A Beta-Beta prime model for rates and their precision with application to small area estimation

Fernando A. S. Moura (IM-UFRJ)

Soraia Perreira (Faculdade de Ciências, Universidade de Lisboa)

Giovani Loiola da Silva (Instituto Superior Técnico da Universidade de Lisboa)

DME – UFRJ 2023, Ciclo de Palestras da Pós-graduação em Estatística do DME-UFRJ

Summary

Resumo

- Agências nacionais de estatística do mundo inteiro necessitam fornecer estimativas confiáveis de índices econômicos e sociais, ao nível de pequenas áreas ou pequenos domínios, a partir de dados de pesquisas amostrais.

- No entanto, devido ao pequeno tamanho da amostra nessas áreas, não é viável obter estimativas com um nível de precisão aceitável sem usar abordagens baseadas em modelos.

- Este trabalho propõe modelar conjuntamente o estimador direto de índices no intervalo (0,1) e suas respectivas precisões utilizando-se as distribuições Beta e Beta prime.

Resumo

- Apresenta-se um estudo de avaliação da metodologia proposta com dados reais.

- Uma aplicação aos dados da Pesquisa Nacional de Orçamentos Familiares (POF-2018) para se estimar o índice de insegurança alimentar para pequenas áreas do Estado de Minas Gerais também é apresentada.

Exemplo: Pesquisa Domiciliar

Objetivo: estimar a proporção de domicílios com uma determinada característica para cada domínio (pequena área) $i = 1, \ldots, I$.

- Estimativas com nível de precisão aceitável para os Estados.
- Estimativas imprecisas para áreas menores (municípios).
- Primeiro estágio: Setores censitários.
- Segundo estágio: Domicílios

Exemplo: Pesquisa Domiciliar

Parâmetros de interesse

$$\mu_i = \frac{\sum_{j \in S_i} \sum_{k=1}^{N_{ij}} y_{ijk}}{\sum_{j \in S_i} N_{ij}} \qquad i = 1, \ldots, I \tag{1}$$

Onde:

S_i conjunto de setores censitários pertencentes ao município i;

j denota a unidade primária de amostragem(setor censitário);

k denota o domicílio pertencente ao setor j do município i;

y_{ijk} é o indicador binário de presença da característica de interesse para o domicílio k pertencente ao setor j do município i;

N_{ij} é o número de domicílios no setor j pertencente ao município i.

Exemplo: Pesquisa Domiciliar

Estimativa direta de μ_i

$$r_i = \frac{\sum_{j \in s_i} \sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk}}{\sum_{j \in s_i} \sum_{k=1}^{n_{ij}} w_{ijk}} \tag{2}$$

Onde:

s_i é o conjunto de setores selecionados que pertencem ao município i;

n_{ij} é o número de domicílios selecionados no setor selecionado $j \in s_i$;

w_{ijk} é o peso amostral do domicílio selecionado k pertencente ao setor selecionado $j \in s_i$;

Exemplo: Pesquisa Domiciliar

Estimativa da Variância de R_i

Assumindo-se as fração de amostragem de primeiro estágio desprezível ($f_i = m_i/M_i \approx 0$)

$$\hat{V}_D(R_i) = \frac{m_i}{(m_i - 1)\left(\sum_{j \in s_i} \sum_{k=1}^{n_{ij}} w_{ijk}\right)^2} \sum_{j \in s_i} \left(\sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk} - r_i \sum_{k=1}^{n_{ij}} w_{ijk} \right)^2 \tag{3}$$

Onde: m_i é o número de setores selecionados no município i.

- $\hat{V}_D(R_i)$ pode ter um vício não desprezível, principamente quando m_i for pequeno.
- $V_D(R_i)$ pode ser estimado por Jackknife ou outros métodos de reamostragem.
- A vantagem de se utilizar métodos de reamostragem é a redução substancial do vício.

Exemplo: Pesquisa Domiciliar

Por que não modelar y_{ijk} e N_{ijk} ao invés de R_i (e) $\hat{V}_D(R_i)$?

- Como μ_i é uma função dos y_{ijk}'s e dos N_{ijk}'s poderíamos pensar num modelo desagregado para prevê-los e então prever os μ_i's.

- Contudo, há duas razões principais para não se utilizarem modelos desagregados:

  - Confidencialidade dos dados.

  - Geralmente não há possibilidade de se identificarem as unidades selecionadas no cadastro (ou registros administrativos) que contem as covariáveis para auxiliarem na predição.

Modelo Beta Hierárquico de Precisão Constante (BH)

Notação

- Seja $0 < \mu_i < 1$ o "verdadeiro valor" desconhecido de uma taxa ou proporção para a pequena área $i = 1, \ldots, m$.
- Seja $0 < R_i < 1$ o estimador empregado de μ_i para cada pequena área $i = 1, \ldots, m$.
- Como a pesquisa foi desenhada para fornecer estimativas precisas para áreas maiores, as estimativas r_i são (para a maioria das áreas) imprecisas.
- Dependendo do desenho amostral empregado, para algumas áreas podem não haver nenhuma informação amostral.
- Mesmo quando a informação amostral for inexistente para uma área i, há ainda a possibilidade de se fazer inferência sobre μ_i.

Modelo Beta Hierárquico de Precisão Constante

Modelo hierárquico parametrizado como em Ferrari and Cribari(2004)

$$R_i|\mu_i, \phi_i \sim Beta(\mu_i, \phi) \tag{4}$$

- Com $E_M(R_i|\mu_i) = \mu_i$ e $V_M(R_i|\mu_i, \phi) = \mu_i(1-\mu_i)(1+\phi)^{-1}$. onde o símbolo M na esperança e variância é para denotar que elas são calculadas sob o modelo assumido.
- O parâmetro de precisão ϕ é positivo.
- Para "trocar informações entre as áreas" pode-se assumir a seguinte função de ligação:$log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T\beta + \nu_i$, onde $\nu_i|\sigma_\nu^2 \sim N(0, \sigma_\nu^2)$ e i.i.d.
- Note que a hipótese $E_M(R_i|\mu_i, \phi) = \mu_i$ implica que o "estimador direto" é não viciado sob o modelo assumido.

Modelo Beta Hierárquico de Benmei et. al. (2006)

- No contexto de estimação de proporções, Benmei et. al. (2006) propuseram $R_i|\mu_i, \phi_i \sim Beta(\mu_i, \phi_i)$ com $\phi_i = n_i.\gamma_i^{-1} - 1$, onde n_i e γ_i são respectivamente o tamanho da amostra da pequena área i e o seu efeito de desenho (geralmente estimado para o plano amostral empregado).
- Logo:

$$V_M(R_i|\mu_i, \gamma_i) = (\mu_i(1-\mu_i)/n_i)\gamma_i \tag{5}$$

- O termo acima entre parenteses é a variância da proporção amostral em cada área obtida quando é selecionada uma amostra aleatória simples com fração amostral desprezível.
- O parâmetro γ_i reflete o efeito na variância em se utilizar um desenho complexo (ex: amostragem de conglomerados).
- Seja $\kappa_i = \gamma_i/n_i$. Note que a condição $\phi_i > 0$, $\forall i$ implica que $0 < \kappa_i < 1$, $\forall i$.

Estimador de ϕ_i

- Um estimador consistente para ϕ_i é dado por:

$$\hat{\phi}_i = \frac{R_i(1 - R_i)}{\hat{V}_D(R_i|\mu_i, \phi_i)} \tag{6}$$

- Contudo sabe-se que o estimador usual $\hat{V}_D(R_i|\mu_i, \phi_i)$ pode ter um vício não desprezível, principalmente para amostras pequenas.

- Uma alternativa é usar métodos de reamostragem, tal como Jackknife ou bootstrap.

- No estudo de simulação apresentado mais adiante, é comparado o método Jackknife com o estimador usual da variância.

Distribuição Beta-Prime

- Denotemos genericamente por $Y|\mu, v \sim BetaP(\mu, v)$ uma Distribuição Beta-prime com média μ e precisão v cuja sua função de densidade é dada por:

$$f(y|\mu, v) \propto y^{\mu(1+v)-1}(1+y)^{-\mu(1+v)-(2+v)} \qquad y > 0; \quad \mu > 0; \quad v > 0.$$

- Se uma variável aleatória X tem uma Distribuição Beta, então $Y = X^{-1} - 1$ tem Distribuição Beta-prime.

- Tem-se que $V(Y|\mu, v) = \frac{\mu(1+\mu)}{v}$.

Distribuição Beta-Prime para $\hat{\phi}_i$

- Como $0 < \kappa_i < 1$, $\forall i = 1, \ldots, m$, um estimador adequado de κ_i também deve estar no intervalo $(0, 1)$.
- Assume-se, então, que os $\hat{\kappa}_i$'s seguem uma distribuição Beta.
- Logo se $\hat{\phi}_i = \hat{\kappa}_i^{-1} - 1$ é um estimador de ϕ_i, vimos que $\hat{\phi}_i$ tem distribuição Beta-prime.
- Assume-se então que $\hat{\phi}_i | \phi_i, a_\phi \sim BetaP(\phi_i, n_i a_\phi)$, $\quad i = 1, \ldots, m$ e condicionalmente independentes, onde $a_\phi > 0$ é um parâmetro desconhecido.

Modelo Beta Beta-Prime Hierárquico (BBP)

Tem-se como input do modelo o vetor bidimensional $D = (r_i, \hat{\phi}_i)$

$$\begin{aligned} R_i|\mu_i, \phi_i &\sim Beta(\mu_i, \phi_i) \\ \hat{\phi}_i|\phi_i, a_\phi &\sim BetaP(\phi_i, n_i a_\phi). \end{aligned} \tag{7}$$

- Função de ligação

$$log\frac{\mu_i}{1-\mu_i} = x_i^T \boldsymbol{\beta}_i \tag{8}$$

onde x_i^T é um vetor de p covariáveis, possivelmente incluindo o intercepto.

- Priori hierárquica

$$\boldsymbol{\beta}_i|\boldsymbol{\beta}, \Omega \sim N(\boldsymbol{\beta}, \Omega_\beta).$$

Modelo Beta Beta-Prime Hierárquico

- Para "trocar informações" entre os parâmetros $\phi_i > 0$; $i = 1, \ldots, m$, considera-se o seguinte modelo hierárquico Beta-prime:

$$\phi_i \sim BetaP(\mu_\phi, v), \quad \forall i = 1, \ldots, m. \tag{9}$$

- Um caso particular importante é conhecido como "Modelo do intercepto aleatório", onde somente o intercepto do parâmetro da função de ligação em (7) é hierarquicamente estruturado.

- Finalmente, prioris próprias e relativamente vagas são assumidas para os hiperparâmetros: $\boldsymbol{\beta} \sim N(0, 10^{-3} I_p)$; $\Omega_\beta^{-1} \sim Wishart(p, I_p)$; $\mu_\phi \sim Ga(0.1, 0.1)$; $a_\phi \sim Ga(0.1, 0.1)$ e $v \sim Ga(0.1, 0.1)$.

Modelo Beta Beta-prime Hierárquico

Várias generalizações do modelo apresentado acima são possíveis:

- Prioris espacialmente estruturadas.

- Covariáveis, $z_i = (z_{i1}, \ldots, z_{iq})^T$ para ajudarem a explicar os parâmetros ϕ_i's:

$$\begin{aligned} \phi_i &\sim BetaP(\mu_{\phi,i}, v), \quad \forall i = 1, \ldots, m \\ \mu_{\phi,i} &\sim LogNormal(z_i^T \boldsymbol{\eta}_i, v_\mu) \\ \boldsymbol{\eta}_i &\sim N(\boldsymbol{\eta}, \Omega_\eta) \end{aligned}$$

Modelo Beta Hierárquico Empírico(BEH)

Com o objetivo de se avaliar a importância de se modelar o estimador do parâmetro de precisão, $\hat{\phi}_i$, considerou-se a seguinte "versão empírica" do Modelo Hierárquico Beta-Beta prime:

$$\begin{aligned} R_i|\mu_i, \hat{\phi}_i &\sim Beta(\mu_i, \hat{\phi}_i) \quad i = 1, \ldots, m \\ log\frac{\mu_i}{1-\mu_i} &= x_i^T \boldsymbol{\beta}_i \\ \boldsymbol{\beta}_i|\boldsymbol{\beta}, \Omega &\sim N(\boldsymbol{\beta}, \Omega_\beta) \end{aligned} \tag{10}$$

- É importante notar que nesta "versão empírica"do modelo proposto, o parâmetro de precisão ϕ_i é substituído por uma estimativa deste, enquanto que na sua "versão bayesiana completa" toda incerteza devida a estimação de ϕ_i é levada em conta pelo modelo.

Estudo de Simulação com dados reais

- Foi realizado um estudo de simulação, utilizando-se os dados do Censo Experimental de Limeira.
- A população (universo) usada consiste de 38740 domicílios distribuídos em 140 setores censitários.
- Os 140 setores censitários foram considerados as pequenas áreas.
- O número de domicílios na população varia de 57 a 588.
- O parâmetro de interesse, μ_i, $i = 1, \ldots, 140$ é a razão entre o total das rendas dos chefes dos domícilios para cada setor i, cujo os chefes têm no máximo quatro anos de escolaridade formal e o total das rendas de todos os chefes de domicílios deste mesmo setor.

Estudo de Simulação com dados reais

- Para cada um dos 140 setores foram selecionados aleatoriamente e sem reposição $n_i = max(10, 0.10N_i)$ domicílios, onde N_i é o número de domicílios no setor $i = 1, \ldots, 140$.

- Todo o processo foi repetido 500 vezes.

- Para cada um dos 500 conjuntos de amostras selecionadas, calcularam-se a estimativa r_i e as duas estimativas das suas variâncias:
  - Método usual, denotada por $\hat{V}_D^1(R_i|\mu_i, \phi)$
  - Método de reamostragem por Jacknife, denotada por $\hat{V}_D^2(R_i|\mu_i, \phi)$.

- As estimativas de ϕ_i foram obtidas então pela equação (6).

Estudo de Simulação com dados reais

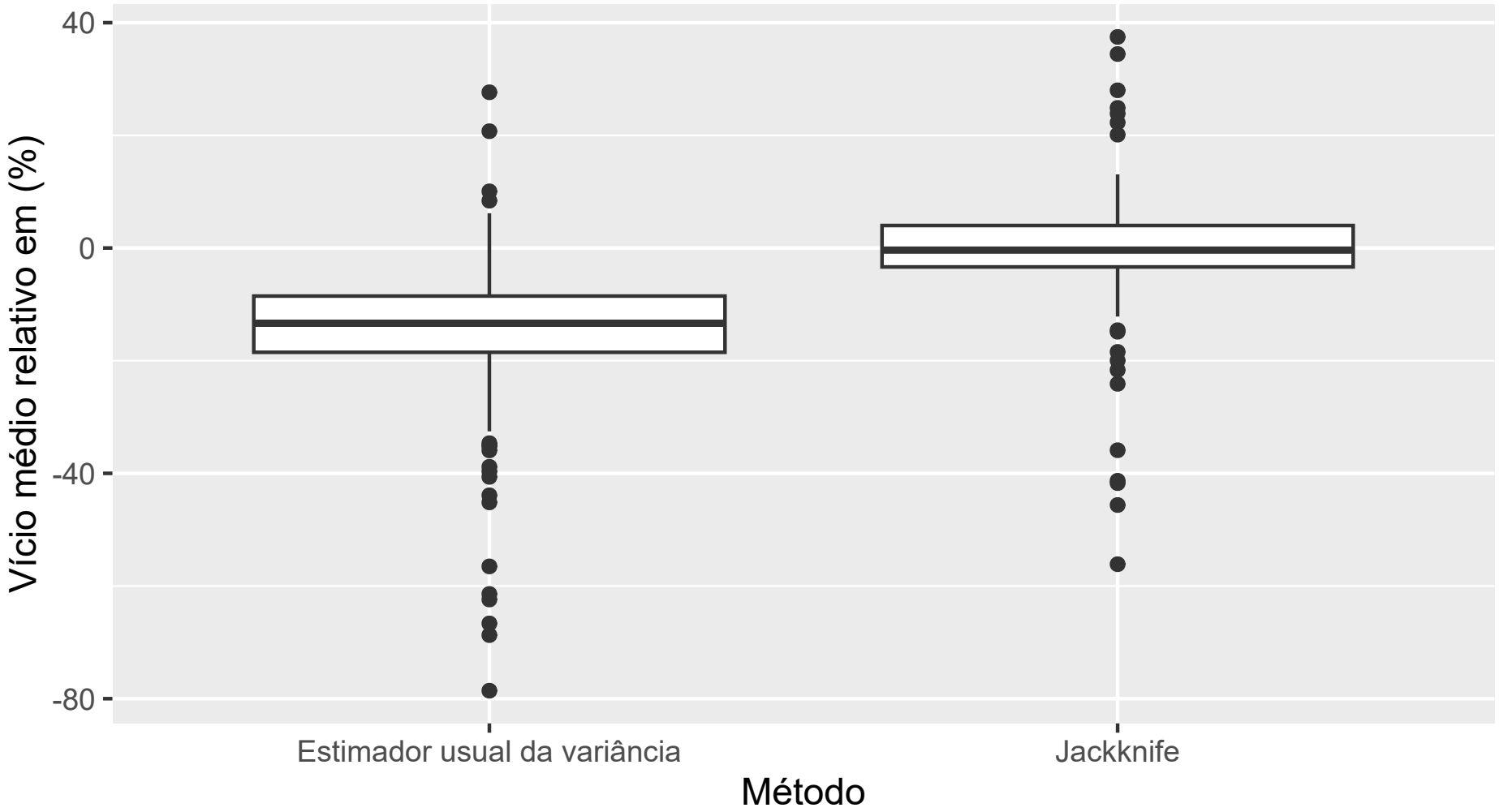

Figure : Boxplots dos vícios médios relativos por setor do estimador usual da variância e do obtido pelo método Jackknife

Estudo de Simulação com dados reais

- Como era esperado, o método de reamostragem de Jacknife apresenta menor vício relativo do que o estimador usual da variância.

- Portanto, foi empregado o método de Jacknife para estimar a variância e consequentemente os parâmetros ϕ_i's dos modelos usados.

Estudo de Simulação com dados reais

- Os três modelos Beta sem inclusão de covariáveis foram ajustados a cada uma das 500 conjuntos de amostras selecionadas.
- O pacote 'Rstan' do R foi usado para selecionar as amostras e ajustar todos os modelos.
- MCMC algorítmo foi usado para gerar uma cadeia de tamanho $200,000$, descartando as primeiras $50,000$ observações e guardando uma observação para cada cinco geradas.
- Portanto, $30,000$ amostras da posteriori de cada parâmetro para cada modelo ajustado foram usadas para inferência.
- As respectivas médias a posteriori calculadas para cada um dos modelos e para cada um dos 140 setores foram consideradas como sendo as estimativas pontuais dos parametros de interesse μ_i, $i = 1, \ldots, 140$.
- Como, para este estudo de simulação, a população é conhecida, foi possível calcular μ_i para cada setor e comparar com as respectivas estimativas obtidas para cada um dos três modelos ajustados.

Estudo de Simulação com dados reais

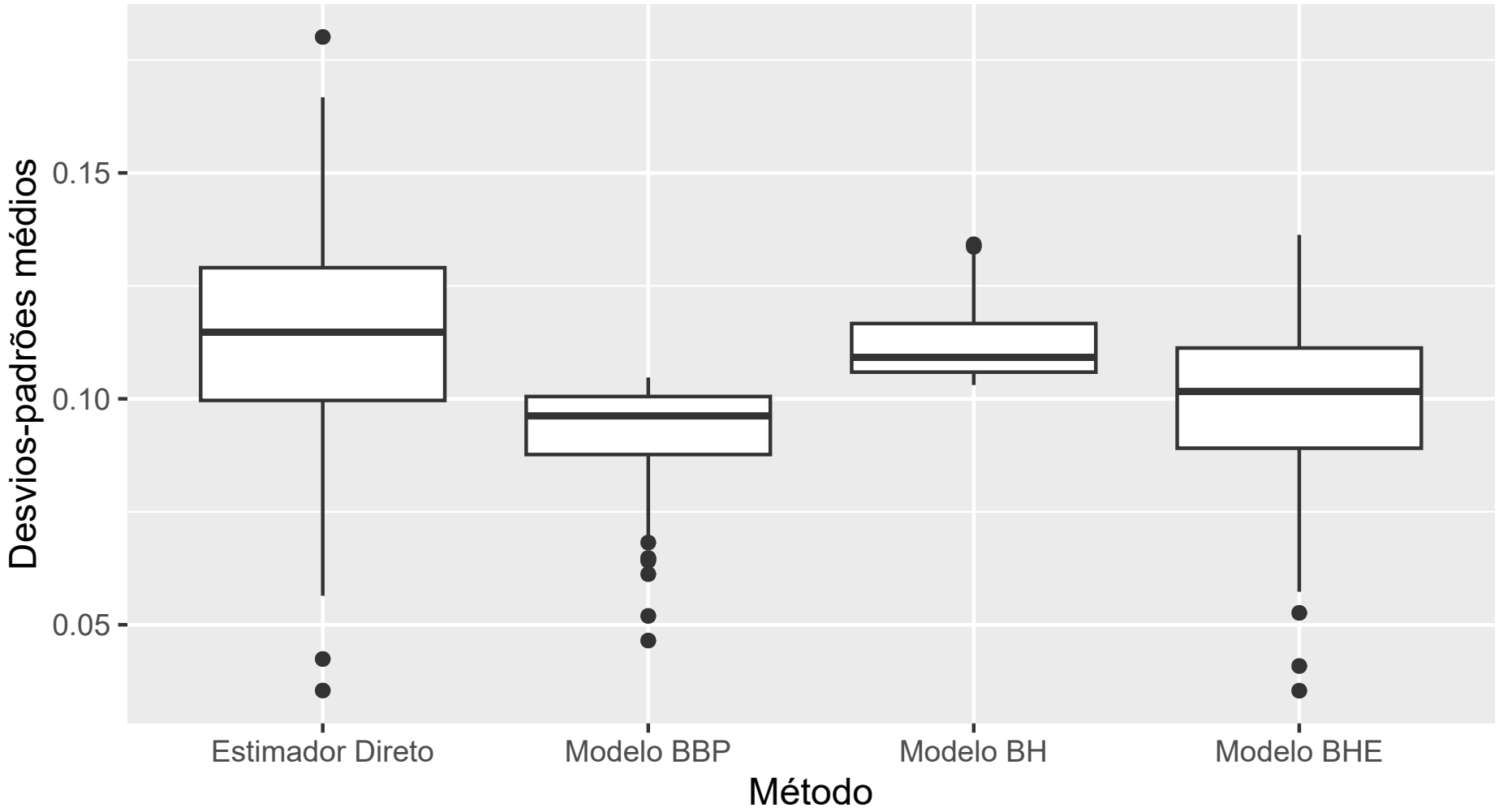

Figure : Boxplot do desvio padrão médio por setor do estimador direto R_i versus os desvios-padrões a posteriori médios por setor para os modelos BH, BHE, BBP

Estudo de Simulação com dados reais

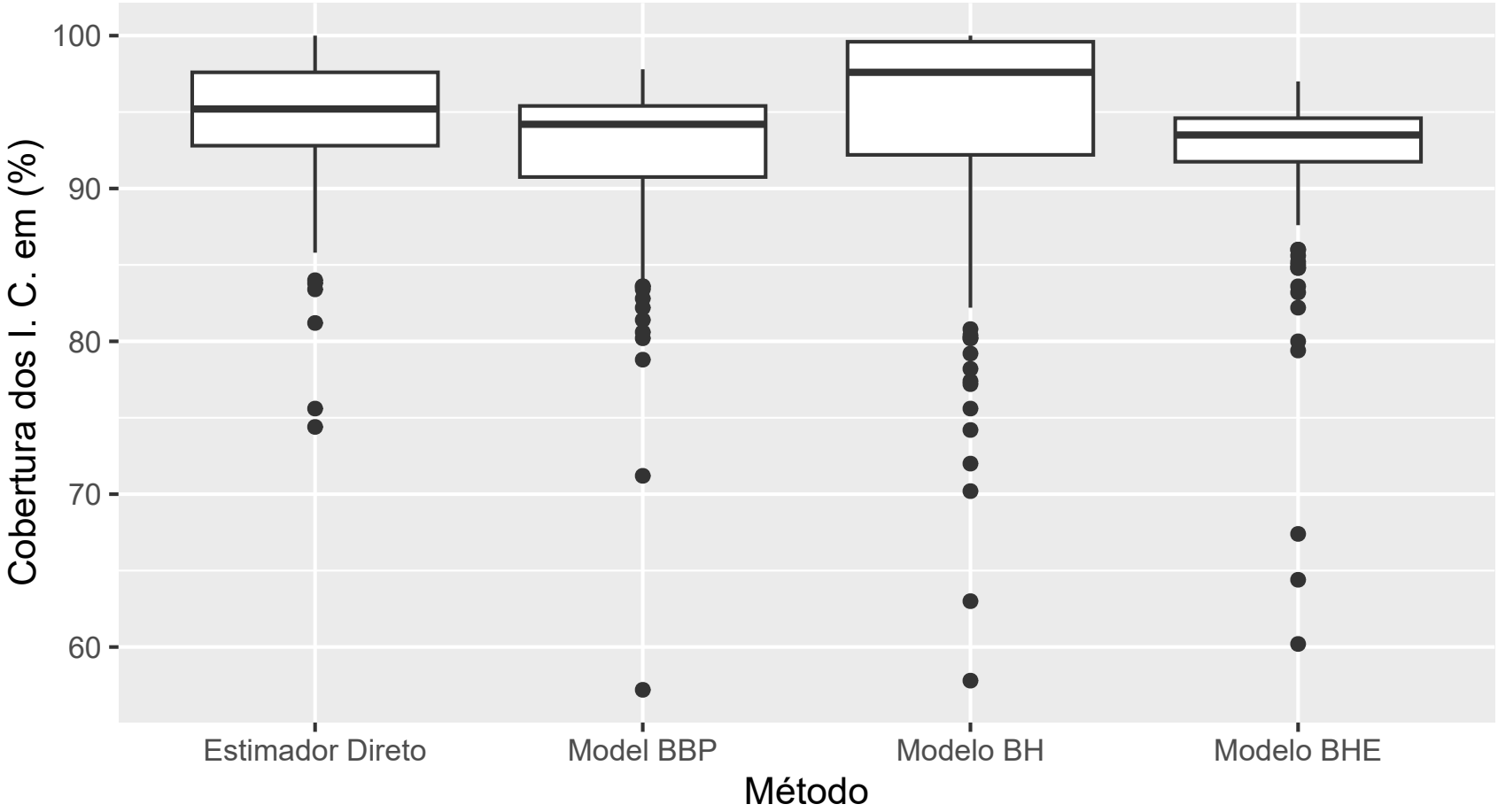

Figure : Boxplot da cobertura empírica dos intervalos de 95 % de credibilidade para os valores de μ_i e os três modelos ajustados e o procedumento de estimação direta

Estudo de Simulação com dados reais

Sumário das medidas de qualidade das estimativas pontuais e intervalares obtidas para os três modelos ajustados

Modelo	Cob. média(%)	C. I. médio	D.P.M	V.R.A.M. (%)
Modelo BH	94	0.42	0.11	17.3
Modelo BHE	92	0.38	0.10	7.4
Modelo BBP	92	0.36	0.09	8.9

Pesquisa de Orçamentos Familiares (2017-2018)

- A Pesquisa de Orçamentos Familiares - POF avalia as estruturas de consumo, despesas, renda e parte da variação patrimonial das famílias.
- Traça um perfil das condições de vida da população com base na análise dos orçamentos familiares.
- A POF 2017-2018 investigou pela primeira vez a prevalência da insegurança alimentar, segundo a Escala Brasileira de Insegurança Alimentar (EBIA).

Pesquisa de Orçamentos Familiares (2017-2018)

Desenho Amostral

- Pesquisa amostral sociodemográfica.
- Amostragem de conglomerados selecionada da "amostra mestre".
- Unidades Primárias de Amostragem (UPAs) são selecionadas com probabilidades proporcionais ao número de domicílios.
- Selecionam-se aleatoriamente domicílios para cada UPA selecionada.

Cobertura da POF

- Brasil e Grandes Regiões.
- Estados.
- Áreas metropolitanas.
- Municípios das Capitais.

Estimação de índices de prevalência de insegurança alimentar para pequenas áreas

Pequenas áreas

- Conjunto contíguos de setores censitários (UPAs) que correspondem aos estratos finais da pesquisa (Estratos-POF)
- 49 áreas pertencentes ao Estado de Minas Gerais

Distribuição do número de UPAs por pequena área (Estrato-POF)

Número de UPAs	Frequencia
3 - 5	27
6 - 10	7
11 - 15	8
16 - 20	2
21 - 25	4
26 - 28	1

Estimação de índices de prevalência de insegurança alimentar para pequenas áreas

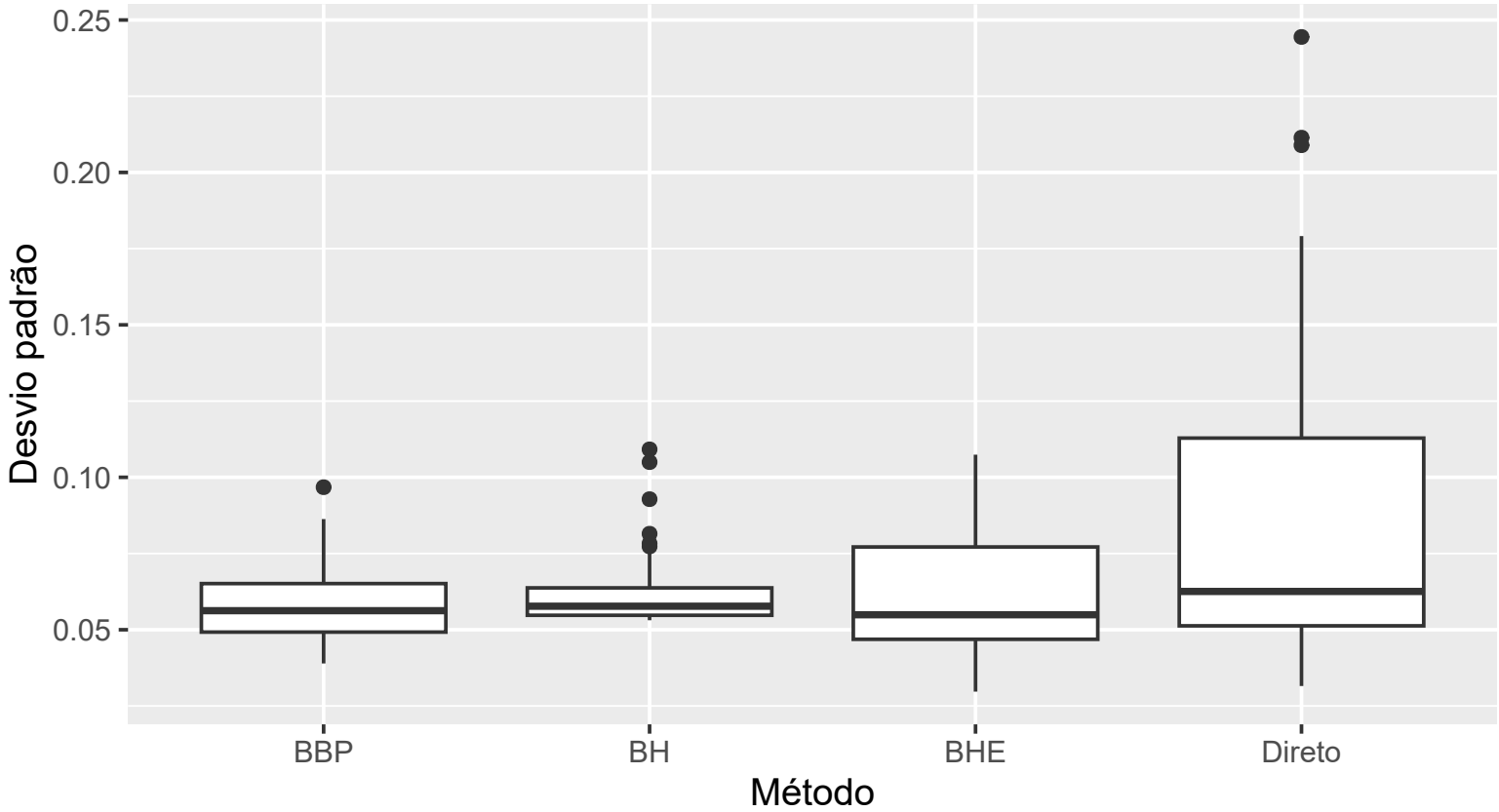

Figure : Boxplots dos Desvios padrões dos estimadores diretos e dos desvios padrões a posteriori de μ_i's para os três modelos ajustados

Comparação dos Modelos

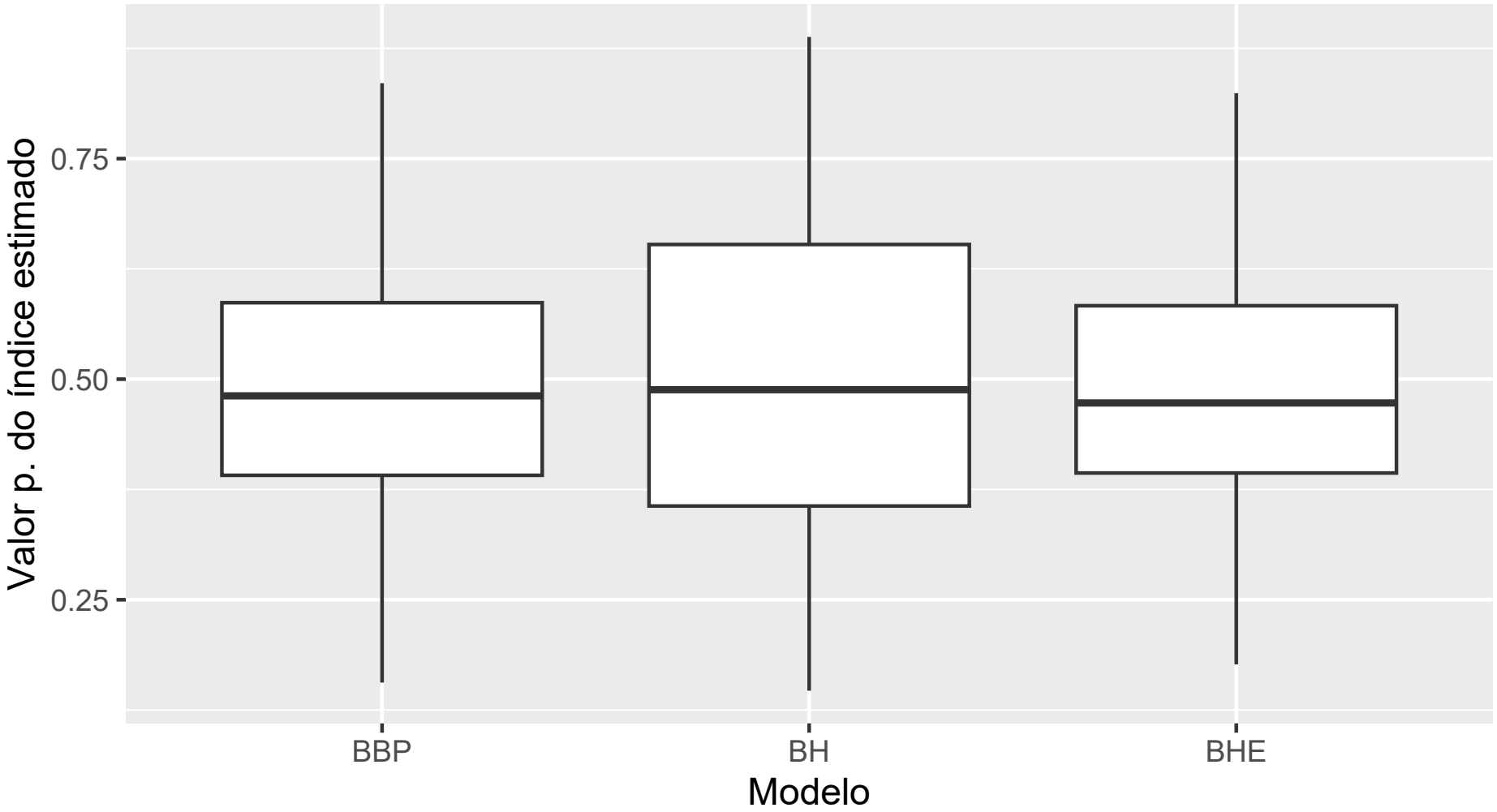

Figure : Boxplot da Probabilidade do índice replicado ser maior do que o valor estimado do índice para cada área e para os três modelos ajustados

Comparação dos Modelos

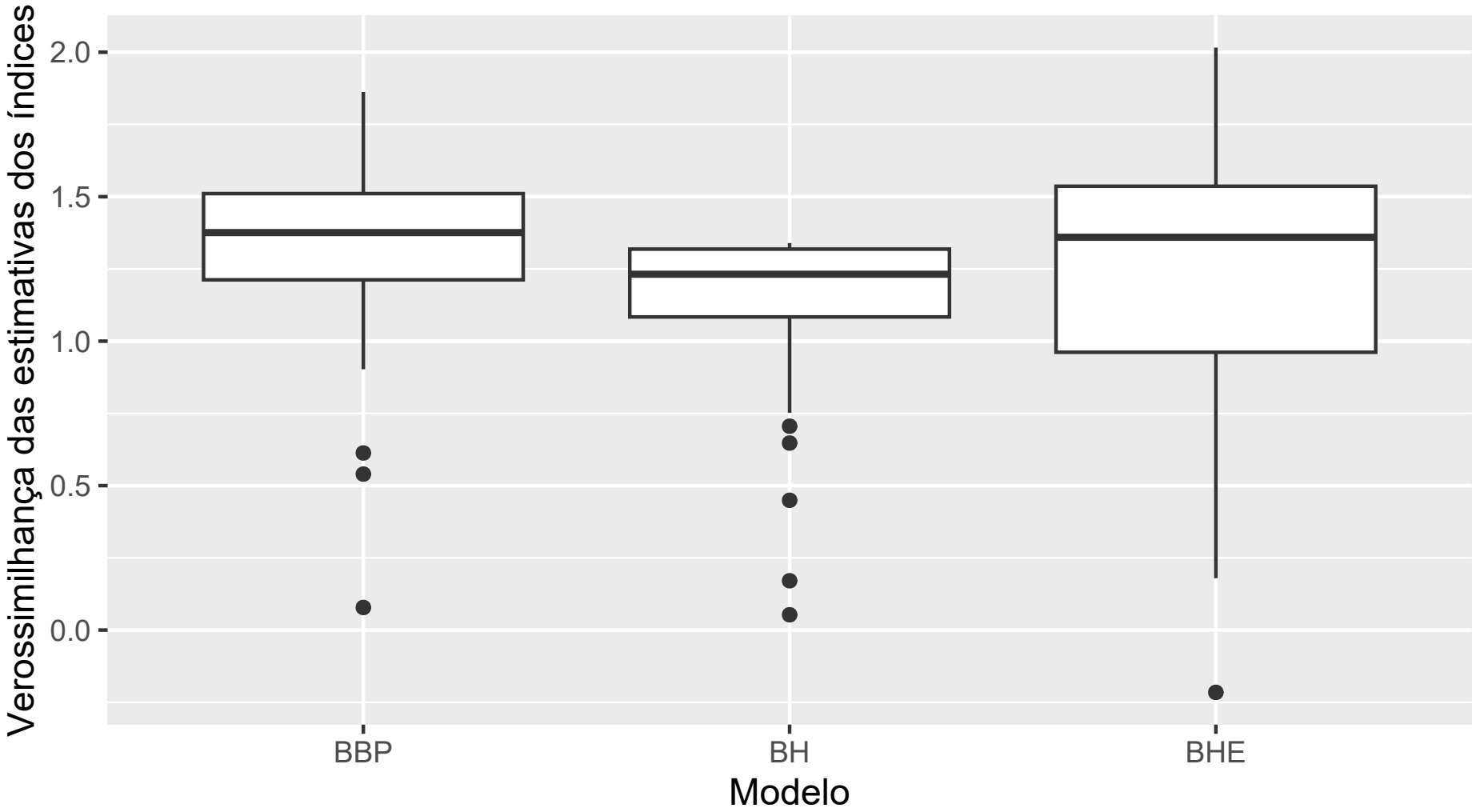

Figure : Boxplot da função de verossimilhança dos índices estimados para os três modelos ajustados

Comparação dos Modelos

Sumário das medidas de qualidade para os três modelos ajustados

Modelo	D.P.(%)	Comp. do I.C.	P-valor	Verossi. (%)
BH	6,3	0,238	0,50	1,12
BHE	6,1	0,238	0,49	1,23
BBP	5,8	0,231	0,49	1,33

Conclusões e Trabalhos Futuros

Estudo de Simulação

- Diferentemente do estimador usual da variância da razão, o método de estimação por reamostragem de Jackknife mostrou-se não tendencioso para a maioria dos setores.

- Os modelos (BHE e BBP) que usam as estimativas das precisões (ϕ_i's) estimam melhor os parâmetros μ_i's do que o modelo BH.

- Além disso, há um ganho extra quando os parâmetros ϕ_i's são hierarquicamente modelados por uma Distribuição Beta-prime.

Conclusões e Trabalhos Futuros

Aplicação

- Os modelos propostos parecem se ajustar melhor aos dados do que o modelo comumente empregado, fornecendo ainda estimativas com razoável precisão mesmo para tamanhos de amostras pequenos.
- É importante notar que nos estudos de simulação e a aplicação apresentada não foi possível utilizar covariáveis.
- A introdução de covariáveis, com bom poder de explicação, melhoraria ainda mais a qualidade do ajuste do modelo e consequentemente aumentaria a precisão das estimativas para as pequenas áreas.
- A introdução de efeitos espacialmente estruturadas poderia também aumentar acurácia das estimativas.

Referências

Benmei, L., Lahiri, P., Kalton, G., 2006. Hierarchical bayes modeling of survey-weighted small area proportions, in: Proceedings of Section on Survey Research Methods of ASA, pp. 3181–3186.

FAO, IFAD, U.W., WHO., 2021. The state of food security and nutrition in the world 2021. transforming food systems for food security, improved nutrition and affordable healthy diets for all. https://doi.org/10.4060/cb4474en.

Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. Journal of Applied Statistics 31, 799–815.

Ferraz, V., Moura, F.A.S., 2012. Small area estimation using skew normal models. Computational Statistics Data Analysis 56, 2864–2874.

Referências

Lohr, S.L., 2021. Sampling: design and analysis. CRC press.

Moura, F.A.S., Holt, D., 1999. Small area estimation using multilevel models. Survey methodology 25,73–80.

Pereira, S., Turkman, F., Correia, L., 2018. Spatio-temporal analysis of regional unemployment rates: A comparison of model based approaches. REVSTAT 16, 515–536.

Pereira, S., Turkman, K.F., Correia, L., Rue, H., 2019. Unemployment estimation: Spatial point referenced methods and models. Spatial Statistics , 100345.

You,Y. and Chapman, B.(2006). Small area estimation using area level models and estimated sampling variances. Survey Methodology, 32, 97-103.