Mapeamento de indicadores usando estimação em pequenos domínios

Kelly Cristina M. Gonçalves

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Métodos Estatísticos

DMEUFRJ

*parceiros neste trabalho: Denise Britz (IBGE), Malay Ghosh (University of Florida) e Maria Eduarda Gallo (UFRJ).

Introdução

- Pesquisas domiciliares são em geral a principal fonte de indicadores relacionados a condições de vida, pobreza e força de trabalho.
- Problemas: amostras muito pequenas para serem representativas em grupos menores (geográficos, demográficas); amostras não cobrem todas os grupos.
- O censo populacional cobre supostamente 100% da população permitindo acesso a pequenos grupos, mas tem em geral pouca informação sobre indicadores de interesse. Além de estar defasado muitas vezes.

Introdução

- Uma solução: combinar dados do censos e de pesquisas para troca de informações.
- A estimação em pequenos domínios é um ramo focado em melhorar a confiabilidade das estimativas e as medidas de incerteza associadas para populações onde as amostras não produzem estimativas suficientemente confiáveis.

Objetivos da apresentação

- Difundir a área de investigação e suas aplicações práticas.
- Apresentar dois trabalhos na área: (1) desenvolvimento de metodologia motivada por problema prático relacionado a políticas públicas; (2) desenvolvimento de metodologia motivada por tornar o problema de estimação menos custoso computacionalmente.

Estimação em pequenos domínios: motivação

- A erradicação da pobreza extrema para todas as pessoas em todos os lugares ainda é um tema importante hoje, sendo este um dos Objetivos de Desenvolvimento Sustentável (ODS).

- Para que isso seja possível, utiliza-se o indicador dado pela proporção de pessoas abaixo da linha da pobreza (por exemplo), definida como US$ 1,90 por dia.

Estimação em pequenos domínios: motivação

- Foster et al. (1984) propuseram a seguinte medida de pobreza:

$$F_{\alpha,i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\frac{z - y_{ij}}{z} \right)^{\alpha} I(y_{ij} < z),$$

em que y_{ij} é a renda familiar per capita para o indivíduo j no estrato i, z é a linha de pobreza e n_i é o número de pessoas na população no estrato i.

- Existem três métricas principais definidas como: Incidência da Pobreza ($\alpha = 0$), *Gap* da Pobreza ($\alpha = 1$) e Severidade da Pobreza ($\alpha = 2$).

Estimação em pequenos domínios: motivação

Instituto Brasileiro de Geografia e Estatística (IBGE) já fornece estimativas da proporção de pessoas abaixo da linha da pobreza, desagregadas por:

- Sexo;
- situação do domicílio (urbano ou rural);
- raça;
- faixas etárias;
- regiões.

Estimação em pequenos domínios: motivação

Instituto Brasileiro de Geografia e Estatística (IBGE) já fornece estimativas da proporção de pessoas abaixo da linha da pobreza, desagregadas por:

- Sexo;
- situação do domicílio (urbano ou rural);
- raça;
- faixas etárias;
- regiões.

Neri (2022), dos Santos e Aruto (2022) e Antonaci (2012) mostram a importância de estimativas mais desagregadas geograficamente para o combate à pobreza. Exemplo: 146 estratos de municípios - conjuntos de municípios (definidos pelo IBGE).

Estimação em pequenos domínios: motivação

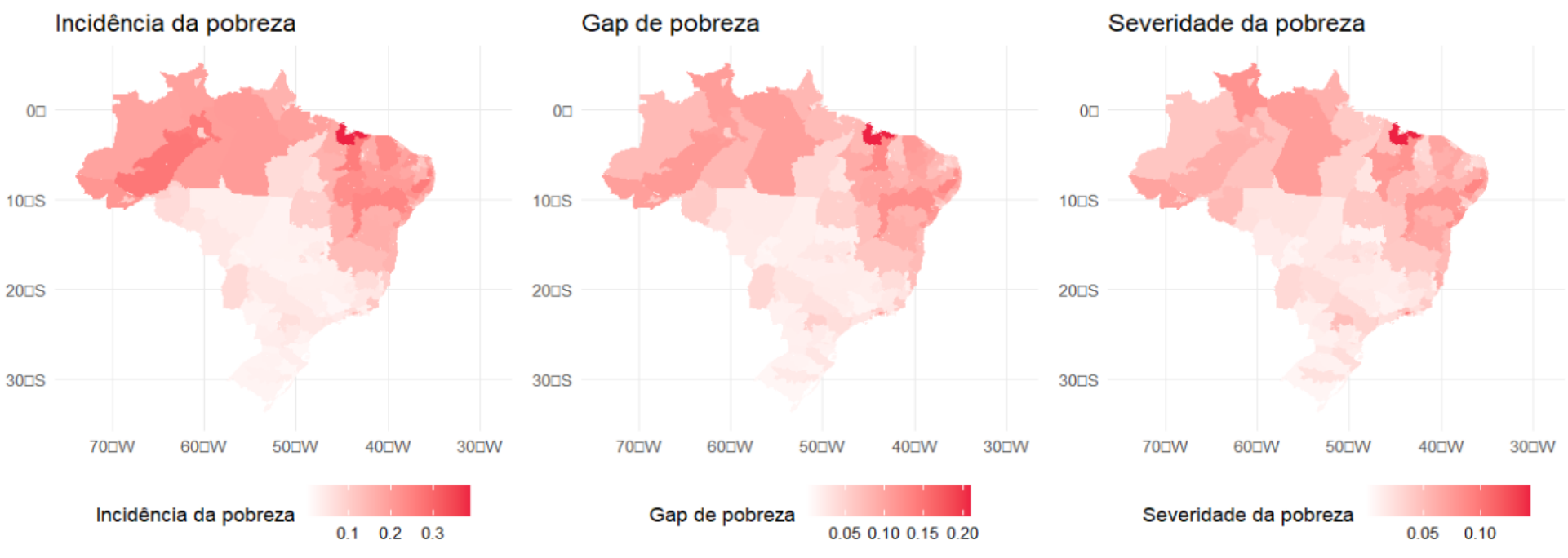

Figura: Estimativa direta das medidas de pobreza por estrato municipal no ano de 2021 no Brasil.

Estimação em pequenos domínios: motivação

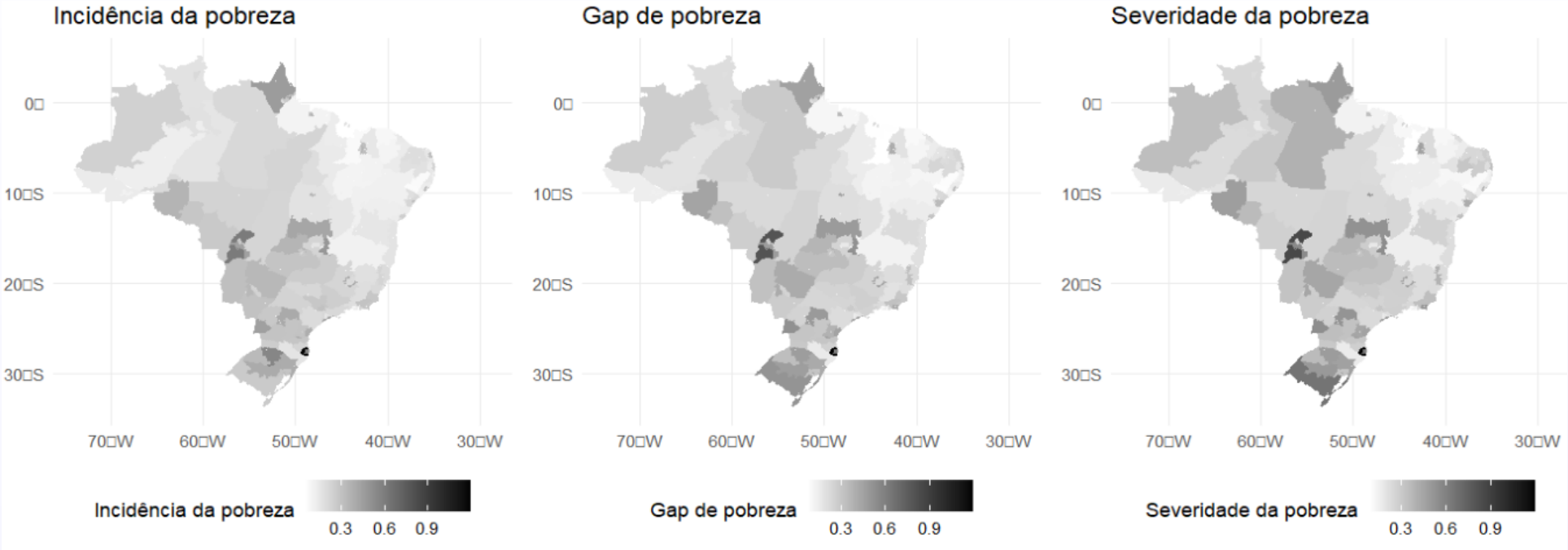

Figura: Coeficiente de variação da estimativa direta das medidas de pobreza por estrato municipal no ano de 2021 no Brasil.

Tabela: Classificação das estimativas quanto ao Coeficiente de Variação (CV)

Valor do CV	Qualidade da estimativa
Zero	"Exata"
Menor que 0.05	Excelente
Entre 0.05 e 0.15	Boa
Entre 0.15 e 0.30	Razoável
Entre 0.30 e 0.50	Baixa precisão
Maior que 0.50	Imprecisa

Fonte: IBGE.

Estimação em pequenos domínios: motivação

Tabela: Classificação das estimativas quanto ao Coeficiente de Variação (CV)

Valor do CV	Qualidade da estimativa
Zero	"Exata"
Menor que 0.05	Excelente
Entre 0.05 e 0.15	Boa
Entre 0.15 e 0.30	Razoável
Entre 0.30 e 0.50	Baixa precisão
Maior que 0.50	Imprecisa

Fonte: IBGE.

Pergunta:

Como tornar as estimativas da pobreza publicáveis para apoiar o planejamento de políticas públicas?

Estimação em pequenos domínios

- Nos últimos anos, a demanda por estatísticas de pequenas áreas aumentou muito em todo o mundo.
- Dependendo do nível de desagregação de uma pesquisa por amostragem, **estimativas diretas** (que não usam modelagem) podem ser **imprecisas** e não confiáveis.
- Uma **área pequena** é uma unidade de amostragem onde a amostra não é grande o suficiente para produzir **estimativas diretas** (*baseadas em desenho*) com precisão adequada.
- As estimativas diretas para essas áreas apresentam altas variâncias e altos coeficientes de variação.

Atenção:

- Uma área pequena não é necessariamente uma área com tamanho pequeno!
- A classificação cruzada (isto é, geográfica, demográfica) geralmente leva a amostras pequenas, mesmo em pesquisas muito grandes!

Estimação em pequenos domínios

- Os métodos de **Estimação em pequenos domínios (SAE)** utilizam outras bases de dados, através de modelos estatísticos, para **resolver ou mitigar este problema** (Rao e Molina, 2015).

- Melhores estimativas para estas áreas são obtidas através da "empréstimo de informação"de outras áreas semelhantes.

Estimação em pequenos domínios

- Os métodos de **Estimação em pequenos domínios (SAE)** utilizam outras bases de dados, através de modelos estatísticos, para **resolver ou mitigar este problema** (Rao e Molina, 2015).
- Melhores estimativas para estas áreas são obtidas através da "empréstimo de informação"de outras áreas semelhantes.
- Exemplos de fontes de empréstimo de informações: Registros Administrativos, Censos, outras pesquisas, etc.

Estimação em pequenos domínios

- Quando implementado corretamente e quando boas informações auxiliares podem ser encontradas, SAE pode:

Estimação em pequenos domínios

- Quando implementado corretamente e quando boas informações auxiliares podem ser encontradas, SAE pode:
  - Diminuir medidas de incerteza (EQM/variâncias/CVs) em relação aos estimadores diretos.
  - Permitir a publicação de estatísticas que de outra forma seriam suprimidas.
  - Permitir a liberação de estimativas em níveis mais baixos de agregação.
  - Permitir a produção de estimativas para áreas sem levantamento de amostra.

Aplicações

- Renda, desigualdade, medidas de pobreza, desemprego, fatores de risco de câncer, proporções de pessoas com deficiência, etc.

Aplicações

- Renda, desigualdade, medidas de pobreza, desemprego, fatores de risco de câncer, proporções de pessoas com deficiência, etc.

Importância de SAE

- A necessidade de estatísticas num nível inferior de agregação tem aumentando em todo o mundo.
- Pode ajudar a informar políticas baseadas em dados e na alocação eficiente de recursos.

Notação

- m domínios de interesse, por ex. geográficos, grupos demográficos, classificação cruzada de cada um (áreas pequenas);
- Estimadores diretos podem ter uma amostra pequena (ou nenhuma) para algumas das as áreas;
- N_i: tamanho da i-ésima pequena área;
- uma amostra de tamanho n_i é retirada da i-ésima área;
- y_{ij} denota a observação da pesquisa para j-ésima unidade no i-ésimo domínio;
- Pesos amostrais associados w_{ij}.

Notação

- Estimador direto: baseado em dados de amostra apenas para o domínio de interesse;
- $\hat{Y}_i = \sum_{j=1}^{n_i} w_{ij} y_{ij}$;
- Associado a $\hat{Y}_i$ está alguma estimativa de seu erro padrão (calculado analiticamente, via linearização, métodos de replicação, etc.).

Estimação em pequenos domínios

- Modelos de pequenas áreas podem ser classificados como:
  - (i) modelos a nível de área;
  - (ii) modelos a nível de unidade.

Estimação em pequenos domínios

- Modelos de pequenas áreas podem ser classificados como:
  - (i) modelos a nível de área;
  - (ii) modelos a nível de unidade.

- O primeiro é utilizado com mais frequência do que o último porque os modelos ao nível da unidade requerem informações para as unidades individuais da amostra, o que normalmente não está disponível para fins secundários.

Estimação em pequenos domínios

- Modelos de pequenas áreas podem ser classificados como:
  - (i) modelos a nível de área;
  - (ii) modelos a nível de unidade.
- O primeiro é utilizado com mais frequência do que o último porque os modelos ao nível da unidade requerem informações para as unidades individuais da amostra, o que normalmente não está disponível para fins secundários.
- A escolha por modelos de área ou unidade depende da aplicação e dos dados disponíveis!

Modelos a nível de área

Trabalho em conjunto com Denise Britz (ENCE/ IBGE e vice-presidente do International Statistical Institute) e Maria Eduarda Gallo (mestranda PPGE UFRJ).

Modelos a nível de área

Trabalho em conjunto com Denise Britz (ENCE/ IBGE e vice-presidente do International Statistical Institute) e Maria Eduarda Gallo (mestranda PPGE UFRJ).

Objetivo: estimação de indicadores de pobreza confiáveis para estratos de municípios no Brasil.

Modelos a nível de área

- Usa apenas estimadores diretos de pesquisa por amostragem e estimativas de variações amostrais associadas (ou tamanho efetivo da amostra) para cada área.
- Não precisa de acesso aos dados em nível de unidade, que muitas vezes podem ser difícil de obter devido à confidencialidade.
- Só precisa de informações auxiliares no nível da área.

Modelos a nível de área

Modelo de Fay-Herriot (Fay and Herriot, 1979) é descrito como: para $i = 1, \ldots, m$

$$\hat{\theta}_i = \theta_i + e_i,$$

$$\theta_i = \boldsymbol{x_i^T \beta} + \nu_i, \text{ onde}$$

- θ_i é a característica populacional de interesse para a área i;
- $\hat{\theta}_i$ é a estimativa direta de θ_i obtida na pesquisa;
- e_i é o erro de amostragem, tal que $e_i \sim Normal(0, \sigma_i^2)$, σ_i^2 fixo;
- ν_i é o efeito aleatório de área, tal que $\nu_i \sim Normal(0, \sigma_\nu^2)$;
- $\boldsymbol{x_i}$ é o $p-$vetor com as variáveis auxiliares para o domínio i;
- $\boldsymbol{\beta}$ é o $p-$vetor dos coeficientes de regressão.

Modelos a nível de área

Modelo de Fay-Herriot (Fay and Herriot, 1979) é descrito como: para $i = 1, \ldots, m$

$$\hat{\theta}_i = \theta_i + e_i,$$

$$\theta_i = \boldsymbol{x_i^T \beta} + \nu_i, \text{ onde}$$

- θ_i é a característica populacional de interesse para a área i;
- $\hat{\theta}_i$ é a estimativa direta de θ_i obtida na pesquisa;
- e_i é o erro de amostragem, tal que $e_i \sim Normal(0, \sigma_i^2)$, σ_i^2 fixo;
- ν_i é o efeito aleatório de área, tal que $\nu_i \sim Normal(0, \sigma_\nu^2)$;
- $\boldsymbol{x_i}$ é o $p-$vetor com as variáveis auxiliares para o domínio i;
- $\boldsymbol{\beta}$ é o $p-$vetor dos coeficientes de regressão.
- Objetivo: estimar θ_i.

Modelos a nível de área

- O primeiro nível é referido como o modelo de amostragem.
- O segundo nível é referido como o modelo de ligação.
- A variância amostral é geralmente assumida como conhecida para identificação, sendo na prática estimada a partir de microdados.

Modelos de área: inferência Bayesiana

- Atribui-se distribuições a priori para os parâmetros do modelo.
- Prioris não informativas tentam fazer suposições mínimas a priori sobre os parâmetros associados.
- Calculamos ou aproximamos distribuições a posteriori de parâmetros de interesse, geralmente usando Markov-Chain Monte Carlo (MCMC).
- Calculamos médias a posteriori de parâmetros de interesse das pequenas áreas.
- Tipicamente reporta-se variâncias ou desvios a posteriori (definindo o grau de incerteza associado às estimativas obtidas com base no modelo). **Neste ponto é que se espera vantagens em usar o modelo e não as estimativas diretas.**

Elementos de sucesso no uso da metodologia

- Pesquisa amostral que mede a quantidade de interesse aproximadamente não viesado, mas com tamanho amostral limitado.
- Rica informação auxiliar e expertise.
- Diversos *softwares* com modelos de SAE implementados e disponíveis para uso.

Elementos de sucesso no uso da metodologia

- Pesquisa amostral que mede a quantidade de interesse aproximadamente não viesado, mas com tamanho amostral limitado.
- Rica informação auxiliar e expertise.
- Diversos *softwares* com modelos de SAE implementados e disponíveis para uso.
- Dica: *toolkit - https://unstats.un.org/wiki/display/SAE4SDG/*

Problema 1: Estimação de indicadores de pobreza

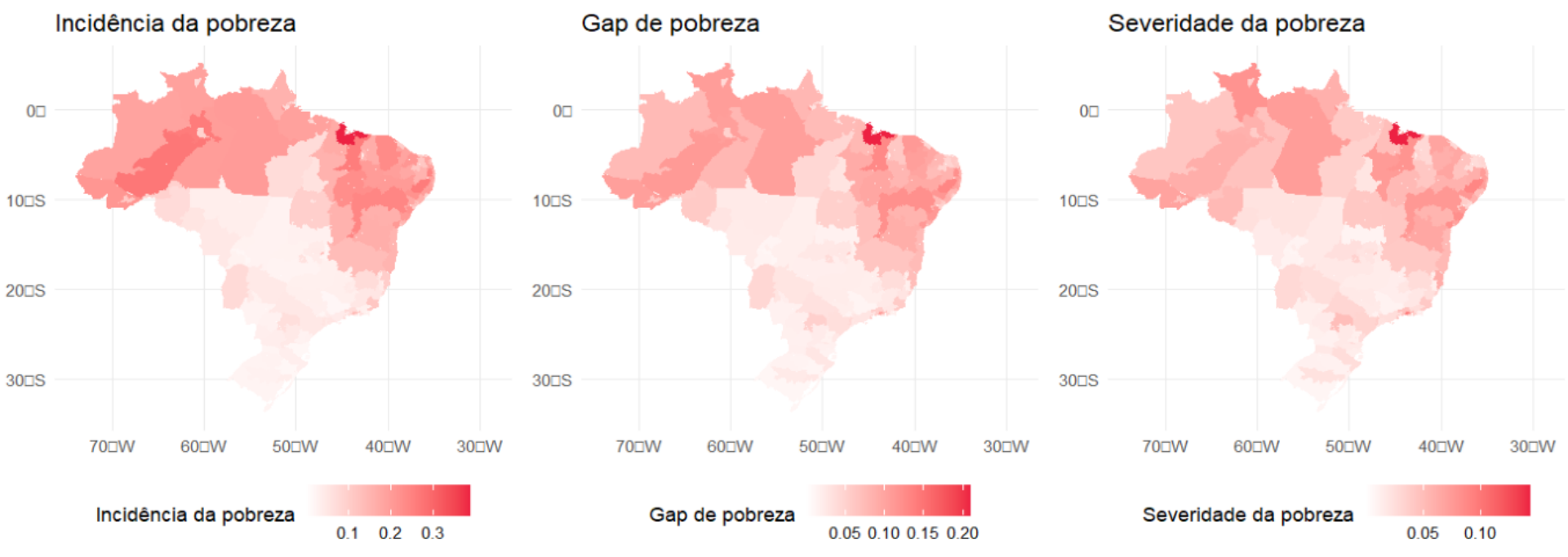

Figura: Estimativa direta das medidas de pobreza por estrato municipal no ano de 2021 no Brasil.

Problema 1: Estimação de indicadores de pobreza

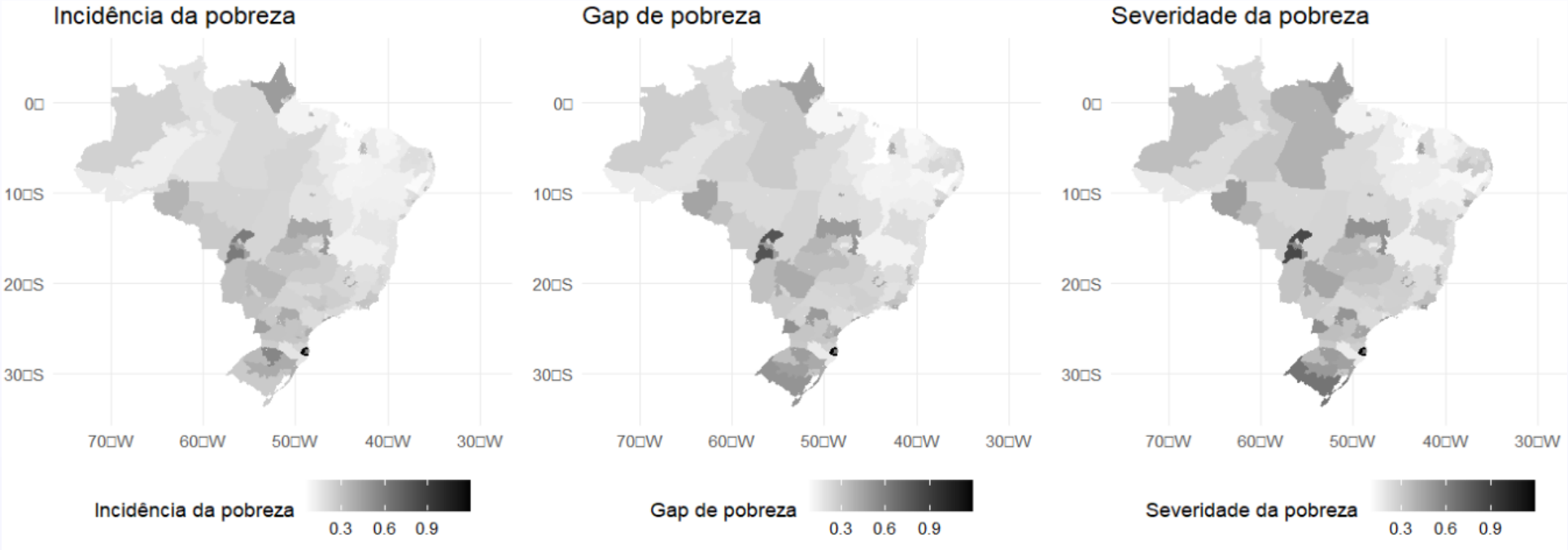

Figura: Coeficiente de variação da estimativa direta das medidas de pobreza por estrato municipal no ano de 2021 no Brasil.

Modelo

Modelo Beta é apropriado para estes indicadores devido ao suporte (Janicki, 2019): para $i = 1, \ldots, m$,

$$\hat{\theta}_i | \theta_i, \phi_i \sim Beta(\theta_i \phi_i, (1 - \theta_i)\phi_i)$$

$$g(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) = \boldsymbol{x_i^T \beta} + \nu_i, \ \nu_i \sim Normal(0, \sigma_\nu^2)$$

Modelo

Modelo Beta é apropriado para estes indicadores devido ao suporte (Janicki, 2019): para $i = 1, \ldots, m$,

$$\hat{\theta}_i | \theta_i, \phi_i \sim Beta(\theta_i \phi_i, (1 - \theta_i)\phi_i)$$

$$g(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) = \boldsymbol{x_i^T \beta} + \nu_i, \ \nu_i \sim Normal(0, \sigma_\nu^2)$$

- $\hat{\theta}_i$ é o estimador direto da medida de pobreza (incidência, gap ou severidade).
- Objetivo: estimar θ_i com variância reduzida!

Dados

- *Pesquisa Nacional por Amostra Domiciliar Contínua* (PNADC) do IBGE.
- A **PNADC** tem como objetivo produzir indicadores que permitam o acompanhamento da força de trabalho e situação socioeconômica da população.
- Variáveis auxiliares por área: proporção de pessoas no **CADUNICO** com rendimento inferior a 0,5SM; proporção de escolas sem esgoto; proporção de matrículas na educação infantil; proporção de matrículas no ensino fundamental *Educação de Jovens e Adultos* (EJA) (**Censo escolar**); participação da indústria no PIB.

Resultados

Modelo	Variáveis
M1	Proporção de pessoas no CADUNICO com renda menor que $0,5SM$
M2	M1 + Indicadora se o estrato de município fica no Norte ou Nordeste
M3	M2 + Interação entre as duas variáveis
M4	M1 + *dummy* para as regiões
M5	Proporção de escolas sem esgoto Proporção de matrículas na educação infantil Proporção de matrículas no EJA Fundamental Participação da Indústria no PIB
M6	M5 + *dummy* para as regiões

Estimação de indicadores de pobreza

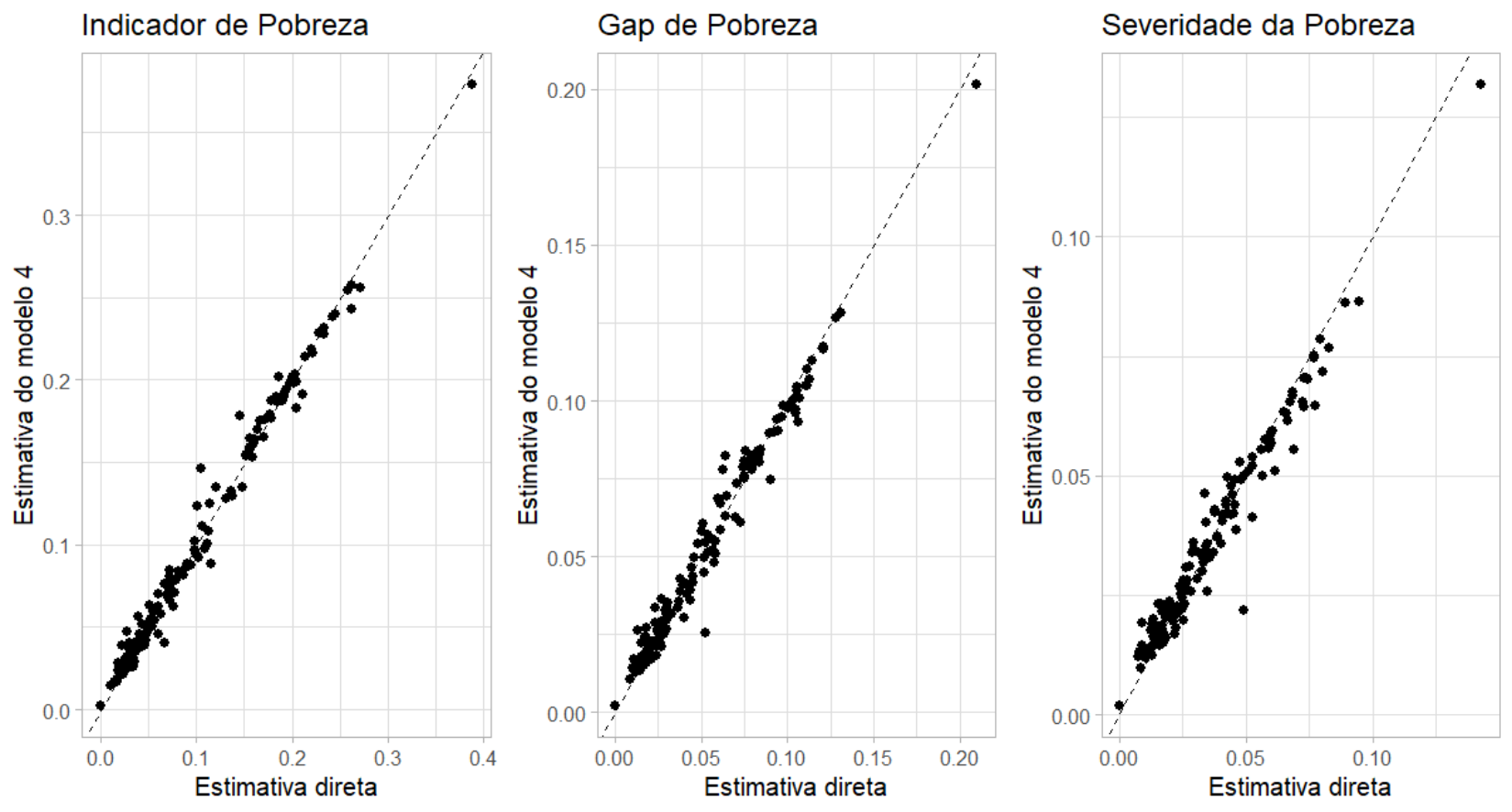

Figura: Estimativa direta vs com base no modelo (M4) para cada medida de pobreza.

Estimação de indicadores de pobreza

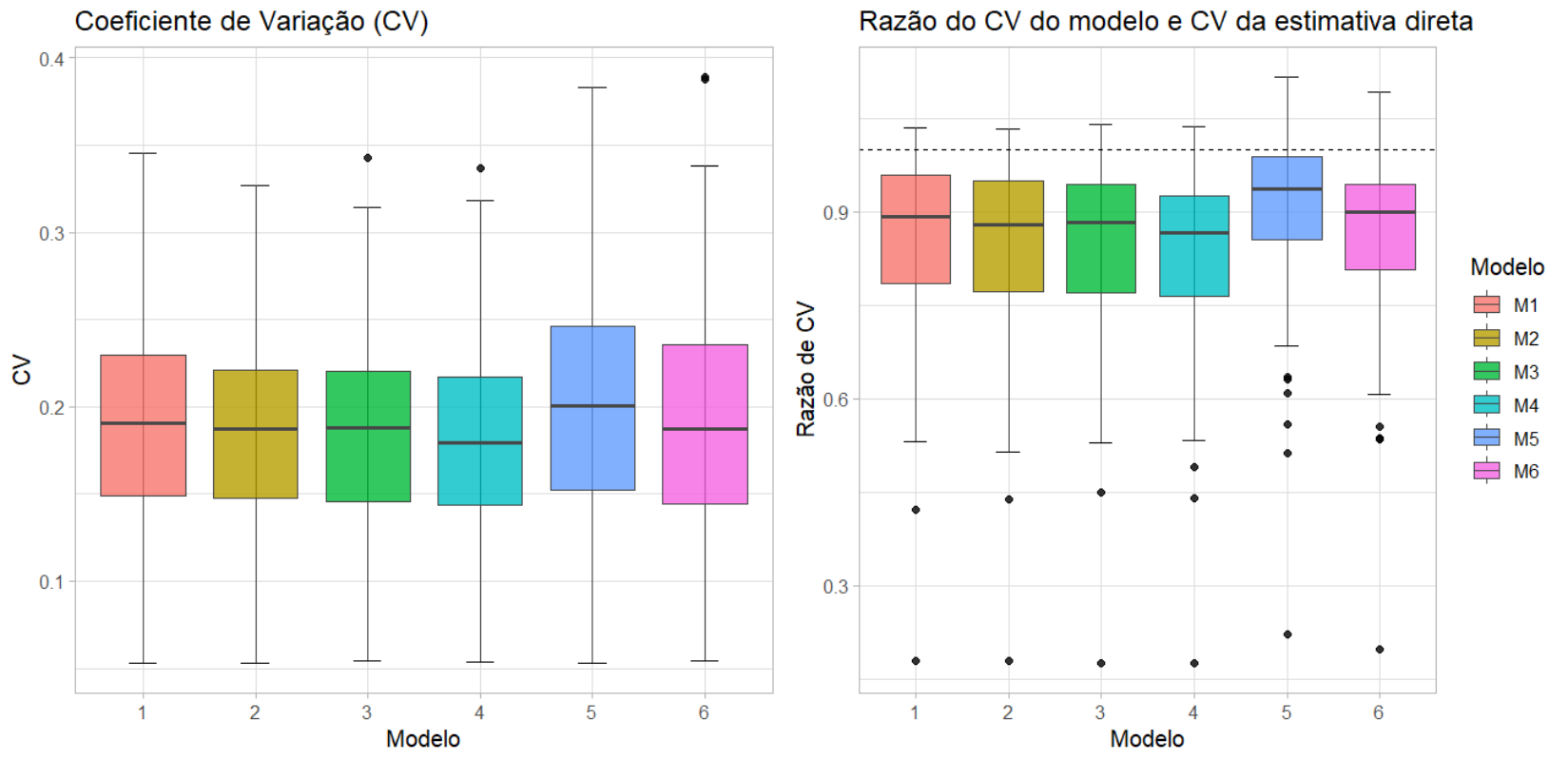

Figura: Boxplot do Coeficiente de Variação (CV) e da Razão de CV (estimativas dos modelos/ direta) para a incidência da pobreza.

Resultados

Tabela: Quantidade de estratos de município por categoria de CV.

Modelo	Boa	Razoável	Pouco precisa	Imprecisa
Estimativa direta	32	82	28	4
M1	38	103	5	0
M2	41	101	4	0
M3	39	102	5	0
M4	**43**	**99**	**4**	**0**
M5	35	99	12	0
M6	42	92	12	0

Resultados

Tabela: Quantidade de estratos de município por categoria de CV.

Modelo	Boa	Razoável	Pouco precisa	Imprecisa
Estimativa direta	32	82	28	4
M1	38	103	5	0
M2	41	101	4	0
M3	39	102	5	0
M4	**43**	**99**	**4**	**0**
M5	35	99	12	0
M6	42	92	12	0

Melhora na qualidade das estimativas de pobreza, tornando-as publicáveis!

Resultados

Tabela: Média do CV da estimativa direta e dos modelos M1 a M6 de incidência da pobreza por tamanho de amostra nos estratos.

Modelo	< 1000	1000 a 2000	2000 a 3000	3000 a 5000	> 5000
N estratos	7	26	38	58	17
Est dir	0.51	0.28	0.25	0.21	0.13
M1	0.24	0.21	0.20	0.18	0.13
M2	0.24	0.20	0.20	0.18	0.12
M3	0.24	0.21	0.20	0.18	0.12
M4	0.23	0.20	0.19	0.18	0.12
M5	0.28	0.23	0.22	0.19	0.13
M6	0.26	0.22	0.21	0.19	0.12

Indicadores de pobreza ao longo do tempo

Série temporal com as estimativas diretas disponível de 2012 a 2022.

Pontos a se pensar

- Avaliar tais indicadores ao longo do tempo pode trazer resultados interessantes?
- Incorporação de série histórica traz maior precisão às estimativas?

Indicadores de pobreza ao longo do tempo

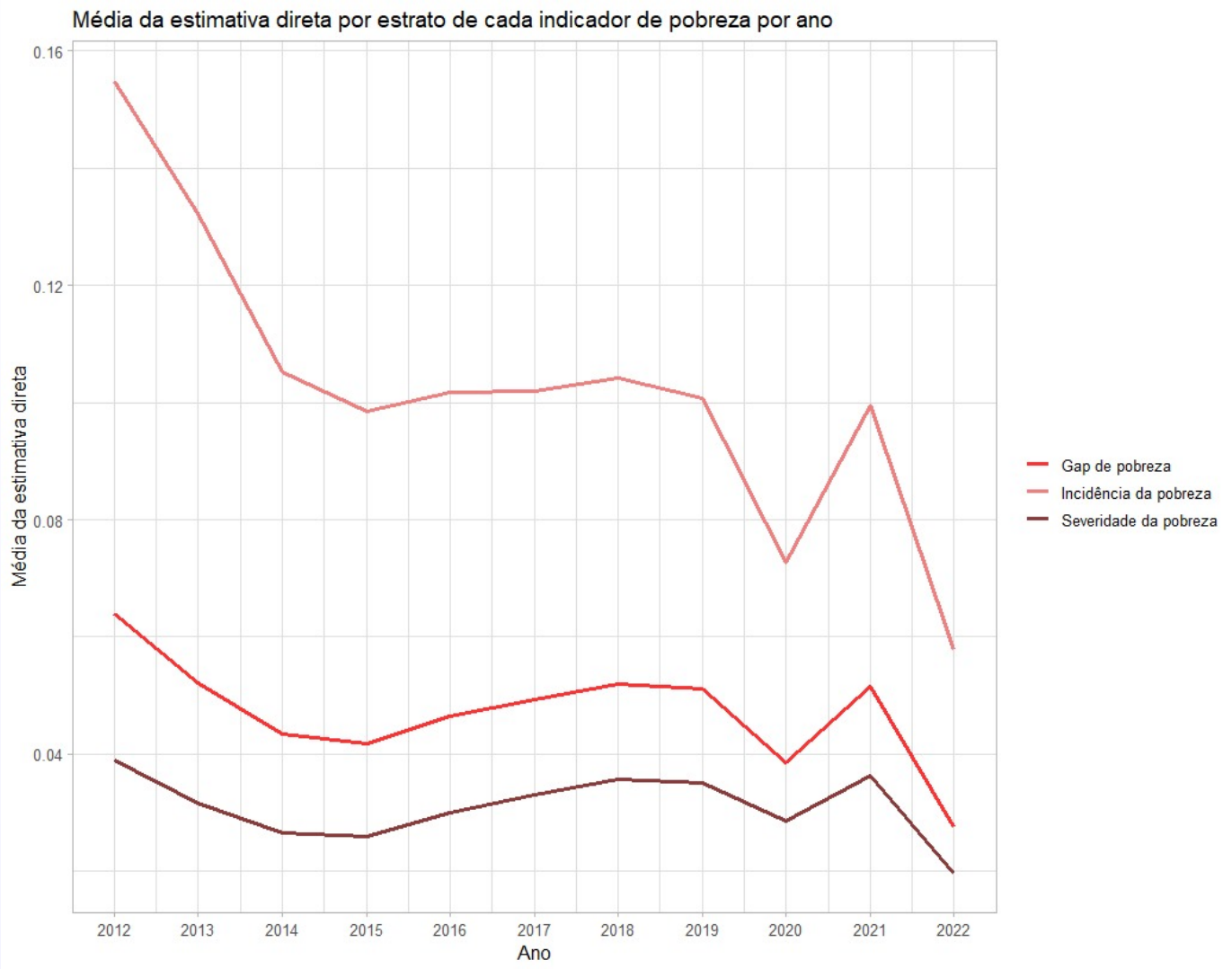

Figura: Estimativas diretas (média por estrato) por ano.

Indicadores de pobreza ao longo do tempo

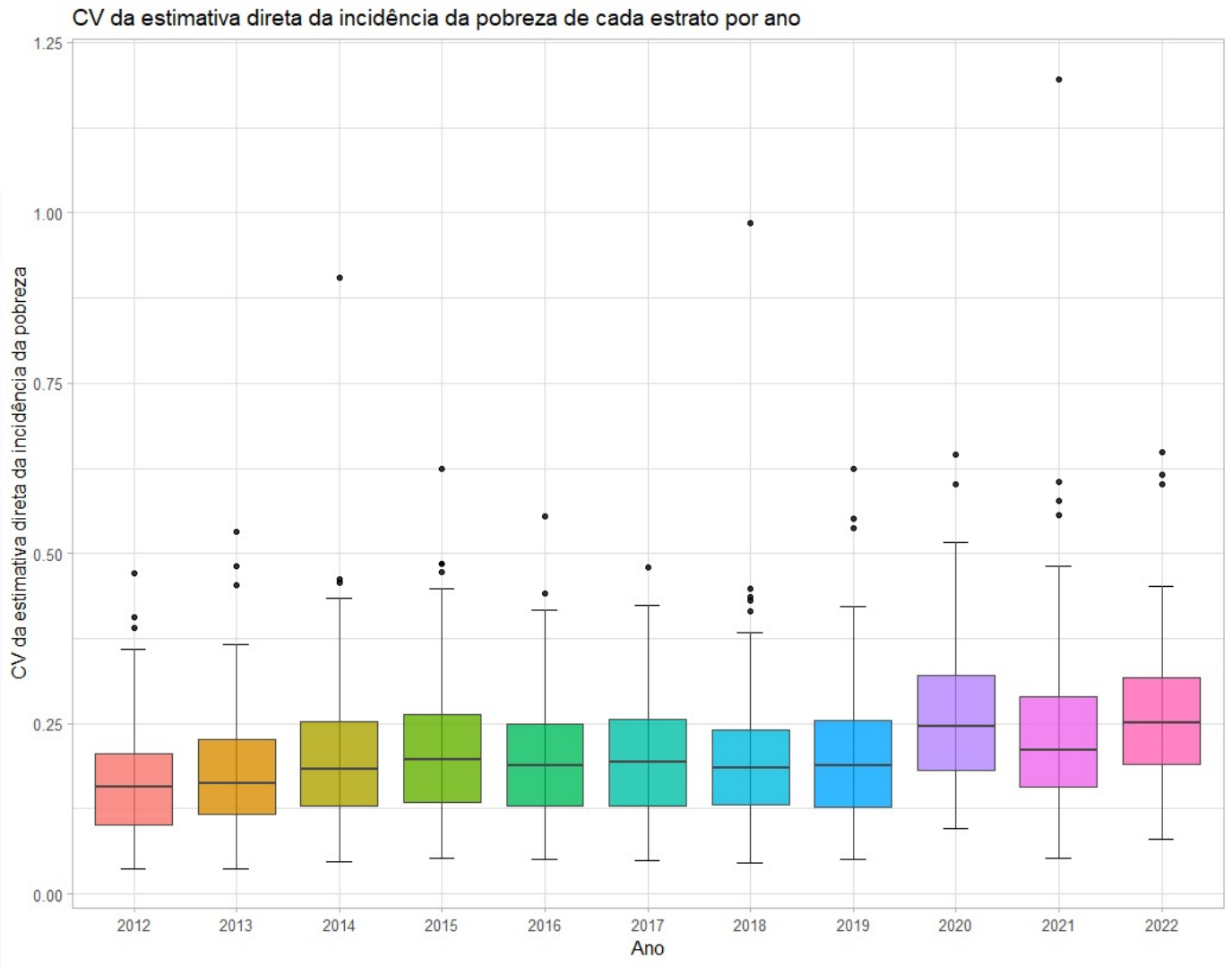

Figura: CV das estimativas diretas da incidência de pobreza por ano para cada estrato.

Com a inclusão da componente temporal o modelo é descrito da seguinte forma: para $i = 1, \ldots, m$ e $t = 1, \ldots, T$

$$\hat{\theta}_{it} | \theta_{it}, \phi_{it} \sim Beta(\theta_{it}\phi_{it}, (1 - \theta_{it})\phi_{it})$$

$$g(\theta_{it}) = log\left(\frac{\theta_{it}}{1 - \theta_{it}}\right) = \boldsymbol{z_{it}^T \beta} + \nu_i + u_t$$

$$\nu_i \sim Normal(0, \sigma_\nu^2)$$

Nesse caso, é adicionado ao modelo a componente temporal u_t.

Indicadores de pobreza ao longo do tempo

Variações do modelo:

- M1t: $u_t = 0$
- M2t: $u_t \sim Normal(0, \sigma_u^2)$
- M3t: $u_t | u_{t-1} \sim Normal(u_{t-1}, \sigma_u^2)$

Variações do modelo:

- M1t: $u_t = 0$
- M2t: $u_t \sim Normal(0, \sigma_u^2)$
- M3t: $u_t|u_{t-1} \sim Normal(u_{t-1}, \sigma_u^2)$

Modelo M2t sofre de problemas de identificabilidade, exigindo imposição de restrição. Impôs-se restrição de soma zero para ν_i e u_t.

Incidência de pobreza ao longo do tempo

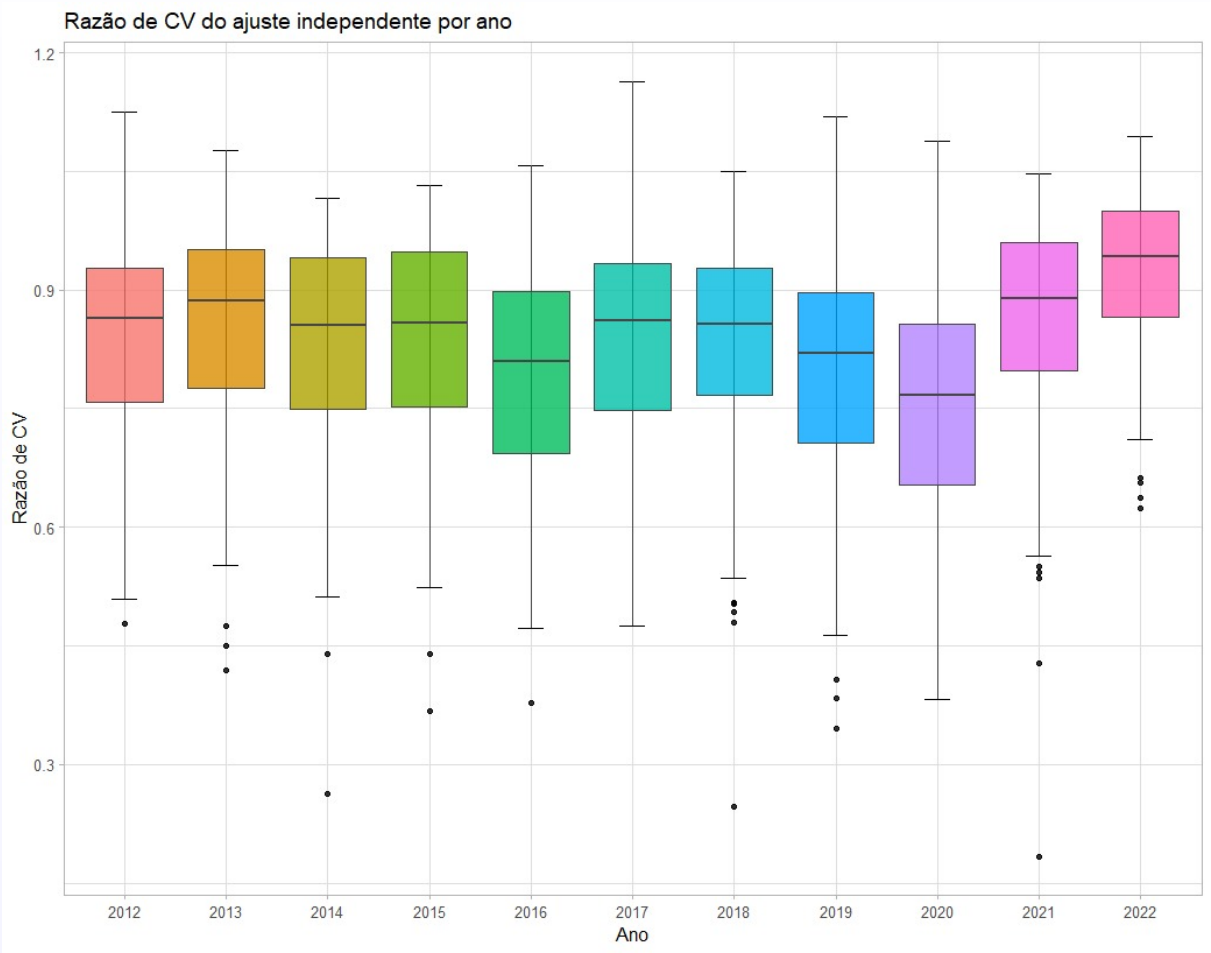

Figura: Razão de CV (estimativas dos modelos/ direta) ao longo do tempo.

Incidência de pobreza ao longo do tempo

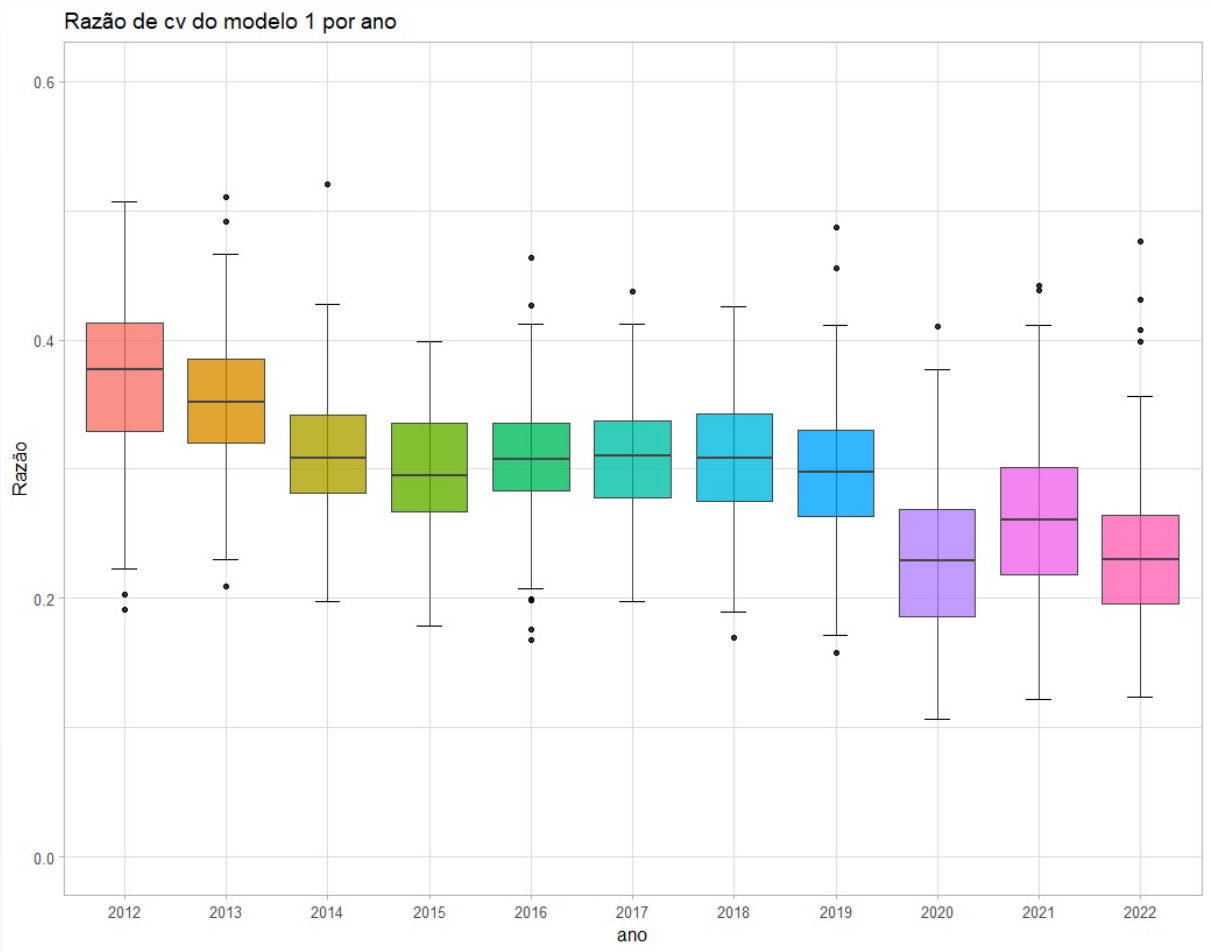

Figura: Mod1: Razão de CV ao longo do tempo.

Incidência de pobreza ao longo do tempo

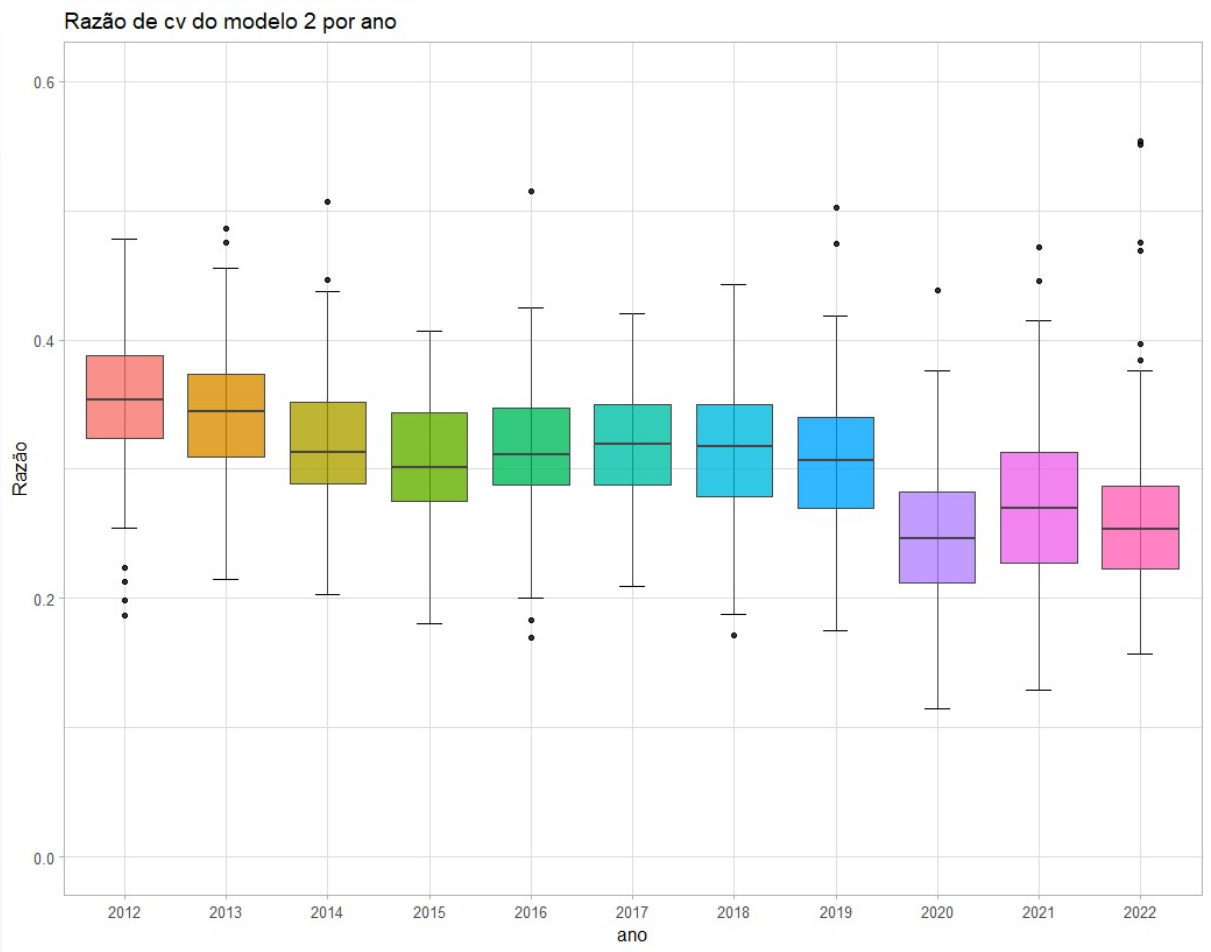

Figura: Mod2: Razão de CV ao longo do tempo.

Incidência de pobreza ao longo do tempo

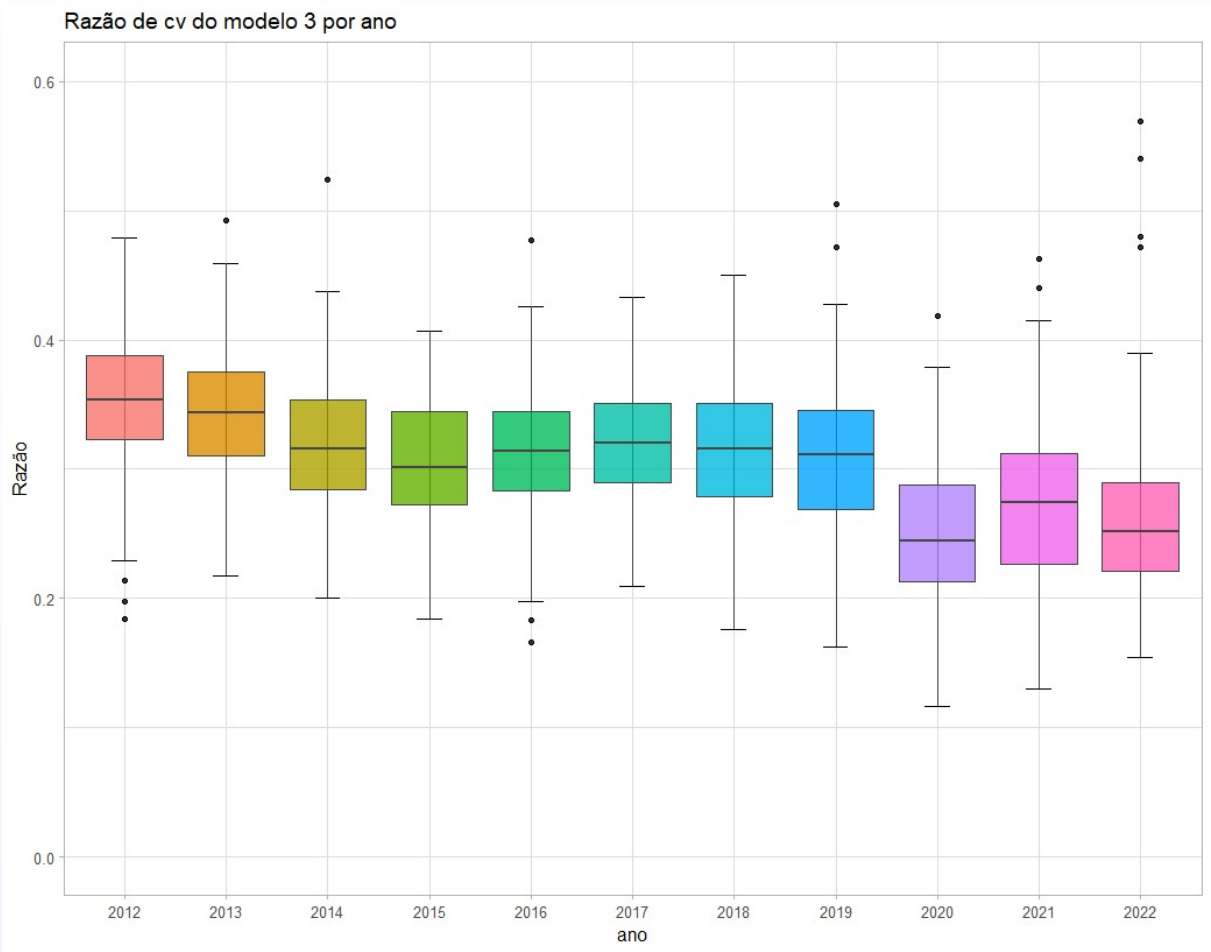

Figura: Mod3: Razão de CV ao longo do tempo.

Incidência de pobreza ao longo do tempo

Conclusão:

- Modelos que incorporem o tempo em sua formulação geram estimativas ainda mais confiáveis (mais informação).
- Modelos 2 e 3 apresentam melhores resultados!

Incidência de pobreza ao longo do tempo

Conclusão:

- Modelos que incorporem o tempo em sua formulação geram estimativas ainda mais confiáveis (mais informação).
- Modelos 2 e 3 apresentam melhores resultados!
- Pergunta: qual o comportamento da precisão das estimativas com a incorporação de cada ano?

Incidência de pobreza ao longo do tempo

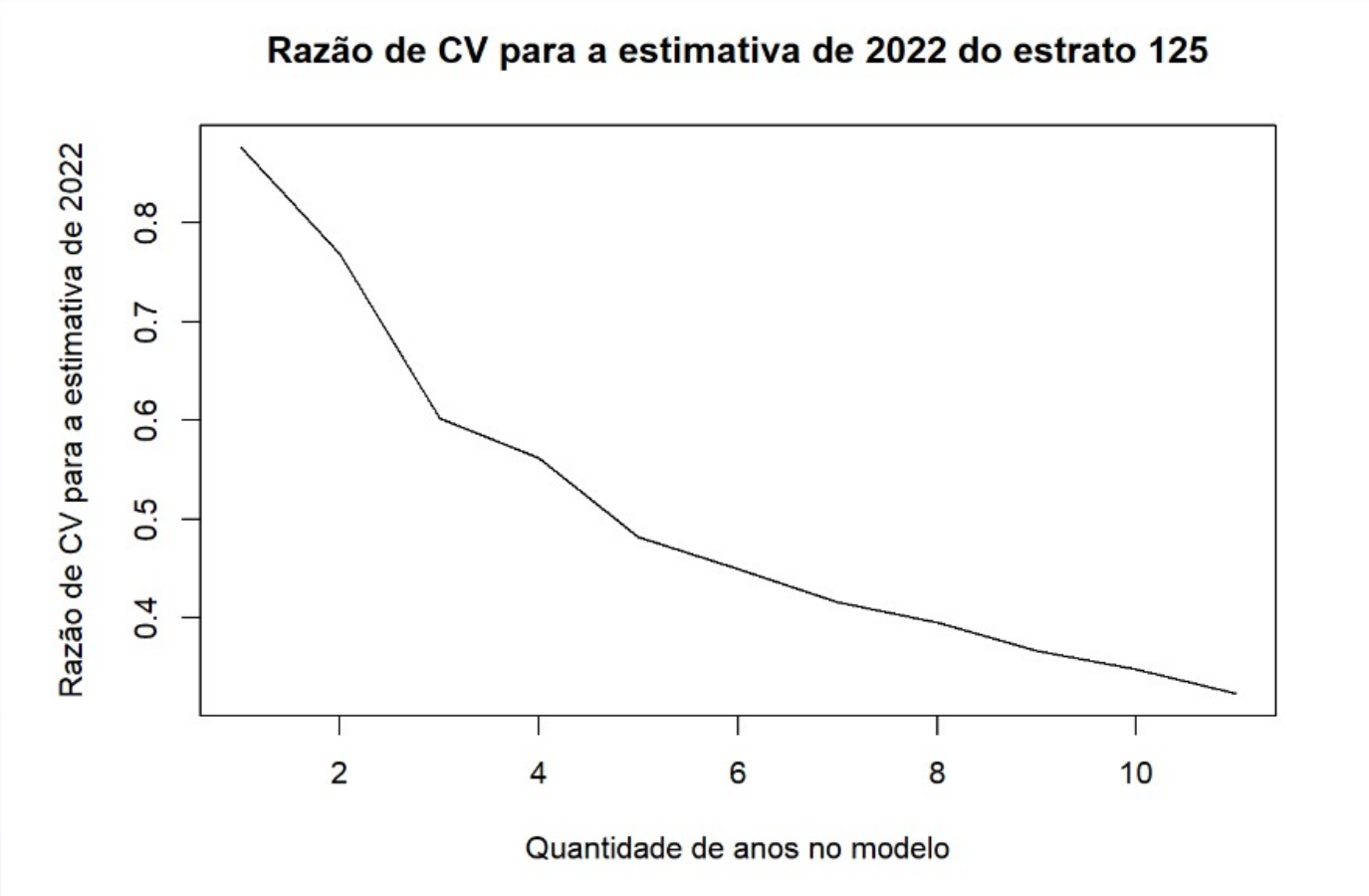

Figura: Mod3: CV com a incorporação de x anos.

Modelos a nível de unidade

Trabalho em conjunto com Malay Ghosh (University of Florida).

Modelos a nível de unidade

Trabalho em conjunto com Malay Ghosh (University of Florida).

Objetivo: propor modelo para dados de contagem que use transformação dos dados originais permitindo trabalhar com modelo Normal e homocedasticidade aproximada nos erros, reduzindo assim o custo computacional e uma camada na estimação.

Modelos a nível de unidade

- Modelos desagregados permitem analisar variáveis com mais detalhes numa população.
- Precisa de acesso aos dados da pesquisa em nível de unidade.
- Precisa de acesso a variáveis auxiliares a nível de unidade (para cada pessoa por exemplo) na população, o que pode ser difícil de obter.
- Também é possível usar covariáveis em nível de área, mas elas podem não ser tão eficazes quanto as de nível de unidade.

Modelos a nível de unidade

- Modelos desagregados permitem analisar variáveis com mais detalhes numa população.
- Precisa de acesso aos dados da pesquisa em nível de unidade.
- Precisa de acesso a variáveis auxiliares a nível de unidade (para cada pessoa por exemplo) na população, o que pode ser difícil de obter.
- Também é possível usar covariáveis em nível de área, mas elas podem não ser tão eficazes quanto as de nível de unidade.

A escolha por modelos de área ou unidade depende da aplicação e dos dados disponíveis!

Modelos a nível de unidade

Modelo de Battese Harter Fuller (BHF) (1988) é descrito da seguinte forma: para $j = 1, \ldots, N_i$, $i = 1, \ldots, m$,

$$y_{ij} = \theta_{ij} + e_{ij},$$

$$\theta_{ij} = x_{ij}^T \beta + \nu_i,$$

- $e_{ij} \sim N(0, \sigma_e^2)$,
- $\nu_i \sim N(0, \sigma_\nu^2)$,
- y_{ij} é variável resposta para unidade j na área i e x_{ij} é a variável auxiliar associada.

Modelos a nível de unidade

Modelo de Battese Harter Fuller (BHF) (1988) é descrito da seguinte forma: para $j = 1, \ldots, N_i$, $i = 1, \ldots, m$,

$$y_{ij} = \theta_{ij} + e_{ij},$$

$$\theta_{ij} = x_{ij}^T \beta + \nu_i,$$

- $e_{ij} \sim N(0, \sigma_e^2)$,
- $\nu_i \sim N(0, \sigma_\nu^2)$,
- y_{ij} é variável resposta para unidade j na área i e x_{ij} é a variável auxiliar associada.
- Objetivo é predizer $\bar{Y}_i = \frac{1}{N_i} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} y_{ij} \right\}$.

Modelos a nível de unidade

Exemplos:

- Número de dias numa semana que uma pessoa faz atividade física moderada a intensa para grupos demográficos dados pelo cruzamento de raça, etnia, faixa etária (18-29, 30-39, 40-49, 50-59 and 60-84) e gênero (2015-2016 US National Health and Nutrition Examination Survey (NHANES)).

Modelos a nível de unidade

Exemplos:

- Número de dias numa semana que uma pessoa faz atividade física moderada a intensa para grupos demográficos dados pelo cruzamento de raça, etnia, faixa etária (18-29, 30-39, 40-49, 50-59 and 60-84) e gênero (2015-2016 US National Health and Nutrition Examination Survey (NHANES)).

- Número de estudantes por escola no sexto ano que atingem performance adequada no exame *Prova Brasil* no estado de Rondônia. Rondônia é composta por 431 escolas distribuídas em 52 municípios (1 a 99 escolas por município).

Modelos de unidade para dados de contagem

Modelo Poisson é apropriado para dados de contagem: for $j = 1, \ldots, N_i$, $i = 1, \ldots, m$,

$$y_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

$$\log(\lambda_{ij}) = x_{ij}^T \beta + \nu_i,$$

where $\nu_i \sim N(0, \sigma_\nu^2)$.

Vvariance-stabilizing transformation (Anscombe, 1952)

Using the transformation $z_{ij} = y_{ij}^{1/2}$, we get:

$$\begin{aligned} z_{ij} &= \theta_{ij} + e_{ij} \\ \theta_{ij} &= x_{ij}^T \beta + \nu_i, \end{aligned}$$

where $\nu_i \sim N(0, \sigma_\nu^2)$, $\theta_{ij} = \lambda_{ij}^{1/2}$ and $e_{ij} \sim N(0, 1/4)$.

Modelos de unidade para dados de contagem

- Inferência deve considerar que objetivo é prever y_{ij} e não z_{ij}, $j \notin s$.

Modelos de unidade para dados de contagem

- Inferência deve considerar que objetivo é prever y_{ij} e não z_{ij}, $j \notin s$.

Passo-a-passo

- Considerando que $z_i^s = (z_{i1}, \ldots, z_{in_i})^T$, $z_i^{\bar{s}} = (z_{in_i+1}, \ldots, z_{iN_i})^T$, $(z_i^s, z_i^{\bar{s}})$ tem distribuição Normal multivariada condicional a β e σ_ν^2;
- A distribuição preditiva condicional $z_i^{\bar{s}} \mid z_i^s, \beta, \sigma_\nu^2$ é Normal multivariada;
- Obter média, variância e covariâncias a posteriori para $y_{ij} = z_{ij}^2$, $j = n_i + 1, \ldots, N_i, i = 1, \ldots, m$, condicional a β e σ_ν^2 (Lema 1);
- Marginalizar as quantidades obtidas no passo anterior via propriedades de esperança e variância condicional.
- Estimativas são obtidas para estas via Gibbs Sampling (permite obter numericamente as distribuições a posteriori de β e σ_ν^2).

Conclusão: A aproximação normal facilita manipulações analíticas e a transformação raiz quadrada atinge erros homocedásticos.

Referências

- Antonaci, G. (2012) Comparação de métodos para estimação de ńdices de pobreza em pequenas áreas. Dissertação mestrado, UFRJ.
- Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83, 28-36.
- Foster, J., Greer, J. e Thorbecke, E. (1984) A class of decomposable poverty measures. Econometrica, 52, p. 761-766.
- IBGE (2006) Pesquisas por amostragem: política de divulgação de estimativas com baixa precisão amostral.

Referências

- Gonçalves, K. C. M.; Ghosh, M. Unit-level model for small area estimation with count data under square root transformation. Brazilian Journal of Probability and Statistics, 36 (1) 1 - 19, 2022.
- Janicki, R. (2019) Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. Communications in Statistics - Theory and Methods, 49(9), 2264-2284.
- Neri, M. (2022) Mapa da nova pobreza. Fundação Getúlio Vargas.
- Rao, J. N. K. and Molina, I. (2015) Small area estimation. Wiley, New York.

Obrigada!

kelly@dme.ufrj.br

https://sites.google.com/dme.ufrj.br/kelly/

Obrigada!

kelly@dme.ufrj.br

https://sites.google.com/dme.ufrj.br/kelly/

- Agradecimento: CAPES e Science.