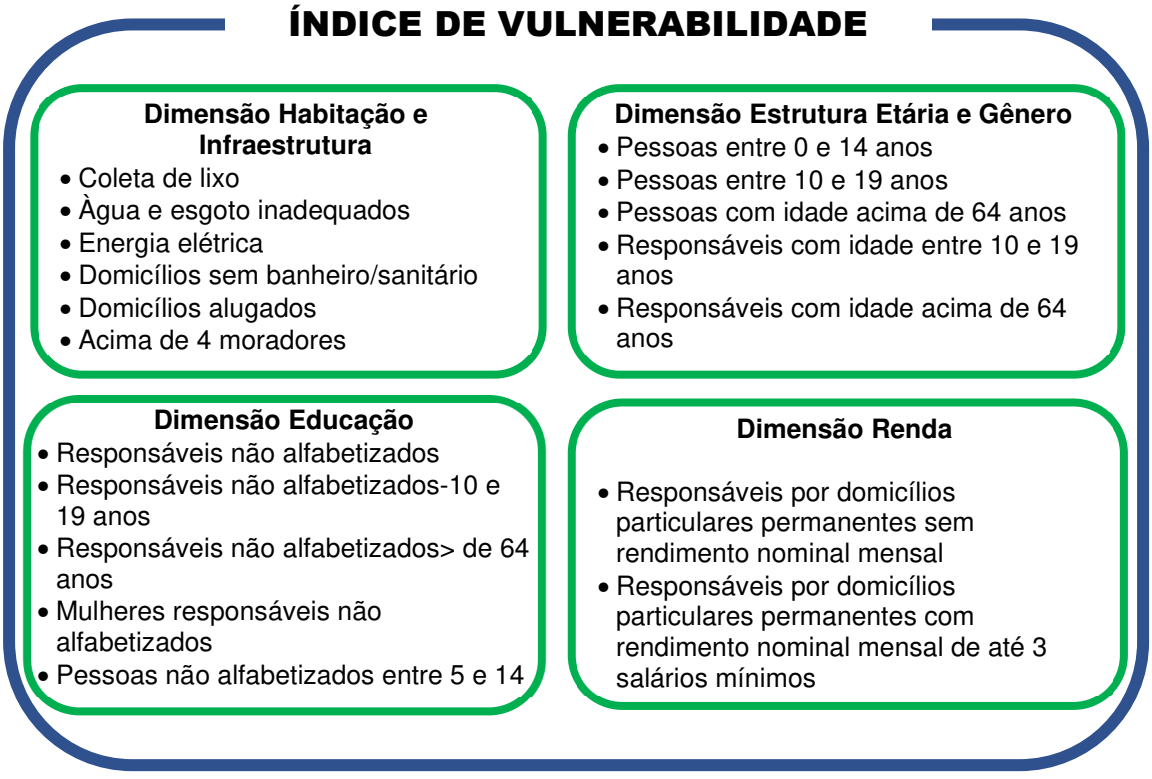
Tipo de Atendimento	CENSO ESCOLAR - 2017				Pop. estimada - 2017	Diferença de Atendimento no Município	Situação
	Município	Estado	Privado	Atendimento 2017			
Ens. Fund. - Anos Iniciais	7534	5195	2994	15723	16521	-798	DÉFICIT
Ens. Fund. - Anos Finais	1663	9039	1518	12220	14494	-2274	DÉFICIT
Ensino Médio	0	7515	704	8219	11690	-3471	DÉFICIT
TOTAL	9197	21749	5216	36162	42705	-6543	

Fonte: Censo Escolar-INEP/Censo Demográfico 2010 e estimativas de população-IBGE

Este déficit de atendimento pode afetar o funcionamento das Unidades Escolares, com a busca de vagas onde estão localizadas, aumentando o atendimento e ocasionando salas superlotadas, o que pode interferir no processo de aprendizagem dos alunos.

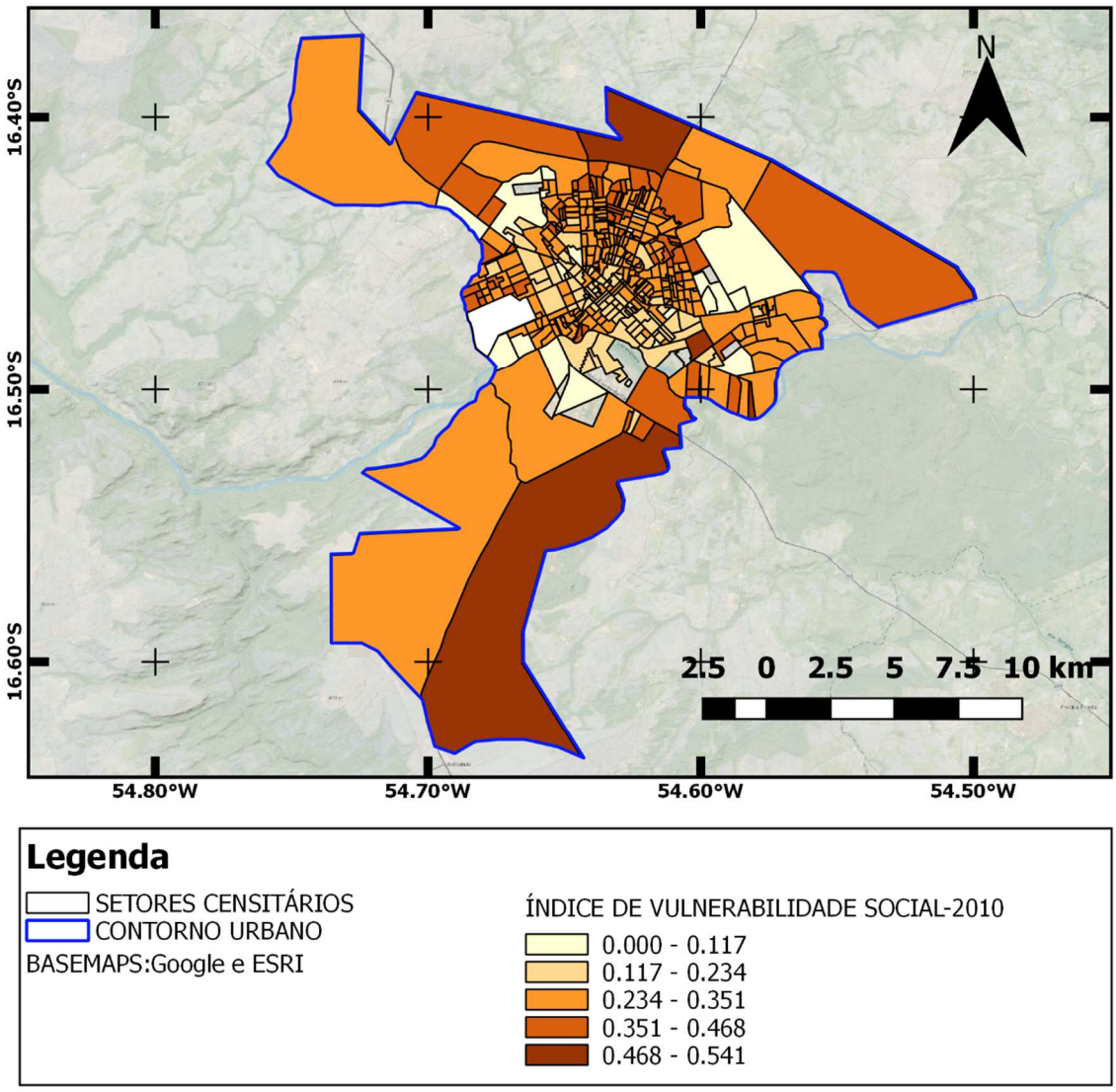
Para verificar se os indicadores educacionais das Unidade Escolares têm correlação com o fenômeno de segregação socioespacial das áreas onde se encontram implantadas as Unidades Escolares, usamos a metodologia de Dias (2013) para calcular o Índice de Vulnerabilidade Social (IVS) dos setores urbanos do Município de Rondonópolis-MT, a partir de 22 variáveis disponibilizadas na da Base de Informações do Censo Demográfico 2010 – IBGE, agrupadas segundo o esquema de blocos da Figura 11.

Figura 11 - Agrupamento das variáveis da Base de Informações do Censo 2010-IBGE para elaboração do Índice de Vulnerabilidade Social (IVS)



Na Figura 12 temos o mapa de vulnerabilidade social para a zona Urbana do Município de Rondonópolis – MT, elaborado com o auxílio do software livre QGis.

Figura 12 - Mapa de Vulnerabilidade Social da área urbana do Município de Rondonópolis-MT



Fonte de dados - Censo Demográfico 2010 – IBGE

Comparando o Índice de Desenvolvimento da Educação Básica, para os anos de 2011 e 2015[1] , das Unidades Escolares, com o Índice de Vulnerabilidade Social para o Município de Rondonópolis-MT (Figuras 13 e 14), verificamos que não há uma correlação aparente, tanto para o IDEB dos Anos Iniciais quanto para os Anos Finais.

[1] O Índice de Desenvolvimento da Educação Básica (IDEB) varia entre 0 e 10 e é calculado pelo Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP)-MEC, sendo divulgado bianualmente, para os Anos Iniciais (1º ao 5º ano) e Finais (6º ao 9º ano) do Ensino Fundamental. Foram coletados os IDEBs disponíveis das Unidades Escolares do município de Rondonópolis-MT.

Figura 13 - Mapa com o IDEB para o Ensino Fundamental – Anos Iniciais, das Unidades Escolares

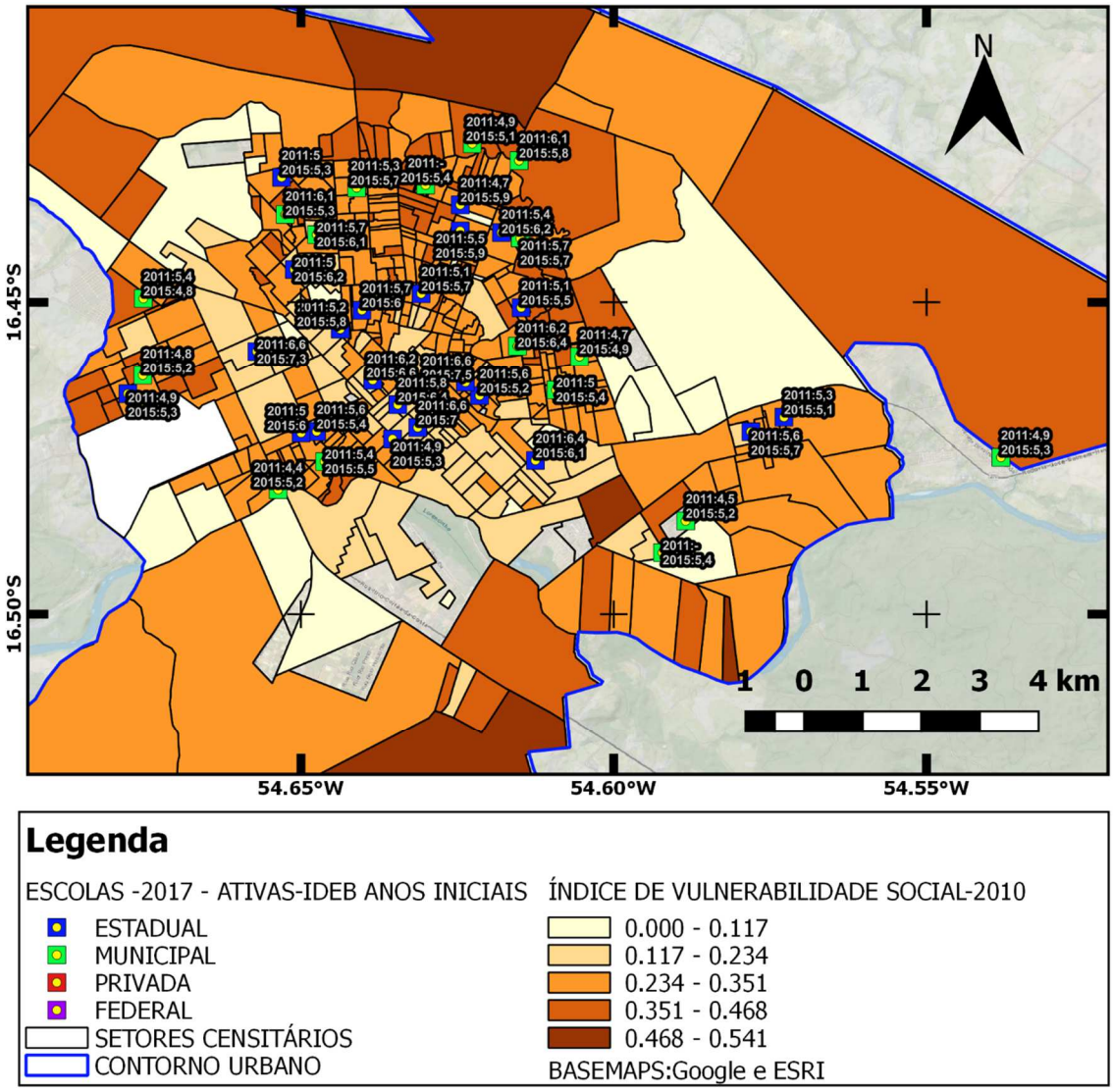
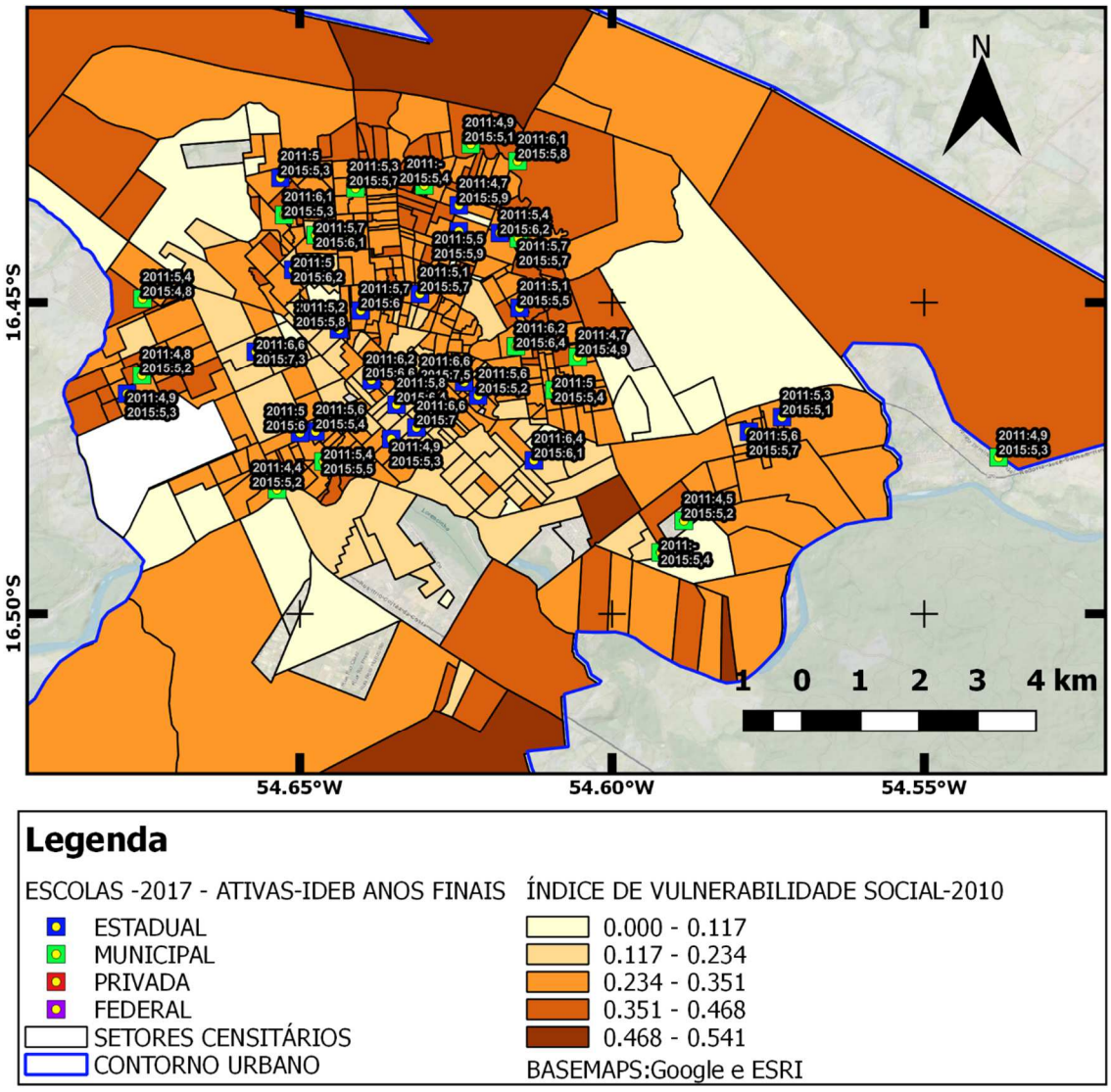


Figura 14 - Mapa com o IDEB para o Ensino Fundamental – Anos Finais, das Unidades Escolares



Usando o EXCEL podemos comparar o Índice de Desenvolvimento da Educação Básica, para os anos de 2011 e 2015[2] , das Unidades Escolares, com o Índice de Vulnerabilidade Social Médio, para o município de Rondonópolis-MT (Figuras 15 e 16), verificamos que há uma correlação fraca, tanto para o IDEB dos Anos Iniciais quanto para os Anos Finais. Para a etapa de atendimento de Ensino Médio não há dados disponíveis de IDEB para esta análise.

[2] O Índice de Desenvolvimento da Educação Básica (IDEB) varia entre 0 e 10 e é calculado pelo Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP)-MEC, sendo divulgado bianualmente, para os Anos Iniciais (1º ao 5º ano) e Finais (6º ao 9º ano) do Ensino Fundamental. Foram coletados os IDEBs disponíveis das Unidades Escolares do município de Rondonópolis-MT.

Figura 15 - IDEB – Anos Iniciais vs. IVS Médio para os anos de 2011 e 2015

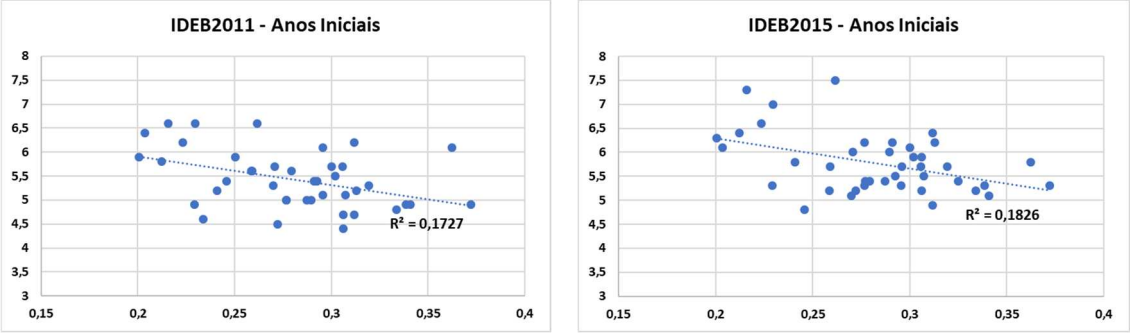
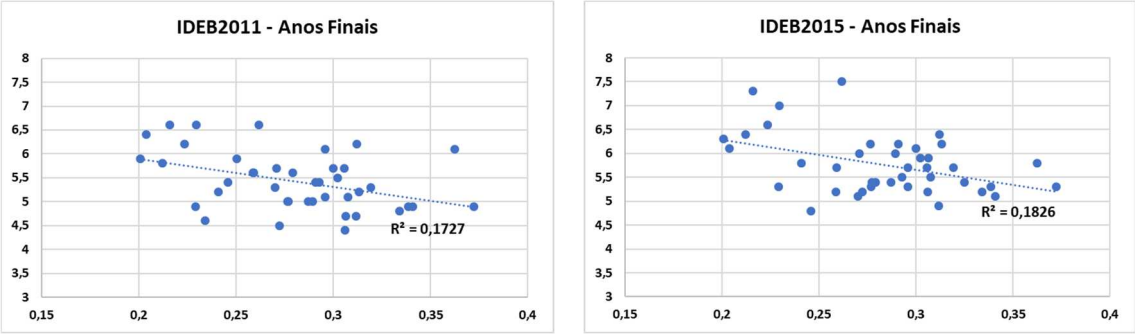


Figura 16 - IDEB – Anos Finais vs. IVS Médio para os anos de 2011 e 2015



4.CONCLUSÃO

A disposição espacial das Unidades Escolares acarreta em diferentes níveis de oportunidades quanto ao acesso às Unidades Escolares, o que vem a corroborar quanto ao papel que o Geoprocessamento ocupa no desenvolvimento regional, Borba & Silva (2012), podendo servir como uma ferramenta de assessoramento técnico às secretarias dos Municípios, oferendo subsídios para tomada de decisões quanto as necessidades de melhor atendimento da rede pública escolar.

Constatou-se que o IDEB das Unidades Escolares, tanto para os Anos Iniciais e Finais do Ensino Fundamental, tem uma correlação fraca aparente com o Índice de Vulnerabilidade Social Médio da área que está inserida as mesmas, para o município de Rondonópolis-MT, e que as metodologias utilizadas associadas com ferramentas de Geoprocessamento torna-se indispensável para verificar o panorama de atendimento escolar, auxiliando as tomadas de decisões sobre as áreas com necessidade de maiores investimentos, seja de recursos financeiros ou pedagógicos.

Como a análise foi feita utilizando dados do Censo Demográfico de 2010, temos que considerar o lapso temporal desses índices. Quanto mais recentes forem os dados sobre a população mais significativo e mais próximo da realidade regional será a análise realizada.

A análise de dados deve ser complementar, usando dados espacializados e ferramentas de análise de dados estatísticos para termos um bom panorama do atendimento escolar no Município de Rondonópolis – MT, podendo a metodologia estendida para os demais Municípios.

5. REFERÊNCIAS BIBLIOGRÁFICAS

JAYARAMAN, V. *Transportation, facility location and inventory issues in distribution network design*.International Journal of Operations and Production Management, Vol. 18, 471-494,1998.

OWEN, S.H. e DASKIN, M.S. *Strategic facility location: A review*. European Journal of Operational Research, Vol. 111, 423-447,1998.

CÂMARA, Gilberto; Davis, Clodoveu e Monteiro, Antônio Miguel. *INTRODUÇÃO À CIÊNCIA DA GEOINFORMAÇÃO*. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2001.

ARANTES, Claudio Oliveira. *Mapeamento Educacional Urbano*. Brasília: MEC , 1991.

MOURA, Ana Clara Mourão. *Geoprocessamento na gestão e planejamento urbano*.3. ed. Rio de Janeiro: Editora Interciência, 2014.

SILVA, LRA et al. *Ferramenta SIG de cálculo de estimativa populacional para o planejamento urbano na cidade do Rio de Janeiro*. Foz do Iguaçú,PR: XVI Simpósio Brasileiro de Sensoriamento Remoto, 2013.

MALLER, Rafaela; GANDOLPHO, André Alves. *LOCALIZAÇÃO DE ESCOLAS DO ENSINO FUNDAMENTAL: CASO DE ITAIPAVA/RJ*. Revista de Engenharia da Universidade Católica de Petrópolis, v. 8, n. 2, 2014.

TORRES, Rubens Petri; NEGRI, Silvio Moises. *ANÁLISE DA SEGREGAÇÃO SÓCIO ESPACIAL URBANA EM RONDONÓPOLIS (MT), A PARTIR DOS EQUIPAMENTOS URBANOS E SOCIAIS INSTALADOS*. VII Congresso Brasileiro de Geógrafos. Vitória-ES. 2014.

DIAS, Gutemberg Henrique. *Identificação da vulnerabilidade socioambiental na área urbana de Mossoró-RN, a partir do uso de técnicas de análises espaciais*. Dissertação de Mestrado em Ciências Naturais. Rio Grande do Norte: UFRN, 2013.

REZENDE, P. S. *Metodologia para avaliação da vulnerabilidade socioambiental: Estudo da cidade de Paracatu(MG).* Dissertação de Mestrado em Geografia. Uberlândia: UFU, 2016.

ÉRNICA, M.; BATISTA, A. A. G. *A escola, a metrópole e a vizinhança vulnerável*. Cadernos de Pesquisa. São Paulo. 2012.

NEGRI, Silvio Moisés et al. *O processo de segregação sócio-espacial no contexto do desenvolvimento econômico da cidade de Rondonópolis-MT*. 2008.

NEVES, Fernando Henrique. *Planejamento de equipamentos urbanos comunitários de educação: algumas reflexões*. Cadernos Metrópole, v. 17, n. 34, 2015.

ISSN 2675-3243