UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

PÓS-GRADUAÇÃO EM ESTATÍSTICA

ESTIMADOR SUBSEMBLE ESPACIAL PARA DADOS MASSIVOS EM GEOESTATÍSTICA

Márcia Helena Barbian

TESE DE DOUTORADO

BELO HORIZONTE
Agosto 2016

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Márcia Helena Barbian

ESTIMADOR SUBSEMBLE ESPACIAL PARA DADOS MASSIVOS EM GEOESTATÍSTICA

Trabalho apresentado ao Programa de PÓS-GRADUAÇÃO EM ESTATÍSTICA do DEPARTAMENTO DE ESTATÍSTICA da UNIVERSIDADE FEDERAL DE MINAS GERAIS como requisito parcial para obtenção do grau de Doutor em ESTATÍSTICA.

Orientador: *Prof. Dr. Renato Martins Assunção*

BELO HORIZONTE
Agosto 2016

Dedico esse trabaho à minha mãe Alice e ao meu pai Deonisio.

Agradecimentos

A caminhada foi longa e muitas foram as dificuldades, em alguns momentos a motivação diminuiu e a falta de vontade de continuar essa jornada apareceu, mas graças a muitas pessoas especiais foi possível superar essas dificuldades e terminar mais essa etapa da minha vida.

Agradeço a Deus por me guiar por esse caminho, agora que olho pra trás consigo ver como tudo se encaixa, como as oportunidades apareceram nos momentos certos e como as pessoas que fazem parte das nossas vidas estão ali por algum motivo.

Quero agradecer imensamente aos meus pais, exemplos de pessoas especiais, de companheirismo em todos os momentos. Minha mãe é muito bondosa, gentil, dedicada à familia, além disso muito determinada. Meu pai sempre forte, presente, nunca desanima. Se não fosse a dedicação e apoio de vocês, tenho certeza que não teria mais essa conquista. Quero agradecer muito aos meus irmãos, Jackson e Eduardo e à minha cunhada Janaína, por todo o apoio e paciência. Além disso, por me proporcionar o convívio com as minhas florzinhas: Carol, Luísa e Laura, que me deram muitos momentos de alegria mesmo quando tudo parecia tão complicado. Não podia deixar de agradecer a Regina, um anjo que Deus colocou em minha vida, que me ajuda muito com a minha mãe.

Agradeço muito ao Renato, pelos anos de orientação e por me motivar sempre. Exemplo de pesquisador e professor, que contribui muito para o meu enriquecimento profissional. Além disso, agradeço muito sua generosidade e grande paciência.

Aos membros da banca de qualificação e de defesa Paulo, Marcelo, Marcos, Thaís, Andrea e Alexandre, pelas valiosas sugestões para melhoria deste trabalho.

Meu grande agradecimento ao programa de Pós-Graduação em Estatística da UFMG, particularmente ao Marcelo, Glaura, Enrico, Rosângela, Denise, Bernardo, Rogéria e Rose.

À CAPES pela bolsa de doutorado e à FAPEMIG por diversos apoios financeiros prestados para participação em eventos.

Quero agradecer a todos os meus colegas do departamento de estatística da UFRGS, pelo ótimo ambiente de trabalho e por todo o apoio enquanto terminava o doutorado. Em especial a minha colega de sala Lu, pelo bom humor contagiante.

Agradeço aos amigos pelas muitos momentos de alegria. Aos colegas pelo companheirismo e por vários momentos divertidos. Em especial quero agradecer ao Rodrigo, Zaida, Paulo, Fabrícia, Wecsley, Paola, Alessandro, Thaís, Lucas, Gustavo, Marquinhos, Letícia, Jacque, Denise, Bebel, Nivia e ao pessoal do leste nova geração. A Aline por estar presente nos momentos mais complicados. A Grazi, pelo abrigo, paciência em me ouvir, obrigada por tudo guria. A Érica por ter sido tão generosa e atenciosa.

E a todos que de uma forma direta ou indireta contribuíram para a realização deste trabalho. Muito Obrigada!

Entrega o teu caminho ao Senhor; confia nele, e ele o fará. — Salmos 37:5

Resumo

Um problema que vem se tornando habitual em análise geoestatística é a quantidade crescente de observações. Em tais casos é comum que estimadores usualmente utilizados não possam ser empregados devido a dificuldades numéricas. Esta tese têm por objetivo propor um novo estimador para massivas observações em geoestatística: o estimador *subsemble* espacial. O estimador *subsemble* espacial seleciona várias subamostras, espacialmente estruturadas, do conjunto completo de dados. Cada subamostra estima com facilidade os parâmetros do modelo e as estimativas resultantes são ponderadas através de um subconjunto de validação. Em estudos simulados, compara-se a metodologia proposta com outros métodos e os resultados apresentam sua acurácia e rapidez. Além disso, uma aplicação em um banco de dados reais, com 11.000 observações, confirma essas características.

Palavras-chave: subamostragem; estatística-U; programação paralela; geoestatística; observações massivas.

Abstract

A problem that is becoming common in geostatistical analysis is the growing number of observations. In such cases, common estimators cannot be used due to numerical difficulties. This thesis proposes a new estimator for massive observations in geostatistics: the spatial subsemble estimator. The estimator selects small spatially structured subset of observations. The model parameters are estimated easily with each subsample, and the resulting estimates are weighted by a subset of validation. We compare the spatial subsemble with competing alternatives showing that it is faster and accurate. In addition, we present an application in a real database with 11000 observations.

Keywords: Large spatial data, Subsampling, U-statistics, parallel computing, geoestatistic.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Com a disseminação do computador pessoal e dos *smartphones*, a revolução computacional se faz presente no dia a dia de toda a população. O desenvolvimento rápido da tecnologia apresenta características que poucas décadas atrás eram vistas como inimagináveis. Uma dessas características é a geração de um volume imenso de informação, o que demanda a análise de conjuntos de dados caracterizados pela presença dos 4'V do *big data*: grande volume, velocidade, variedade e veracidade. Os mecanismos geradores desses dados estão presentes em todos os lugares, através de sensores, satélites, celulares e internet das coisas.

Em particular, a quantidade de observações espacialmente referenciadas aumentou drasticamente em tamanho nos últimos anos, exemplos são: imagem de satélite, modelos de monitoramento de clima e dados de meteorologia ou qualidade do ar que são observados globalmente. O NOAA (National Oceanic & Atmospheric Administration), uma organização que faz parte do Departamento de Comércio dos Estados Unidos, possui um *website* onde terabytes de dados são gerados para representar o clima, atmosfera e meteorologia do planeta.

Uma das implicações dessa evolução na análise da informação é a necessidade de estatísticos trabalharem com dificuldades numéricas e computacionais, métodos usuais tipicamente falham ou demoram muito tempo para gerar resultados. Por exemplo, um modelo geoestatístico Gaussiano no qual deseja-se calcular a predição de n estações irregularmente espaçadas, envolve a inversão de uma matriz de covariância de dimensão $n \times n$, que requer o cálculo de $O(n^3)$ operações e $O(n^2)$ de espaço em memória (Stein, 2008). Esses números aumentam rapidamente, quando $n = 1000$ a maioria dos *softwares* de estatística espacial resolvem o problema facilmente. Todavia, quando $n = 10.000$ a inversão da matriz de covariância torna-se um desafio para a maioria dos *softwares* e *hardwares*.

Em vista disso, estatísticos tentam se adaptar a essa demanda, buscando métodos para analisar grandes conjuntos de dados espaciais. Com esse intuito, é possível distinguir duas grandes classes de metodologias: a primeira adota um ponto de vista bayesiano possibilitando utilizar processos latentes com dimensão reduzida (Banerjee *et al.*, 2008; Datta *et al.*, 2016; Finley *et al.*, 2009) ou a aproximação de campos Gaussianos através de campos markovianos (Lindgren *et al.*, 2011; Rue e Tjelmeland, 2002). A segunda é mais variada:

- considerar que observações distantes de um campo Gaussiano são independentes e consequentemente igualar à zero alguns valores da matriz de covariância (Furrer *et al.*, 2006; Kaufman *et al.*, 2008);
- truncar a representação espectral (Fuentes, 2007);
- utilizar verossimilhança condicional (Stein *et al.*, 2004; Vecchia, 1988);
- trabalhar com a função escore, em que a inversa da matriz de covariância é aproximada por uma matriz esparsa (Sun e Stein, 2016);
- utilizar um conjunto de contrastes multinível (Castrillón-Candás *et al.*, 2015).

Todos esse métodos têm em comum uma característica: ignorar algum aspecto do modelo, a fim de reduzir a complexidade numérica. Uma revisão dessas técnicas pode ser encontrada em Sun *et al.* (2012).

Recentemente, problemas envolvendo *big data* para dados não espaciais têm sido abordados na literatura (Bühlmann *et al.*, 2016; Schifano *et al.*, 2016), algumas dessas técnicas utilizam o conceito *subsampling* em suas metodologias (Kleiner *et al.*, 2014; Sapp *et al.*, 2014). Em dados espaciais, Liang *et al.* (2013) desenvolve um método de estimação baseado em subamostragem, em que pequenas subamostras são sequencialmente selecionadas e os parâmetros do modelo são atualizados através de um algoritmo de aproximação estocástica (Andrieu *et al.*, 2005; Robbins e Monro, 1951). O método é consistente e, mais importante, é rápido, o que o torna uma ótima ferramenta para analisar grandes conjuntos de dados. Entretanto, essa metodologia possui alguns inconvenientes. O primeiro é a impossibilidade do uso de paralelismo devido a estrutura sequencial de estimação. O segundo é que ele depende de parâmetros de *tuning*, que são difíceis de selecionar. O terceiro, é a necessidade de checar a convergência estocástica do algoritmo.

Neste trabalho, propõem-se uma nova metodologia; o *subsemble* espacial. A proposta é de simples aplicação, computacionalmente rápida (pode-se paralelizar a estimação) e requer pouca memória. Além disso, a técnica proposta permite o cálculo de intervalos de confiança e possui boas propriedade tanto para *infill asymptotics* assim como *increasing domain asymptotics*. O método é baseado em subamostragem para pequenos subconjuntos de dados espacialmente estruturados. Em cada subamostra, os parâmetros são calculados através de algum estimador de preferência e as estimativas resultantes são ponderadas, utilizando um subconjunto de validação.

Um estudo de Monte Carlo é realizado para avaliar o comportamento do estimador *subsemble* espacial. Além disso, compara-se a proposta com o estimador de máxima verossimilhança e com o estimador RSA, proposto por Liang *et al.* (2013). Para a aplicação, utiliza-se um conjunto de 11.000 observações localizadas nos Estados Unidos, em que a variável de interesse é a quantidade de precipitação.

O trabalho terá a seguinte estrutura: no capítulo 2 o modelo geoestatístico básico é definido; no capítulo 3 apresenta-se uma breve descrição sobre importantes metodologias para estimar os parâmetros do modelo geoestatístico em grandes bancos de dados; no capítulo 4 o estimador *subsemble* espacial será apresentado; no capítulo 5 as características assintóticas do estimador proposto; no capítulo 6 simulações; e no 7 uma aplicação a dados reais.

Capítulo 2

Modelo Geoestatístico

2.1 O modelo Geoestatístico Gaussiano

Suponha que o vetor $\mathcal{Y} \equiv (Y(s_1), Y(s_2), \ldots, Y(s_n))^t$ são n valores observados de um processo aleatório $\{Y(s) : s \in D \subset \mathbb{R}^2\}$, onde o índice espacial s varia continuamente através da região D. Seja o modelo:

$$Y(s_i) = \mu(s_i) + X(s_i) + \epsilon(s_i), \qquad i = 1, \ldots, n \tag{2.1}$$

onde $\mu(s_i) = \mathbb{E}[Y(s_i)]$ e $(\epsilon(s_1), \epsilon(s_2), \ldots, \epsilon(s_n))$ são variáveis aleatórias independentes e identicamente distribuídas com função de densidade normal de média igual à zero e variância τ^2, também conhecido como efeito pepita. As variáveis aleatórias $X(s_1), X(s_2), \ldots, X(s_n)$ originam-se de um processo Gaussiano $\{X(s) : s \in D \subset \mathbb{R}^2\}$ com $\mathbb{E}[X(s)] = 0$ e matriz de covariância igual a $\sigma^2\mathbf{R}$. A matriz $\mathbf{R}$ de dimensão $n \times n$ possui elementos $R_{ij} = \rho(||s_i - s_j||; \boldsymbol{\lambda})$, representando a correlação entre $X(s_i)$ e $X(s_j)$. A distância Euclidiana entre as observações s_i e s_j é $||s_i - s_j||$ e $\boldsymbol{\lambda}$ são os parâmetros da função de correlação. Nesse modelo, $\{Y(s)\}$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu} = \mathbf{1}\mu$ e matriz de covariância $\boldsymbol{\Sigma} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}$, em que $\mathbf{I}$ é uma matriz identidade $n \times n$ e $\mathbf{1}$ é um vetor n dimensional composto de 1s.

Pode-se generalizar ainda mais o modelo acima, permitindo que a média não seja constante,

$$\mu(s_i) = \beta_0 + \sum_{j=1}^{c} \beta_j \mathbf{v}_j(s_i), \tag{2.2}$$

$\mathbf{v}_j(s_i)$ denota a j-ésima variável exploratória observada no ponto s_i e β_j é o correspondente coeficiente de regressão (Cressie, 2015).

Seja $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p, \tau^2, \sigma^2, \boldsymbol{\lambda})$ o vetor de parâmetros do modelo (2.1), a função de verossimilhança é dada por:

$$L(\boldsymbol{\theta}, \mathbf{y}) = -\frac{n\log(2\pi)}{2} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \tag{2.3}$$

Para obter o estimador de máxima verossimilhança é necessário calcular o inverso da matriz de covariância $\boldsymbol{\Sigma}$ e o seu determinante $|\boldsymbol{\Sigma}|$. Esse processo exige uma complexidade computacional da ordem $O(n^3)$, tornando-se um problema quando o número de observações é massivo, pois o tempo necessário para obter os resultados é proibitivo e requer uma grande quantidade de memória RAM.

2.2 Cálculo da Variância do Estimador de Máxima Verossimilhança

Para estimar a variância do EMV no caso espacial é necessário levar em consideração as localizações das observações $\{Y(s)\}$, o que dificulta a obtenção de propriedades teóricas. Por exemplo, mesmo quando os pontos são distribuídos conforme um *lattice*, ainda não foi possível encontrar uma

forma fechada para a matriz de informação de Fisher dos estimadores de máxima verossimilhança de funções de covariância comumente utilizadas, como a família Matérn ou como caso particular a função de covariância exponencial ($\sigma^2\rho(||\mathbf{s}_i - \mathbf{s}_j||;\boldsymbol{\lambda}) = \sigma^2 \exp^{-||\mathbf{s}_i-\mathbf{s}_j||/\phi}$).

Além disso, em análise geoestatística existem duas visões assintóticas: *increasing domain asymptotics*, na qual mais dados são coletados por aumentar o domínio e o *infill assymptotics*, em que os dados são coletados mais densamente em um domínio fixo. Propriedades assintóticas dos estimadores são diferentes nos dois domínios.

Para o caso de *increasing domain asymptotics*, os autores Mardia e Marshall (1984) provam a consistência e normalidade assintótica do estimador de máxima verossimilhança para dados regularmente espaçados de campos Gaussianos. Condicionado a algumas suposições, a forma funcional da matriz de informação de Fisher assintótica dos EMV é:

$$IF(\boldsymbol{\beta}, \boldsymbol{\varrho}; \mathbf{Y}) = diag(G_{\boldsymbol{\beta}}, G_{\boldsymbol{\varrho}}), \tag{2.4}$$

em que $G_{\boldsymbol{\beta}} = \mathbf{V}^t\boldsymbol{\Sigma}\mathbf{V}$, $\boldsymbol{\varrho} = (\sigma^2, \tau^2, \boldsymbol{\lambda})$ e o (i,j)-ésimo elemento de $G_{\boldsymbol{\varrho}}$ é dado por

$$G_{\boldsymbol{\varrho}}(i,j) = \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varrho_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varrho_j}\right) = \frac{1}{2}tr\left(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\varrho_i}\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\varrho_j}\right). \tag{2.5}$$

No domínio da frequência, Guyon (1982) sugere utilizar a aproximação de Whittle (1954) para o logaritmo da função de verossimilhança. Em seu artigo, ele demonstra que os estimadores de máxima verossimilhança baseados nessa metodologia são consistentes. Uma extensão dessa aproximação para observações irregularmente espaçadas é proposta por Fuentes (2007).

Ying (1993) demonstrou a normalidade assintótica dos estimadores de máxima verossimilhança de uma função de covariância exponencial multiplicativa e Van Der Vaart (1996) complementou o seu trabalho, demonstrando a eficiência assintótica desses estimadores.

Em uma situação muito específica, Loh (2005) conseguiu determinar de forma fechada a função de verossimilhança e consequentemente derivar sua matriz de informação de Fisher. Para isso, o autor considera uma subclasse de campos aleatórios Gaussianos do tipo Matérn, propostos por Stein *et al.* (1989).

Para dados selecionados no *infill asymptotics*, o EMV dos parâmetros da função de covariância não podem ser consistentemente estimados. Como alternativa, Zhang (2004) mostrou que a razão dos parâmetros da função de covariância Matérn é consistente, por exemplo (ϕ/σ^2). Sobre a distribuição assintótica dessa razão, o mesmo autor indica através de simulações que o estimador pode ser assintoticamente normal. Entretanto, ainda não foi possível provar matematicamente essa propriedade.

Capítulo 3

Estado da Arte

Neste capítulo a descrição de algumas propostas para lidar com grandes bancos de dados em geoestatística será apresentada, as metodologias de *tapering*, verossimilhança condicional e SPDE serão descritas brevemente. Um maior enfoque será dado para a técnica baseada em reamostragem de Liang *et al.* (2013), pois é um método que se assemelha a proposta desse trabalho, visto que utiliza subamostras no cálculo do estimador.Dada essa característica comum entre as duas metodologias, acredita-se que a comparação entre o estimador *subsemble* espacial e o método proposto por Liang *et al.* (2013) seja mais adequada, por isso no capítulo 6 as simulações serão calculadas levando em consideração essas duas alternativas para lidar com *big data* em geoestatística.

3.1 *Covariance Tapering*

Kaufman *et al.* (2008) apresenta duas aproximações para a função de verossimilhança utilizando o método conhecido como *tapering*. A ideia do *tapering* é assumir que pares distantes de observações podem ser considerados independentes. A matriz de covariância assume valores iguais a 0 para grandes distâncias, quando a correlação entre as observações é muito baixa. A nova matriz de covariância é esparsa, o que permite a utilização de algoritmos específicos, reduzindo o tempo de processamento para maximizar a verossimilhança.

Seja $\mathbf{Y}(\mathbf{s})$ observações de um campo aleatório Gaussiano, estacionário e isotrópico. A função de correlação $T(\gamma)_{ij} = \rho_{taper}(||s_i - s_j||; \boldsymbol{\lambda}; \gamma)$ assume valor 0 quando a distância entre as observações $||s_i - s_j||$ é maior que um limiar γ, gerando a matriz de covariância *tapering* $\boldsymbol{\Sigma} \circ T(\gamma)$, em que $\circ$ é o produto de Shur. O artigo propõe substituir a matriz de covariância $\boldsymbol{\Sigma}$ da verossimilhança 2.3 por $\boldsymbol{\Sigma} \circ T(\gamma)$.

Pequenos valores de γ tornam a matriz de covariância mais esparsa, quando $\gamma = 0$ as observações são tratadas como independentes e quando $\gamma \to \infty$ é o mesmo que a função de verossimilhança original. Logo, a escolha do valor de γ é crucial na estimativa dos parâmetros e no tempo de processamento. Quanto maior o valor de γ, maior demanda computacional; quanto menor o valor de γ, mais esparsa é a matriz e mais rápido o cálculo dos resultados.

Uma grande desvantagem é que o método só considera os pontos mais próximos, o que pode ser uma alternativa ruim quando as observações apresentam uma dependência de maior alcance.

3.2 Verossimilhança Condicional

Sabe-se que a densidade conjunta do modelo 2.1 pode ser escrito como o produto de densidades condicionais. A idéia do artigo de Vecchia (1988) é que a densidade conjunta pode ser descrita como o produto de densidades condicionais baseadas em algum ordenamento das observações, além disso, para diminuir o tempo de processamento, as distribuições condicionais dependem somente de $m << n$ observações vizinhas de y_i. Com esse objetivo, uma sequência $L_m(\boldsymbol{\theta}, \mathbf{y})$ é definida, em que $L_m(\boldsymbol{\theta}, \mathbf{y})$ se aproxima de $L(\boldsymbol{\theta}, \mathbf{y})$ quando m se aproxima de n.

A função de verossimilhança é dada por

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} f(y_i | y_j, 1 \leq j \leq i-1)$$

$f(y_i|\cdot)$ denota a função densidade de probabilidade condicional de y_i. Se o tamanho da amostra é grande, espera-se que para grandes valores de i, as condicionais contenham informação redundante. Isso significa que i-ésima densidade condicional da equação pode ser aproximadamente equivalente a $f(y_i|\mathbf{y}_{im})$, em que $\mathbf{y}_{im}$ é um vetor composto por algumas observações entre $\{y_j, 1 \leq j \leq m\}$.

Uma maneira de decidir quais observações devem constituir o vetor $\mathbf{y}_{im}$ é selecionar observações amostrais que estão mais próximas de y_i, no sentido da distância euclidiana, isto é, os vizinhos mais próximos. A equação da verossimilhança aproximada de ordem m é

$$L_m(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} f(y_i | \mathbf{y}_{im})$$

em que $\mathbf{y}_{im}$ é um vetor consistindo de $\min(i-1, m)$ observações. Se $i - 1 > m$, seleciona-se as m observações que são mais próximas de y_i. Quando $i - 1 < m$, seleciona-se o vetor $y_1, \ldots, y_{i-1}$.

Uma dificuldade na aplicação dessa metodologia é decidir o tamanho do vetor $\mathbf{y}_{im}$, isto é, o número de vizinhos m considerados na distribuição condicional. Ademais, Stein *et al.* (2004) afirmam que o método perde informação quando estima os parâmetros da função de covariância, posto que só considera os pontos mais próximos. Por isso, Stein *et al.* (2004) adaptaram a proposta de Vecchia (1988) para aproximar a verossimilhança restrita de observações de um processo Gaussiano condionando observações próximas e distantes de y_i. Além disso, sugerem um método para verificar a eficiência dessa aproximação.

3.3 Equações Diferenciais Parciais Estocásticas

O principal resultado do estudo de Lindgren *et al.* (2011) é que pode ser construído um link entre campos Gaussianos e Campos Aleatórios Markovianos Gaussianos (CAMG) através de uma aproximação de equações diferencias parciais estocásticas. Esse método, também conhecido como SPDE, permite que algoritmos de matrizes esparsas sejam aplicados, o que torna o processo de estimação muito mais rápido. Essa metodologia tem sido utilizada principalmente no contexto bayesiano, em que além da economia de tempo devida à utilização de campos Markovianos, a distribuição à *posteriori* pode ser calculada de forma ainda mais otimizada, por meio da aproximação integrada de Laplace (INLA).

A representação de um CAMG pode ser construída explicitamente por meio de uma determinada SPDE com ruído Gaussiano. Essas equações diferenciais têm como solução um campo Gaussiano com função de covariância de Matérn. O resultado é uma aproximação do campo Gaussiano por meio da triangulação geral da área D, onde essas áreas triangulares geram um CAMG.

O lado negativo dessa nova metologia é que a implementação do modelo é muito complexa. O pré-processamento para criação das SDPE, triangulações e representações do CAMG não são tarefas triviais.

3.4 Aproximação Estocástica Baseada em Reamostragem

Os autores Liang *et al.* (2013) propõem estimar os parâmetros do modelo 2.1 através de um algoritmo de aproximação estocástica (RSA), proposto por Robbins e Monro (1951), em que para cada iteração o algoritmo seleciona uma pequena subamostra m do conjunto de dados $\mathcal{Y}$. Como o estimador utiliza somente uma pequena proporção dos dados em cada passo do algoritmo, ele evita a inversão de grandes matrizes de covariância, tornando-se uma ótima ferramenta para grandes bancos de dados. Além disso, os autores provam propriedades importantes do estimador.

Seja $\mathbf{Z}$ uma amostra aleatória de m observações selecionada aleatoriamente do vetor $\mathcal{Y}$, o qual possui tamanho n. Para estimar $\boldsymbol{\theta}$ utiliza-se a divergência de Kullback-Leibler,

$$KL(f_{\boldsymbol{\theta}}, g) = -\int\int \log\left(\frac{f_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{s})}{g(\mathbf{z}|\mathbf{s})}\right) g(\mathbf{z}|\mathbf{s})g(\mathbf{s})dzds,$$

em que $f_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{s})$ é a densidade da normal multivariada, $g(\mathbf{z}|\mathbf{s})$ denota a verdadeira função densidade geradora dos dados e $g(\mathbf{s})$ é a distribuição de $\mathbf{s}$. Utilizando subamostras selecionadas aleatoriamente de $\mathcal{Y}$ a divergência de Kullback-Leibler pode ser aproximada por

$$\hat{KL}(f_{\boldsymbol{\theta}}, g|\mathcal{Y}) = C - \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \log f_{\boldsymbol{\theta}}(z_i|s_i),$$

em que C denota uma constante relacionada com a entropia de $g(\mathbf{z}, \mathbf{s})$. Para atualizar as estimativas de $\boldsymbol{\theta}$, o algoritmo de aproximação estocástica utiliza o sistema de equações

$$\frac{\partial \hat{KL}(f_{\boldsymbol{\theta}}, g|\mathcal{Y})}{\partial \boldsymbol{\theta}} = -\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} H(\boldsymbol{\theta}, z_i, s_i) = 0, \tag{3.1}$$

A função $H(\boldsymbol{\theta}, z_i, s_i)$ é a derivada de primeira ordem do $\log f_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{s})$ com respeito a $\boldsymbol{\theta}$, o vetor (z_i, s_i) denota uma amostra aleatória selecionada de $\mathcal{Y}$.

Como exemplo, para a função de covariância exponencial os respectivos componentes de $H(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s})$ da equação (3.1) são dados por

$$\begin{aligned}
H_{\beta_0}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}) &= \mathbf{1}_m^t \boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \\
H_{\beta_i}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}) &= v_i^t \boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \\
H_{\phi}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}) &= -\frac{1}{2} tr\left(\boldsymbol{\Sigma}_z^{-1}\sigma^2 \frac{\partial \mathbf{R}_z}{\partial \phi}\right) + \sigma^2(\mathbf{z} - \boldsymbol{\mu}_z)^t \boldsymbol{\Sigma}_z^{-1} \frac{\partial \mathbf{R}_z}{\partial \phi} \boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \\
H_{\sigma^2}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}) &= -\frac{1}{2} tr(\boldsymbol{\Sigma}_z^{-1}\mathbf{R}_z) + \frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)^t \boldsymbol{\Sigma}_z^{-1}\mathbf{R}_z\boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \\
H_{\tau^2}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}) &= -\frac{1}{2} tr(\boldsymbol{\Sigma}_z^{-1}) + \frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)\boldsymbol{\Sigma}_z^{-2}(\mathbf{z} - \boldsymbol{\mu}_z),
\end{aligned}$$

em que $\partial \mathbf{R}_z / \partial \phi$ é uma matriz $m \times m$ com o elemento (i, j) dado por $||s_i - s_j||/\phi^2 \exp^{-||s_i - s_j||/\phi}$.

Outra característica importante da metodologia RSA é a definição do espaço paramétrico. Inicialmente o algoritmo permite que as estimativas de $\boldsymbol{\theta}$ sejam avaliadas em pequenos intervalos, em cada passo do algoritmo esses intervalos podem ou não aumentar. Conforme as sequencias de estimativas são obtidas, espera-se que o espaço paramétrico convirja para intervalos que contenham os verdadeiros valores dos parâmetros. Essa modificação do algoritmo de Robbins e Monro (1951) foi proposta por Andrieu *et al.* (2005) e pode ser definida da seguinte forma: seja Θ o espaço paramétrico de $\boldsymbol{\theta}$, além disso, considere que $\{\mathcal{K}_s, s \geq 0\}$ é uma sequência de conjuntos compactos de Θ, tal que

$$\bigcup_{s \geq 0} \mathcal{K}_s = \Theta, \qquad \text{e} \qquad \mathcal{K}_s \subset int(\mathcal{K}_{s+1}), \qquad s \geq 0,$$

em que $int(A)$ denota o interior do conjunto A. Seja π_t o número de truncações realizadas até a iteração t, o algoritmo é iniciado no espaço paramétrico $\mathcal{K}_0$, esse espaço aumenta dado o valor de π_t.

O algoritmo do estimador RSA pode ser explicado da seguinte maneira: o valor inicial $\boldsymbol{\theta}_0$ é definido de forma aleatória tal que $\boldsymbol{\theta}_0 \in \mathcal{K}_0$, ademais, $\pi_0 = 0$. Dados esses valores iniciais, o processo de iteração possui os seguintes passos:

1. Selecione aleatoriamente e sem reposição a subamostra $(\mathbf{Z}_{t+1}, \mathbf{S}_{t+1})$ do vetor $\{Y(s_1), \ldots, Y(s_n)\}$.

2. Atualize cada componente de $\boldsymbol{\theta}_t$ por meio das seguintes equações

$$\begin{aligned}
\beta_0^{t+1/2} &= \beta_0^t + a_{t+1} H_{\beta_0}(\boldsymbol{\theta}_t, \mathbf{Z}_{t+1}, \mathbf{S}_{t+1}) \\
\beta_i^{t+1/2} &= \beta_i^t + a_{t+1} H_{\beta_i}(\boldsymbol{\theta}_t, \mathbf{Z}_{t+1}, \mathbf{S}_{t+1}) \\
\phi^{t+1/2} &= \phi^t + a_{t+1} H_{\phi}(\boldsymbol{\theta}_t, \mathbf{Z}_{t+1}, \mathbf{S}_{t+1}) \\
(\sigma^2)^{t+1/2} &= (\sigma^2)^t + a_{t+1} H_{\sigma^2}(\boldsymbol{\theta}_t, \mathbf{Z}_{t+1}\mathbf{S}_{t+1}), \\
(\tau^2)^{t+1/2} &= (\tau^2)^t + a_{t+1} H_{\sigma^2}(\boldsymbol{\theta}_t, \mathbf{Z}_{t+1}\mathbf{S}_{t+1}),
\end{aligned}$$

3. Se $||\boldsymbol{\theta}_{t+\frac{1}{2}} - \boldsymbol{\theta}_t|| \leq b_t$ e $\boldsymbol{\theta}_{t+\frac{1}{2}} \in \mathcal{K}_{\pi_k}$ então $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t+\frac{1}{2}}$ e $\pi_{t+1} = \pi_t$; caso contrário $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$ e $\pi_{t+1} = \pi_t + 1$. A notação $||\cdot||$ denota a norma euclidiana de um vetor

Observe que para o cálculo do algoritmo é necessário definir dois valores de *input*. O b_t é utilizado para definir a taxa de aceitação do algoritmo. O segundo valor, a_t, é utilizado nas equações de estimação, indicando a diferença entre as atualizações. A escolha desses parâmetros de otimização é uma das grandes desvantagens do estimador RSA. Outra desvantagem, é que é necessário verificar a convergência do algoritmo.

Capítulo 4

O Estimador *Subsemble* Espacial

4.1 A Idéia do Estimador *Subsemble* Espacial

A estratégia de *divide and conquer* adaptada para o contexto espacial é intuitivamente simples e pode ser explicada por meio da Figura 4.1. Os pontos mostrados no quadrado formam uma grande coleção de n observações, os pontos em laranja indicam as estações que fazem parte de uma subamostra espacialmente estruturada de tamanho $m = jk$, composta por $j = 4$ centros de $k = 12$ estações próximas. Essas estações são utilizadas na estimação do vetor de parâmetros $\boldsymbol{\theta}$, resultando em $\hat{\boldsymbol{\theta}}_i$.

O próximo passo é o cálculo do peso w_i associado à estimativa $\hat{\boldsymbol{\theta}}_i$, para tal é necessário escolher um conjunto de dados de validação. Primeiramente propõem-se selecionar aleatoriamente uma observação e seus k' vizinhos mais próximos, depois dividir aleatoriamente essas observações em dois grupos, representados pelos dados coloridos em azul ($\mathbf{Z}^v$) e vermelho ($\mathbf{Z}^p$). Utilizando os pontos azuis $\mathbf{Z}^v$ e $\hat{\boldsymbol{\theta}}_i$ calcula-se a predição $\hat{\mathbf{Z}}^p$ das observações em vermelho. O erro quadrático médio de predição $(\mathbf{Z}^p - \hat{\mathbf{Z}}^p)^t(\mathbf{Z}^p - \hat{\mathbf{Z}}^p)$ fornece uma medida de qualidade do estimador $\hat{\boldsymbol{\theta}}_i$ e a ponderação w_i é inversamente proporcional à essa medida.

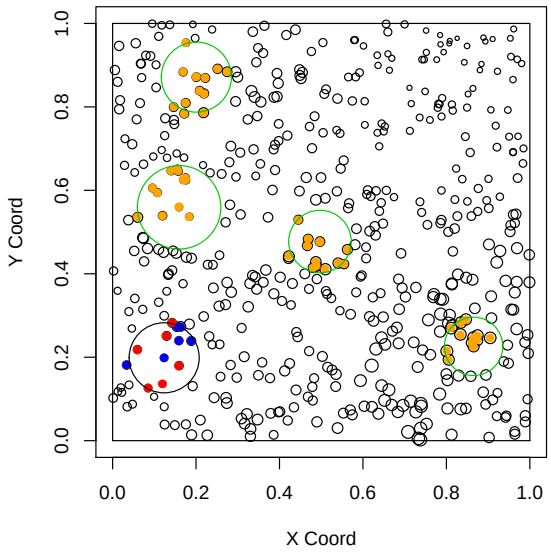

Figura 4.1: *Exemplo de subamostra de estimação e de validação do estimador subsemble espacial: pontos laranjas representam uma subamostra de $m = 48$ observações, em que $j = 4$ e $k = 12$, pontos azuis representam o subconjunto de validação e pontos em vermelho o subconjunto de predição.*

Para compensar o pequeno tamanho da subamostra, o estimador *subsemble* espacial repete

independentemente este procedimento B vezes, as B estimativas $\hat{\boldsymbol{\theta}}_i$ são combinados por meio da média dos pesos w_i ou pela média dos B valores de $\hat{\boldsymbol{\theta}}_i$.

4.2 Definição do Estimador Subsemble Espacial

Seja $\mathbf{Z}(\mathbf{s}) = (Y(s_1^*), Y(s_2^*), \ldots, Y(s_m^*))^T$ uma subamostra de tamanho $m << n$ do conjunto de dados $\mathcal{Y} = \{Y(s_1), Y(s_2), \ldots, Y(s_n)\}$, dado $\mathbf{s} = (s_1^*, s_2^*, \ldots, s_m^*)$, o vetor $\mathbf{Z}$ tem o mesmo modelo subjacente de $\mathbf{Y}$:

$$\mathbf{Z}|\mathbf{S} \sim N_m(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \tag{4.1}$$

em que $\boldsymbol{\Sigma}_z = \sigma^2\mathbf{R}_z + \tau^2\mathbf{I}_m$, $\mathbf{R}_z$ é uma matriz de correlação de dimensão $m \times m$ de $\mathbf{Z}(\mathbf{s})$ e $\boldsymbol{\mu}_z = (\mu(s_1^*), \mu(s_2^*), \ldots, \mu(s_m^*))^T$ é seu valor esperado.

Com o objetivo de calcular os parâmetros do modelo (2.1), os autores Liang *et al.* (2013) uilizaram uma aproximação para a divergência de Kullback-Leibler, onde sua minimização leva a uma equação baseada em uma estatística-U definida como

$$\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \frac{\partial \log f_\theta(z_i|s_i)}{\partial\theta} = 0,$$

em que $f_\theta(z_i|s_i)$ é a densidade da distribuição normal multivariada do modelo (4.1) baseada em uma subamostra de tamanho m. Esta soma leva em consideração todas as subamostras de tamanho m das n observações, contudo ao invés de utilizar todas as subamostras, Liang *et al.* (2013) utilizaram o algoritmo de aproximação estocástica de Robins-Moro (Robbins e Monro, 1951). O método amostra sequencialmente subconjuntos de tamanho m e atualiza as estimativas do vetor de parâmetros, depois de várias atualizações espera-se que as estimativas tenham convergido para o verdadeiro vetor de parâmetros. Apesar da idéia dos autores ser muito interessante, o método apresenta algumas desvantagens, uma delas é definir os valores iniciais do algoritmo de maximização. Outra importante desvantagem é que as m estações são escolhidas por meio de uma amostragem aleatória simples de $\mathcal{Y}$, portanto, a subamostra não possui qualquer estrutura espacial. Isso resulta em menor informação sobre a dependência de observações adjacentes, especialmente para localizações muito próximas.

Como alternativa, propõem-se selecionar B vezes, independentemente, uma subamostra espacialmente estruturada de m observações, em seguida, calcular as estimativas associadas a cada uma das subamostras. Dado que em cada subamostra os parâmetros são estimados através do EMV, seja a i-ésima subamostra $\mathbf{Z}(\mathbf{s}_i) = (Y(s_{i1}^*), Y(s_{i2}^*), \ldots, Y(s_{im}^*))^T$ nas localizações $\mathbf{s}_i = (s_{i1}^*, \ldots, s_{im}^*)$, o i-ésimo estimador de máxima verossimilhança é definido como

$$\hat{\boldsymbol{\theta}}_i = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}_i|\mathbf{s}_i). \tag{4.2}$$

Esse processo resulta em B estimativas para cada um dos parâmetros do modelo, que devem ser combinadas de alguma forma para gerar o estimador. Uma possibilidade é ponderar os B resultados, através de pesos associados à qualidade de predição da estimativa $\hat{\boldsymbol{\theta}}_i$, por exemplo, o erro quadrático médio. Como o objetivo é aplicar a metodologia quando existe um número muito grande de observações, não seria interessante calcular essa medida considerando a amostra completa, visto que o esforço computacional seria elevado. Por esse motivo, propõem-se novamente selecionar uma subamostra dos dados e calcular o erro de predição para esse subgrupo. Essas observações são denominadas subamostra de validação e são selecionadas da seguinte forma: sorteia-se um ponto central, em seguida seleciona-se k' vizinhos mais próximos ao ponto central. As $k'+1$ estações são aleatoriamente divididas em dois subconjuntos, um de validação $\mathbf{Z}^v$ e outro de predição $\mathbf{Z}^p$. O subconjunto de validação calcula o melhor preditor linear não viciado (BLUP) para as localizações do vetor $\mathbf{Z}^p$:

$$\widehat{\mathbf{Z}^p} = \hat{\boldsymbol{\mu}}_{z^p} + \mathbf{r}'\boldsymbol{\Sigma}_{z^v}\left(\mathbf{Z}^v - \hat{\boldsymbol{\mu}}_{z^v}\right),$$

em que $\mathbf{r}$ é a covariância entre $\mathbf{Z}^v$ e $\mathbf{Z}^p$ é estimado por $\hat{\boldsymbol{\theta}}_i$. Os pesos da qualidade de predição são dados por

$$w_i^{-1} = ||\mathbf{Z}^p - \widehat{\mathbf{Z}^p}||^2.$$

Seja $i = 1, \ldots, B$ o índice da subamostra e $\mathbf{s}_i$ as estações que compõem a subamostra i, o estimador *subsemble* espacial é dado por

$$\hat{\boldsymbol{\theta}} = \frac{\sum_i w_i \hat{\boldsymbol{\theta}}_i}{\sum_i w_i}, \tag{4.3}$$

uma alternativa simplificada é calcular a média das estimativas $\hat{\boldsymbol{\theta}}_i$, o que equivale a considerar o mesmo peso para cada uma das B subamostras,

$$\tilde{\boldsymbol{\theta}} = \frac{\sum_i \hat{\boldsymbol{\theta}}_i}{B}, \tag{4.4}$$

nesse caso, como não é necessário calcular os pesos w_i, a subamostra de validação não é utilizada no cálculo do estimador $\tilde{\boldsymbol{\theta}}$.

Outro passo importante no cálculo do estimador *subsemble* é a forma de seleção da subamostra que será utilizada na estimação de $\hat{\boldsymbol{\theta}}_i$. Como o objetivo é estimar $\boldsymbol{\theta}$ tão bem quanto possível dado o tamanho da subamostra $m = jk$, seria interessante levar em consideração a relação espacial entre as observações. Para gerar melhores estimativas para os parâmetros da função de covariância é necessário combinar observações próximas e distantes, onde as k observações localizadas próximas umas às outras fornecem informação sobre a correlação de curto alcance $\rho(\cdot; \boldsymbol{\lambda})$; as estações que estão afastadas permitem estimar o *range* do modelo.

Por meio da teoria de estatísticas-U e *subsampling* é possível obter propriedades teóricas de $\tilde{\boldsymbol{\theta}}$, propriedades equivalentes para $\hat{\boldsymbol{\theta}}$ são muito mais difíceis de provar, devido à presença de pesos estocásticos em (4.3). Todavia, em estudos simulados, foi possível observar que $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ possuem desempenhos semelhantes.

Os estimadores $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ são exemplos da estratégia conhecida como *divide and conquer* (Guha *et al.*, 2012), no qual o problema é dividido em muitos pedaços, onde a estimação é uma tarefa que pode ser facilmente resolvida. Após essa etapa, os resultados desses pequenos pedaços são combinados, produzindo uma única estimativa. Uma das principais vantagens dessa estratégia é a possibilidade de processar cada pedaço separadamente (paralelismo), tornando possível o uso de programação *multithreading* (Tanenbaum, 2009). Cada *thread* pode ser enviado e estimado separadamente por um *core* do CPU, logo, quanto maior a quantidade de *cores*, mais rápido é o cálculo da estimativa. Outra vantagem é o cálculo de intervalos de confiança.

4.3 Cálculo da Variância do Estimador *Subsemble* Espacial

Através da matriz de informação de Fisher assintótica pode-se mensurar, aproximadamente, o valor da variância do EMV. Propõem-se calcular $Var(\hat{\boldsymbol{\theta}}_i)$ para cada respectiva subamostra i, por meio da equação (2.4). Feito isso, as B variâncias amostrais são combinadas para formar as variâncias dos estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$. Como estamos preocupados com dados que apresentam massivas observações, a covariância entre os diferentes $\hat{\boldsymbol{\theta}}_i$ será desconsiderada.

A variância dos estimadores *subsemble* espacial são aproximadas pelas seguintes equações:

$$Var(\hat{\boldsymbol{\theta}}) = \frac{\sum_i w_i^2 \mathcal{I}(\hat{\boldsymbol{\theta}}_i)}{\left(\sum_i w_i\right)^2} \tag{4.5}$$

e

$$Var(\tilde{\boldsymbol{\theta}}) = \frac{\sum_i \mathcal{I}(\hat{\boldsymbol{\theta}}_i)}{B}, \tag{4.6}$$

onde $\mathcal{I}(\hat{\boldsymbol{\theta}}_i)$ é a aproximação da matriz de covariância de $\hat{\boldsymbol{\theta}}_i$, dada pela inversa da equação (2.4).

4.4 Subamostra m

A acurácia das estimativas baseadas em subamostras dependem crucialmente da escolha do tamanho da subamostra, logo, faz-se necessário algum critério que auxilia nessa decisão. Alguns trabalhos na área de reamostragem espacial apontam que o tamanho da subamostra ótima depende de características do campo aleatório subjacente, tal como força de dependência espacial, geometria da região, além da geometria da subamostra (Nordman e Lahiri, 2004; Nordman *et al.*, 2007; Politis e Sherman, 2001; Politis *et al.*, 1999b; Sherman, 1996). Apesar de haver referências sobre o assunto, nenhum desses métodos apresenta uma maneira simples de fazer tal escolha. Com o objetivo de tornar a metodologia mais acessível, propõem-se determinar o tamanho da subamostra m, através da utilização de um *trade-off* entre o tempo e a qualidade das estimativas.

Para mensurar o tempo utiliza-se a seguinte lógica: sabe-se que o número de operações necessárias para calcular o EMV é da $O(n^3)$ e o tempo para seu cálculo é proporcional à quantidade de operações. Consequentemente, definido o valor de B e m é possível quantificar, aproximadamente, o tempo de espera para gerar os resultados. A qualidade das estimativas é calculada a partir da aproximação da matriz de informação de Fisher assintótica ($\mathcal{I}(\hat{\boldsymbol{\theta}}_i)$). Através da matriz $\mathcal{I}(\hat{\boldsymbol{\theta}}_i)$ é possível mensurar quanto da variância do estimador de cada parâmetro aumenta dado o tamanho de uma particular subamostra. Dessa forma, torna-se possível medir a acurácia do $\hat{\boldsymbol{\theta}}_i$.

A utilização do *trade-of* entre tempo e qualidade da estimativa na escolha do tamanho m da subamostra pode ser exemplificada pelas Figuras 4.2 e 4.4. Os pontos pretos da Figura 4.2 representam uma amostra de tamanho $n = 2000$ de um campo gaussiano de média 0 e função de covariância exponencial de parâmetro $\phi = 25$ e $\sigma^2 = 1$. Primeiro, seleciona-se aleatoriamente os j centros que farão parte da subamostra. Dados os centros, considere os k vizinhos mais próximos, essas jk observações (pontos vermelhos) representam uma subamostra dos dados. O segundo passo é calcular a estimativa $\mathcal{I}(\hat{\boldsymbol{\theta}}_i)$ para a subamostra escolhida. O terceiro passo é repetir esse procedimento para valores crescentes de k, mantendo os centros fixos. Na Figura 4.2, ilustra-se diferentes subamostras para os mesmos $j = 5$ pontos centrais, em que $k = 19$ (esquerda), $k = 59$ (centro) e $k = 99$ (direita).

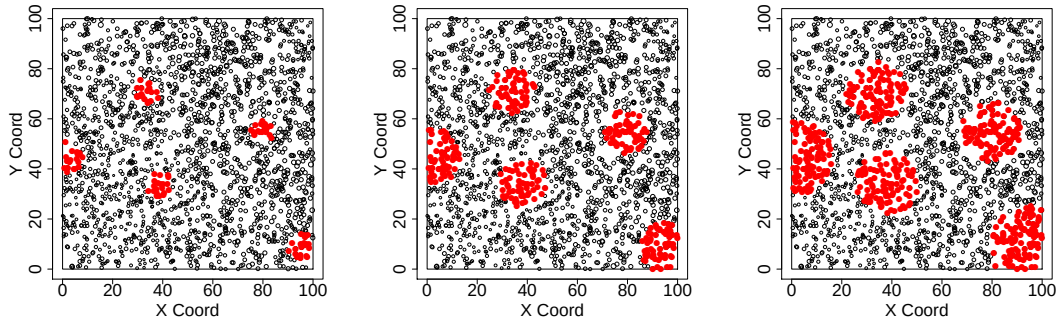

Figura 4.2: *Exemplo de subamostras de tamanho* $m = \{100, 300, 500\}$ *considerando os mesmos pontos centrais* j.

A Figura 4.3 mostra os valores de $\mathcal{I}(\hat{\boldsymbol{\theta}}_i)$ para β_0, ϕ, σ^2 e τ^2, onde os tamanhos das subamostras são $m = \{100, 200, 300, 400, 500, 600, 700\}$ e cada uma das retas representam diferentes pontos centrais j. Através dessa figura, é possível observar que conforme o tamanho da subamostra aumenta, isto é, maior o número de observações consideradas na estimação, menor é a variância dos parâmetros estimados. Essa tendência fica mais clara para os parâmetros $\beta_0, \sigma^2 e \tau^2$, para o parâmetro ϕ há maior instabilidade para menores valores da subamostra.

Na Figura 4.4 é possível observar como o aumento da subamostra influência no tempo de processamento. Pela análise visual desses resultados, pode-se afirmar que um tamanho de subamostra $m = 400$ é uma boa escolha, visto que a $\mathcal{I}(\hat{\boldsymbol{\theta}}_i)$ não diminui consideravelmente com o aumento de m.

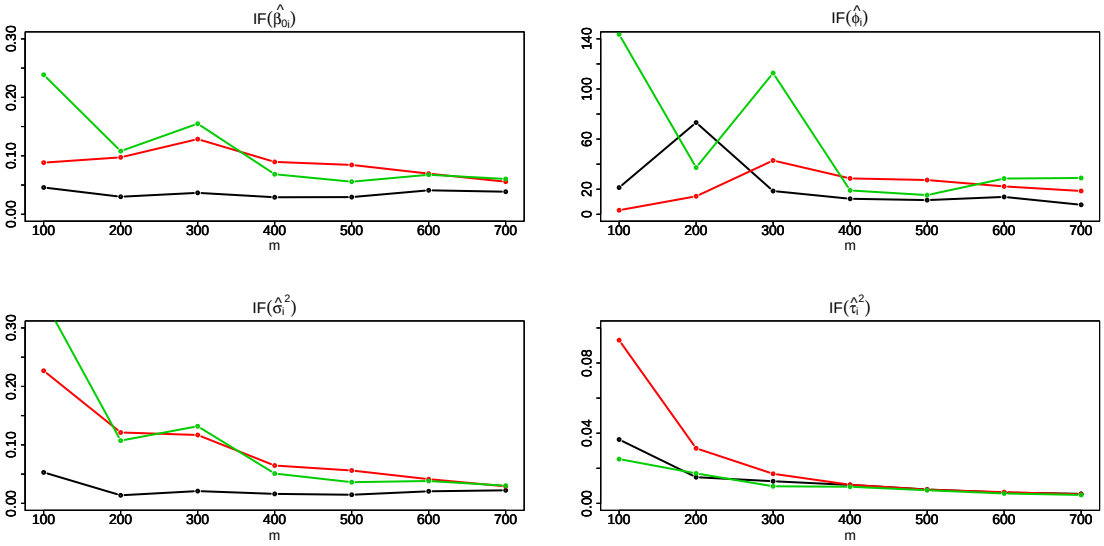

Figura 4.3: *Valores estimados do inverso da matriz de informação de fisher dado o tamanho da subamostra m. As cores representam diferentes pontos centrais j.*

Lembramos que a escolha do tamanho da subamostra é um tanto subjetiva e cada usuário pode considerar um *trade-off* diferente.

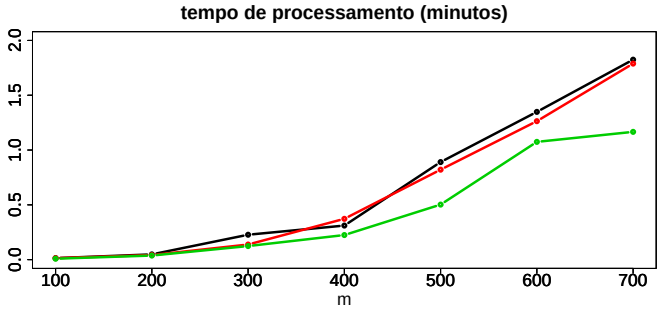

Figura 4.4: *Relação do tempo com o tamanho da subamostra m. Diferentes cores representam diferentes pontos centrais j.*

Além do tamanho da subamostra, é necessário definir uma metodologia para escolha das observações $\mathbf{Z}(\mathbf{s})$. Há vários métodos que podem ser utilizados para esse fim, visto a impossibilidade de análise dos inúmeros possíveis esquemas de seleção, no capítulo 6 apenas três sugestões de algoritmos para escolha da subamostra $\mathbf{Z}(\mathbf{s})$ serão abordados:

- Um algoritmo é denominado 1 centro, em que seleciona-se aleatoriamente um centro ($j = 1$) e suas $k - 1$ observações vizinhas. Essa idéia é baseada na teoria de *subsampling*, que será abordada mais detalhadamente no capítulo 5. Esse método de seleção da subamostra pode não ser o mais adequado, devida à falta de observações distantes, que são importantes no cálculo dos parâmetros da função de covariância.

- Com o objetivo de tentar sanar a falha do algoritmo anterior, propõem-se selecionar uma subamostra em que pontos distantes também são considerados. Esse algoritmo será denominado 1 centro pontos distantes, em que seleciona-se aleatoriamente 1 observação central ($j = 1$), seus k vizinhos mais próximos e além disso $m - j - k$ observações que estão localizadas à grandes distâncias do centro.

- A úlltima possibilidade abordada é aquela em que a subamostra é composta por vários centros ($j > 1$), ou seja, vários subgrupos em diferentes partes do campo aleatório são selecionados.

Ao contrário dos algoritmos anteriores, que utilizam um conjunto de dados concentrados em uma região, esse algoritmo tenta representar o campo aleatório através de vários pedaços de D.

Acredita-se que as subamostras selecionadas através desses algoritmos sejam adequadas para representar um campo aleatório de interesse. Entretanto, a superioridade do desempenho entre os diferentes métodos de seleção e número de centros j vai depender das características específicas dos dados que serão analisados. Por isso, pode ser que em alguns casos, dado o mesmo tamanho da subamostra, $j = 5$ centros possuam uma *perfomance* superior à $j = 1$ e $j = 3$; em outras situações o desempenho de diferentes números de centros podem ser semelhantes. No capítulo 6, um estudo de dados simulados será realizado para melhor entendimento dos diferentes algoritmos, valores de m e B na estimação dos parâmetros do modelo (2.1).

4.5 Número de Repetições B

A técnica de reamostragem bootstrap necessita que vários conjuntos de dados independentes sejam gerados para avaliar a distribuição amostral de algum estimador. Na literatura (Efron) é sabido que depois de uma certa quantidade, aumentar o número de amostras bootstrap não melhora a qualidade da aproximação da distribuição amostral. O grande ganho está no tamanho da amostra original.

Acredita-se que a mesma interpretação é válida para os estimadores *subsemble*. Em estudos simulados observou-se que o tamanho da subamostra m exerce maior influência na qualidade das estimativas, não sendo necessário que o B assuma um valor muito alto.

Essa afirmação é corroborada pelas Figuras 4.5 e 4.6, onde o comportamento dos estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ para diferentes valores de B é reportado. Para representar diferentes dimensões da subamostra, utilizou-se diferentes cores: preta ($m = 100$), vermelha ($m = 300$), verde ($m = 500$) e azul ($m = 700$). É possível observar que, a partir de $B = 25$, os valores estimados de β_0, ϕ, σ^2 e τ^2 não se modificam muito.

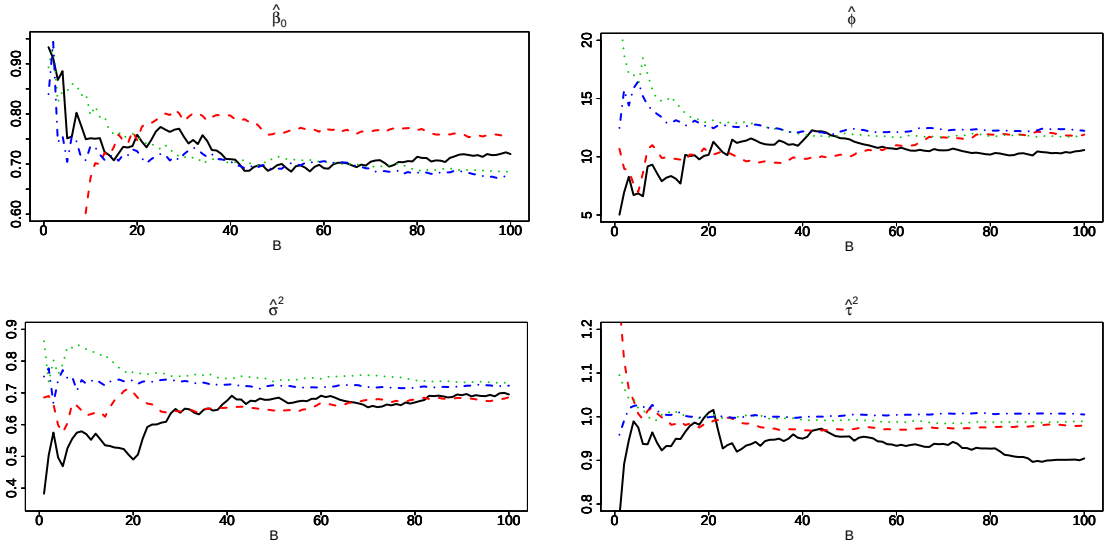

Figura 4.5: *Valores estimados de β_0 (superior direita), ϕ (superior esquerda), σ^2 (inferior esquerda) e τ^2 (inferior direita) para o estimador $\hat{\boldsymbol{\theta}}$ em um campo aleatório com 2000 observações.*

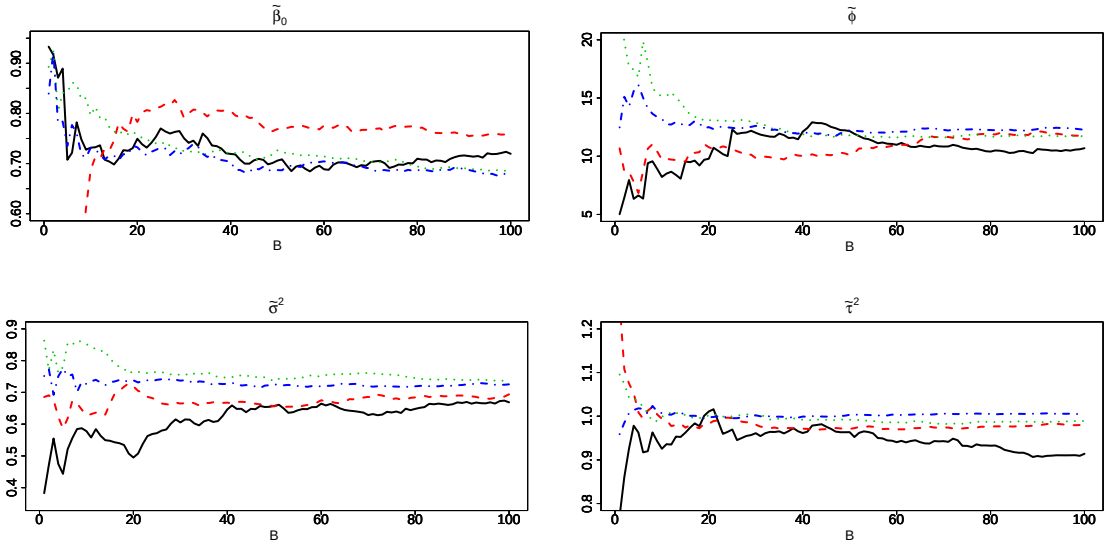

Figura 4.6: *Valores estimados de β_0 (superior direita), ϕ (superior esquerda), σ^2 (inferior esquerda) e τ^2 (inferior direita) para o estimador $\tilde{\boldsymbol{\theta}}$ em um campo aleatório com 2000 observações.*

Capítulo 5

Resultados Assintóticos

Foi mencionado no capítulo 2 que o EMV das duas visões assintóticas em geoestatística possuem propriedades diferentes. Mardia e Marshall (1984) provaram que o EMV dos parâmetros da função de covariância podem ser consistentemente estimados quando estamos no caso do *increasing domain asymptotic*, para *infill asymptotic* não é possível estimar consistentemente tais parâmetros. Entretanto, Zhang (2004) demonstrou que a razão entre os parâmetros da função de covariância Matérn pode ser consistentemente estimada pelo EMV.

Os lemas e teoremas desta seção provam a consistência assintótica do estimador proposto $\tilde{\boldsymbol{\theta}}$. Esses resultados são baseados no trabalho de Liang *et al.* (2013) para *infill asymptotic* e Politis e Romano (1994b) e Politis *et al.* (1999b) para *increasing domain asymptotic*.

5.1 Infill Asymptotics

Para provar os lemas e teoremas apresentados a seguir, utiliza-se um recurso conhecido como amostra auxiliar, na qual amostras infinitas podem ser selecionadas de uma população finita, por meio da reamostragem com reposição. Dada essa interpretação, os resultados assintóticos do estimador de $\boldsymbol{\theta}$ podem ser estudados pelo uso da teoria de estatísticas-U para dados dependentes (Borovskikh, 1996). No restante da seção faz-se uma breve mudança de notação, seja $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = \tilde{\boldsymbol{\theta}}$.

Seja $\mathcal{S}_n = \{Y_1, Y_2, \ldots, Y_n\} = \{Y(s_1), Y(s_2), \ldots, Y(s_n)\}$ uma amostra de um campo aleatório estacionário, definido sobre uma área limitada. Seleciona-se aleatoriamente e com reposição uma amostra de tamanho $N-n$ do vetor $\mathcal{S}_n$. O novo vetor $\mathcal{S}_N = (Y(s_1), Y(s_2), \ldots, Y(s_n), Y(s_{n+1}), \ldots, Y(s_N))$ é a união dos n dados observados com os $N-n$ dados reamostrados.

Seja $\hat{\boldsymbol{\theta}}_i = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}_i|\mathbf{s}_i)$, a equação

$$U_n(\boldsymbol{\theta}) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \hat{\boldsymbol{\theta}}_i = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \psi(Y_1^{(i)}, \ldots, Y_m^{(i)}), \tag{5.1}$$

define uma estatística-U de kernel $\psi = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}_i|\mathbf{s}_i)$ definida sobre a amostra $Y_1, \ldots, Y_n$. Além disso, utiliza-se a seguinte notação:

$$U(\boldsymbol{\theta}) = E[\psi(Y_1, \ldots, Y_m)]. \tag{5.2}$$

Lema 5.1.1. *Seja $\{Y_1, \ldots, Y_n\}$ uma amostra aleatória de um campo aleatório estacionário limitado. Se o mapeamento $(y_1, \ldots, y_m) \mapsto \psi(y_1, \ldots, y_m)$ é contínuo quase certamente e $E|\psi(Y_1, \ldots, Y_m)| < \infty$, então quando $n \to \infty$,*

$$U_n(\boldsymbol{\theta}) \quad \xrightarrow{p} \quad U(\boldsymbol{\theta}).$$

Prova Lema 5.1.1. Segundo Borovskikh (1996), para provar esse Lema é suficiente mostrar que $Var[U_n(\boldsymbol{\theta})] \to 0$ quando $n \to \infty$. Para calcular $Var[U_n(\boldsymbol{\theta})]$, utiliza-se uma amostra auxiliar $\mathcal{S}_N = \{Y_1, \ldots, Y_n, Y_{n+1}, \ldots, Y_N\}$.

Condicionado a essa suposição, S_n pode ser visto como uma simples amostra aleatória de $\mathcal{S}_N$ e $U_n(\boldsymbol{\theta})$ como uma estatística-U definida sobre a população finita $\mathcal{S}_N$. Seguindo o artigo de Zhao e Chen (1990), tem-se o seguinte resultado:

$$Var(U_n|\mathcal{S}_N) = \frac{qm^2}{n}\sigma^2_{1,N} + O\left(\frac{q}{nN}\right) + O\left(\frac{q^2}{n^2}\right), \qquad q = 1 - \frac{n}{N}. \tag{5.3}$$

Visto que $\psi(\cdot)$ é contínuo quase certamente e segundo Lahiri (1996), $E|\psi(Y_1,\ldots,Y_m)|^2 < \infty$ e

$$\sigma^2_{1,N} \xrightarrow{p} \sigma^2_1, \qquad \text{quando} \qquad N \to \infty. \tag{5.4}$$

Um dos resultados do trabalho de Lahiri (1996) indica que para um campo aleatório estacionário limitado, a média amostral de qualquer função contínua converge em probabilidade para a verdadeira média, desde que esta exista.

Portanto, pelas equações (5.3) e (5.4):

$$Var[U_n(\boldsymbol{\theta})|\mathcal{S}_\infty] = \lim_{N\to\infty} Var[U_n(\boldsymbol{\theta})|\mathcal{S}_N] = O_p\left(\frac{1}{n}\right), \tag{5.5}$$

em que $O_p(c_n) = c_n O_p(1)$ e $O_p(1)$ denota uma sequência de variáveis convergindo em probabilidade para uma constante.

Por outro lado, $U_n(\boldsymbol{\theta})$ como uma função linear de $\psi(Y_1^{(i)},\ldots,Y_m^{(i)})$ é também contínuo quase certamente e possui momentos de segunda ordem. E segue de Lahiri (1996) que, $Var[U_n(\boldsymbol{\theta})|\mathcal{S}] \xrightarrow{p} Var[U_n(\boldsymbol{\theta})]$ quanto $n \to \infty$, isto é,

$$Var[U_n(\boldsymbol{\theta}|\mathcal{S})] \xrightarrow{p} Var[U_n(\boldsymbol{\theta})] + o_p(1), \tag{5.6}$$

$o_p(1)$ é uma sequência de variáveis aleatórias que convergem para zero em probabilidade. Por (5.5) e (5.6)

$$Var[U_n(\boldsymbol{\theta})] = O_p\left(\frac{1}{n}\right) - o_p(1) = o_p(1), \tag{5.7}$$

como $Var[U_n(\boldsymbol{\theta})]$ é uma sequência constante, (5.7) implica que $Var[U_n(\boldsymbol{\theta})] \to 0$. □

Seja $\boldsymbol{\Theta}$ o espaço paramétrico de $\boldsymbol{\theta}$ e seja $\boldsymbol{\theta}_0 = \{\boldsymbol{\theta}^* \in \boldsymbol{\Theta} : U(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}} U(\boldsymbol{\theta})\}$ denota o conjunto de máximos globais de $U(\boldsymbol{\theta})$.

Lema 5.1.2. *Seja $\{Y_1,\ldots,Y_m\}$ uma amostra aleatória de um campo aleatório estacionário limitado. Assuma que as seguintes condições sejam verdadeiras:*

i. O mapeamento $\theta \mapsto \psi(Y_1,\ldots,Y_m)$ é contínuo para quase todo $(Y_1,\ldots,Y_m)$ e satisfaz

$$E|\psi(Y_1,\ldots,Y_m)|^2 < \infty. \tag{5.8}$$

ii. Para todo círculo, suficientemente pequeno $O \subset \Theta$, o mapeamento $(y_1,\ldots,y_m) \mapsto \sup_{\boldsymbol{\theta}\in O}\psi(y_1,\ldots,y_m)$ e

$$E|\sup_{\boldsymbol{\theta}\in O}\psi(Y_1,\ldots,Y_m)|^2 < \infty. \tag{5.9}$$

Então para qualquer estimador $\tilde{\boldsymbol{\theta}}_n$ tal que $U_n(\tilde{\boldsymbol{\theta}}_n) \geq U_n(\boldsymbol{\theta}^*) + o_p(1)$ para algum $\boldsymbol{\theta}^* \in \Theta_0$, para todo $\epsilon > 0$ e todo conjunto compacto $\mathcal{K} \subset \Theta$,

$$P(dist(\tilde{\boldsymbol{\theta}}_n, \Theta_0) \geq \epsilon \quad e \quad \tilde{\boldsymbol{\theta}}_n \in \mathcal{K}) \to 0,$$

no qual $dist(\cdot,\cdot)$ denota a métrica de distância.

Para provar o 5.1.2, utilizaremos como ferramenta o lema a seguir (Liang *et al.*, 2013):

Lema 5.1.3. *Seja $Y_i^{(k)}$, $k = 1, \ldots, q$ denotando q sequências de variáveis aleatórias. Se q é finito e para todo $1 \leq k \leq q$,*

$$Y_i^{(k)} \xrightarrow{p} a_k, \qquad \textit{quando} \qquad i \to \infty,$$

então

$$\max Y_i^{(k)} \xrightarrow{p} \max_k a_k, \qquad \textit{quando} \qquad i \to \infty.$$

Prova do Lema 5.1.2. Fixe algum $\boldsymbol{\theta}$ e seja $O_l \downarrow \boldsymbol{\theta}$ uma sequência decrescente de círculos abertos em torno de $\boldsymbol{\theta}$ com o diâmetro convergindo para zero. Seja $\psi_O = \sup_{\boldsymbol{\theta} \in O} \psi$, a sequência ψ_{O_l} é decrescente e não é menor que $\psi_{\boldsymbol{\theta}}$ para todo l. Visto que ψ é contínuo em $\boldsymbol{\theta}$, $\psi_{O_l} \downarrow \psi$ quase certamente, pela condição (5.9) e pelo teorema da convergência monótona, obtêm-se $E\psi_{O_l} \downarrow E\psi$.

Para $\boldsymbol{\theta} \notin \Theta_0$, $E\psi < E\psi_{\boldsymbol{\theta}^*}$, seguindo argumentos do parágrafo anterior, para todo $\boldsymbol{\theta} \notin \Theta_0$ existe um círculo aberto $O_{\boldsymbol{\theta}}$ por volta de $\boldsymbol{\theta}$ com $E\psi_O < E\psi_{\boldsymbol{\theta}^*}$. O conjunto $G = \{\boldsymbol{\theta} \in \mathcal{K} : d(\boldsymbol{\theta}, \boldsymbol{\Theta_0}) \geq \epsilon\}$ é compacto e é coberto pelos círculos $\{O_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in G\}$. Seja $O_{\boldsymbol{\theta}_1}, \ldots, O_{\boldsymbol{\theta}_q}$ um subconjunto finito de G, pelo Lema (5.1.2), condição da equação (5.9):

$$U_{O_{\boldsymbol{\theta}_j}} \xrightarrow{p} E\psi_{O_{\boldsymbol{\theta}_j}},$$

em que

$$U_{O_{\boldsymbol{\theta}_j}} = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \psi_{O_{\boldsymbol{\theta}_j}}(Y_1^{(i)}, \ldots, Y_m^{(i)}),$$

visto que q é finito, pelo lema 5.1.3

$$\sup_{\boldsymbol{\theta} \in B} U_n(\boldsymbol{\theta}) \leq \max_{j=1,\ldots,q} U_{O_{\boldsymbol{\theta}_j}} \xrightarrow{p} \max_{j=1,\ldots,q} E\psi_{O_{\boldsymbol{\theta}_j}} < E\psi_{\boldsymbol{\theta}^*} = U(\boldsymbol{\theta}^*).$$

Portanto,

$$\sup_{\boldsymbol{\theta} \in B} U_n(\boldsymbol{\theta}) < U(\boldsymbol{\theta}^*) + o_p(1). \tag{5.10}$$

Se $\tilde{\boldsymbol{\theta}}_n \in G$, então

$$\sup_{\boldsymbol{\theta} \in G} U_n(\boldsymbol{\theta}) \geq U_n(\tilde{\boldsymbol{\theta}}_n) \geq U_n(\boldsymbol{\theta}^*) + o_p(1),$$

a última desigualdade segue da definição de $\tilde{\boldsymbol{\theta}}_n$ (que é o maximizador global de U_n). Pelo lema 5.1.2, (condição 5.8), $U_n(\boldsymbol{\theta}^*) \xrightarrow{p} U(\boldsymbol{\theta}^*)$ e

$$\sup_{\boldsymbol{\theta} \in B} U_n(\boldsymbol{\theta}) \geq U_n(\boldsymbol{\theta}^*) + o_p(1).$$

Então, pela equação (5.10)

$$P(\tilde{\boldsymbol{\theta}}_n \in B) \leq \{\sup_{\boldsymbol{\theta} \in B} U_n(\boldsymbol{\theta}) \geq U(\boldsymbol{\theta}^*) + o_p(1)\} \to 0,$$

quando $n \to \infty$. □

Teorema 5.1.4. *Seja $\mathcal{Y} = \{Y(s_1), \ldots, Y(s_n)\}$ uma amostra aleatória de um modelo gaussiano espacial definido sobre uma região limitada, representado pela equação (2.1). Seja $\tilde{\boldsymbol{\theta}}_n$ a solução da equação (4.2), considere $\Theta_0 = \{\boldsymbol{\theta}^* \in \Theta : E[L(\boldsymbol{\theta}^*, \mathbf{Z}|\mathbf{S})] = \sup_{\boldsymbol{\theta} \in \Theta} E[L(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{S})]\}$, onde $(\mathbf{Z}|\mathbf{S})$ denota uma amostra aleatória de tamanho m do modelo (2.1). Assuma que Θ é compacto, então para todo $\epsilon > 0$*

$$P(dist(\tilde{\boldsymbol{\theta}}_n, \Theta_0) \geq \epsilon) \to 0, \quad \textit{quando} \quad n \to \infty.$$

Prova do teorema 5.1.4. Como $\mathbf{Z}$ possui distribuição normal multivariada $L(\boldsymbol{\theta}, \mathbf{z}|\mathbf{s})$ é contínuo e não positivo,

$$L(\boldsymbol{\theta}, \mathbf{z}|\mathbf{s}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_z|) - \frac{1}{2}(\mathbf{z} - \mu_{\mathbf{z}})^T \boldsymbol{\Sigma}_z (\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}). \tag{5.11}$$

Visto que a distribuição normal possui momentos finitos de qualquer ordem, as suposições (5.8) e (5.9) são satisfeitas. Pelo lema 5.1.2, $dist(\tilde{\boldsymbol{\theta}}_n, \Theta_0) \xrightarrow{p} 0$. □

Nas provas mencionadas acima, o cálculo do EMV deveria ser repetido $\binom{n}{m}$ vezes. Para massivas observações há um grande número de combinações, tornando o cálculo dessa média inviável.

Como já é conhecido na literatura, para aproximar uma estatística-U não é necessário calcular todas as combinações. Segundo Lee (1990), na maioria dos casos uma quantidade de fatores dessa soma pode ser omitida, sem inflar excessivamente a variância do estimador. Esse tipo de estatística-U é denominada estatística-U incompleta. Segundo o autor a variância da estatística-U incompleta é sempre maior que com todas as combinações, mas sua eficiência assintótica é razoável até mesmo com um subgrupo B muito menor que $\binom{n}{m}$. Essa afirmação é corroborada pelo corolário 2.4.1 de Politis *et al.* (1999a).

5.2 Increasing Domain

Note que, quando seleciona-se uma subamostra de tamanho m de um conjunto de n observações, pode-se interpretar que esse subconjunto também é uma amostra do modelo subjacente. É intuitivo pensar que os dois conjuntos com m e n sejam parecidos, assim como suas distribuições amostrais. Logo, pode-se pensar em aproximar a distribuição amostral da estatística, através de uma amostra menor de tamanho m.

Esse tipo de metodologia é conhecido na literatura como *subsampling*. Entre os precursores dessa ideia está o método jacknife (Efron), que estima o vício e a variância de alguma estatística através de amostras de tamanho $n-1$. Politis e Romano (1994a) descreveram o método de *subsampling* em que o tamanho da subamostra é $m << n$.

Quando dados espaciais são utilizados, deve-se levar em consideração a dependência dos dados. O mesmo é válido para as técnicas de reamostragem, que foram adaptadas por vários autores para serem utilizadas no contexto de observações correlacionadas. Entre os pesquisadores que trabalham com reamostragem considerando as observações dependentes, pode-se citar os que abordam bloco *bootsrap* Kunsch (1989), *moving block bootstrap* Liu e Singh (1992) e *subsampling* para séries temporais e campos gaussianos Hall *et al.* (1995) Politis e Romano (1994a), Lahiri *et al.* (1999), Sherman (1996), Politis *et al.* (1999b), Sherman e Carlstein (1994), Nordman *et al.* (2007) e Lahiri (2013).

O *bootstrap* espacial e *subsampling* espacial são técnicas de reamostragem com pequenas diferenças metodológicas. A idéia do bloco *bootstrap* espacial é dividir em pedaços a região observada D, após, esses blocos são selecionados aleatoriamente e concatenados, uma nova área D' é criada, com mesmo tamanho que a original. A metodologia de *subsampling* também divide D em blocos, entretando, esses blocos não são contatenados, pois o objetivo é utilizar cópias de escala reduzida de D.

As provas e teoremas do que apresenta-se a seguir são extraídos dos trabalhos de Politis e Romano (1994b) e Politis *et al.* (1999b), os autores demonstraram como a metodologia de subamostragem pode ser estendida para o contexto de séries temporais e campos aleatórios estacionários. Eles aproximam a distribuição amostral de uma estatística, através de valores da estatística recalculada em pequenos subconjuntos dos dados. Esses valores recalculados, são adequadamente normalizados e aproximam a verdadeira distribuição amostral da estatística de interesse. A principal condição para a prova dos teoremas é que a estatística, adequadamente normalizada, possua uma distribuição limite não degenerada.

Primeiramente define-se a notação para a amostra completa, isto é, composta por n observações. Seja $\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{y}) = \tilde{\boldsymbol{\theta}}_{\mathbf{n}}$ e suponha que $\{\mathbf{Y}(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^d\}$ é um campo aleatório observado em d

dimensões definido sobre um espaço de probabilidade $(\Omega, \mathcal{A}, P)$, um ponto $\mathbf{s}$ em D é o vetor linha $\mathbf{s} = (s_1, s_2, \ldots, s_d)$. O tamanho amostral é novamente denotado por n, mas com uma mudança de notação: $\mathbf{n} = (n_1, n_2, \ldots, n_d) \in D$ representam o tamanho da amostra, em que n_i é o número de valores distintos que a i-ésima coordenada assume.

O conjunto $E_{\mathbf{n}} \subset \mathbb{R}^d$ representa o retângulo que contém todos os pontos observados, o qual será denotado por

$$E_{\mathbf{n}} = \{\mathbf{s} \in D : 0 < s_k \leq n_k, \text{ para } k = 1, \ldots d\}.$$

Além disso, $|E_{\mathbf{n}}|$ indica a medida de Lebesgue (cardinalidade no caso discreto) do conjunto $E_{\mathbf{n}}$.

Considere que $\tilde{\boldsymbol{\theta}}_{\mathbf{n}}$ e $\boldsymbol{\theta}$ assumem valores em um espaço de Banach separável, (endowed) com norma $||\cdot||$. Define-se $J_{\mathbf{n},||\cdot||}(P)$ como a distribuição amostral de $\upsilon_n||\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}||$ baseada em uma amostra de tamanho $\mathbf{n}$ de (P) e $\upsilon_{\mathbf{n}}$ é uma constante de normalização. A função de distribuição amostral acumulada da estatística, considerando a amostra "completa" é

$$J_{\mathbf{n},||\cdot||}(y, P) = Prob_P\{\upsilon_{\mathbf{n}}||\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}|| \leq y\}.$$

A notação a seguir refere-se a subamostra. Seja o bloco $Z_{\mathbf{u}} = \{Y(\mathbf{s}), \mathbf{s} \in E_{\mathbf{u},\mathbf{m}}\}$, o conjunto

$$E_{\mathbf{u},\mathbf{m}} = \{\mathbf{s} = (s_1, \ldots, s_d) : u_j < s_j \leq u_j + m_j, j = 1, \ldots, d\}$$

representa o retângulo $E_{\mathbf{n}-\mathbf{m}}$ formado pelos pontos $\mathbf{s} = (s_1, \ldots, s_d) \in \mathbb{R}^d$ tal que $0 < s_k \leq n_k + m_k$ para $k = 1, \ldots, d$. Além disso, o vetor $\mathbf{m}$ indica a forma e tamanho do retângulo e o vetor $\mathbf{u}$ indica sua posição na área D. A notação fica mais clara por meio da Figura 5.1, onde cada retângulo representa diferentes subamostras. A figura da esquerda indica que os valores de m_1 e m_2 são próximos, na figura da direita, a diferença entre os valores de m_1 e m_2 é mais acentuada.

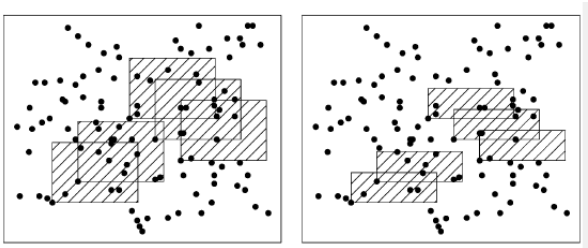

Figura 5.1: *Exemplos de blocos $E_{\mathbf{u},\mathbf{m}}$, para 5 vetores $\mathbf{u}$ e 2 vetores $\mathbf{m}$. (Fonte: Lahiri e Zhu (2006)).*

Seja o valor calculado da estatística para a subamostra Z_u denotado por $\tilde{\boldsymbol{\theta}}_{\mathbf{n},\mathbf{m},\mathbf{u}} = \tilde{\boldsymbol{\theta}}_{\mathbf{m}}(\mathbf{Z}_{\mathbf{u}})$. A aproximação para $J_{\mathbf{n},||\cdot||}(y, \mathbf{P})$ é

$$\hat{J}_{\mathbf{n},\mathbf{m},||\cdot||}(y) = |E_{\mathbf{n}-\mathbf{m}}|^{-1} \int_{E_{\mathbf{n}-\mathbf{m}}} 1\{\upsilon_{\mathbf{m}}(||\tilde{\boldsymbol{\theta}}_{\mathbf{n},\mathbf{m},\mathbf{u}} - \tilde{\boldsymbol{\theta}}_{\mathbf{n}}|| \leq y\} d\mathbf{u}. \tag{5.12}$$

É assumido que o campo aleatório satisfaça uma condição fraca de dependência. Pela literatura de mistura de campos aleatórios, para dois pontos $\mathbf{s} = (s_1, \ldots, s_d)$ e $\mathbf{s}' = (s'_1, \ldots, s'_d)$ em D, defina o *sup* da distância em $\mathbb{R}^d$ por $\delta(\mathbf{s}, \mathbf{s}') = \sup_j |s_j - s'_j|$ e para dois conjuntos E_1, E_2 em $\mathbb{R}^d$, seja $\delta(E_1, E_2) = \inf\{\delta(\mathbf{s}, \mathbf{s}') : \mathbf{s} \in E_1, \mathbf{s}' \in E_2\}$. Defina os coeficientes de mistura forte por:

$$\begin{aligned} \alpha_Y(k; l_1, l_2) \equiv \sup_{E_1, E_2 \subset D} \{&|P(A_1 \cap A_2) - P(A_1)P(A_2)| : \\ &|E_i| \leq l_i, i = 1, 2, \delta(E_1, E_2) \geq k\} \end{aligned} \tag{5.13}$$

em que $\mathcal{F}(E_i)$ é uma sigma álgebra gerada por $\{Y(\mathbf{s}), \mathbf{s} \in E\}$. Uma condição fraca de dependência é formulada quando $\alpha_Y(k; l_1, l_2)$ converge para zero com alguma taxa, quando k tende a infinito e l_1 e l_2 ou permanecem fixos ou tendem a infinito também. Como exemplos de campos aleatórios de mistura forte, pode-se citar os campos Gaussianos com função de densidade espectral estritamente positiva e contínua (Rosenblatt, 2012).

Suposiçao 5.2.1. *$J_{\mathbf{n},||\cdot||}(P)$ converge fracamente para uma lei limite $J_{||\cdot||}(P)$ com correspondente função de distribuição $J_{||\cdot||}(\cdot, P)$, quando $n_i \to \infty$, para $i = 1, \ldots, d$.*

Agora que a notação foi definida e a suposição foi apresentada, podemos enunciar o teorema que prova a consistência da distribuição amostral acumulada da estatística beseada na subamostra $\hat{J}_{\mathbf{n},\mathbf{m},||\cdot||}(y)$ (Politis e Romano, 1994b).

Teorema 5.2.2. *Assuma que 5.2.1 seja verdade, e que $\upsilon_{\mathbf{m}}/\upsilon_{\mathbf{n}} \to 0$, $m_i \to \infty$ e $n_i \to \infty$, para $i = 1, 2, \ldots, d$. Também assuma que exite um vetor $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_n)$, dependendo de $\mathbf{n}$, tal que $2 \leq \Delta_i \leq (n_i - m_i)/m_i$, para todo $i = 1, \ldots, d$, assim como $|\Delta| = \prod_i \Delta_i \to \infty$ e*

$$|\Delta|\alpha_Y \left(\min_i \left[\frac{n_i - m_i}{\Delta_i} - m_i \right] ; (2^{-d}|\Delta| - 1)C(\mathbf{n}, \mathbf{m}, \boldsymbol{\Delta}), 2^d C(\mathbf{n}, \mathbf{m}, \boldsymbol{\Delta}) \right) \to \infty \tag{5.14}$$

onde $C(\mathbf{n}, \mathbf{m}, \boldsymbol{\Delta}) = \prod_i \left(\frac{n_i - m_i}{\Delta_i} + m_i \right)$.

i. Seja y ponto de continuidade de $J_{||\cdot||}(\cdot, P)$, então $\hat{J}_{\mathbf{n},\mathbf{m},||\cdot||}(y) \to J_{||\cdot||}(\cdot, P)$ em probabilidade

ii. Se $J_{||\cdot||}(\cdot, P)$ é contínuo, então $\sup_y |\hat{J}_{\mathbf{n},\mathbf{m},||\cdot||}(y) - J_{\mathbf{n},||\cdot||}(y, P)| \to 0$ em probabilidade.

iii. Seja

$$c_{\mathbf{n},\mathbf{m},,||\cdot||}(1 - \alpha) = \inf\{y : \hat{J}_{\mathbf{n},\mathbf{m},,||\cdot||}(y) \geq 1 - \alpha\}.$$

Correspondentemente, defina

$$c_{||\cdot||}(1 - \alpha, P) = \inf\{y : J_{||\cdot||}(y, P) \geq 1 - \alpha\}.$$

Se $J_{||\cdot||}(\cdot, P)$ é contínuo em $c_{||\cdot||}(1 - \alpha, P)$, então

$$Prob_P\{\upsilon_{\mathbf{n}}||\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}|| \leq c_{\mathbf{n},\mathbf{m},||\cdot||}(1 - \alpha)\} \to 1 - \alpha.$$

Então a probabilidade de cobertura sobre P de $\{\boldsymbol{\theta} : \upsilon_{\mathbf{n}}||\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}|| \leq c_{\mathbf{n},\mathbf{m},||\cdot||}(1 - \alpha)\}$ é de nível nominal $1 - \alpha$.

Demonstração. Seja $\tilde{\boldsymbol{\theta}_{\mathbf{n}}}$ o EMV, segundo Mardia e Marshall (1984) o estimador é assintoticamente normal, logo o processo possui distribuição estacionária e a suposição 5.2.1 é verificada. Além do que, a subamostra m pode ser escolhida tal que $m/n = o(1)$ e $m \to \infty$. □

O teorema acima assume que as subamostras possuem sobreposição máxima. Também, por conveniência de apresentação, a região observada D tinha forma retangular e os blocos da subamostra também foram considerados de formato retangular. Contudo, essa suposição não é necessária. Diferentes formas da amostra e subamostra podem ser observados mantendo as propriedades teóricas do estimador, para isso supõem-se que os dados observados são gerados por processos pontuais marcados. Os autores Lahiri e Zhu (2006), Politis *et al.* (1999b) e Politis e Sherman (2001) abordam esse tipo de situação frequentemente observada.

Capítulo 6

Simulações

6.1 Descrição Geral

Através de simulações de Monte Carlo tem-se por objetivo investigar algumas características relacionadas ao estimador proposto. Pretende-se descobrir como a escolha do número de repetições B, o tamanho da subamostra m e diferentes tipos de seleção da subamostra afetam as estimativas do estimador *subsemble* espacial. Ademais, seu desempenho será comparado com o EMV e com o estimador RSA, proposto por Liang *et al.* (2013). A título de comparação, os cenários das simulações foram inspirados nesse artigo.

Como o estimador permite a utilização de programação paralela, muito tempo de processamento pode ser economizado com o aumento da economia proporcional ao aumento do número de processadores. Para avaliar esse característica, foram utilizadas duas configurações de *hardware*, uma configuração possuía 4 núcleos e a outra 24 núcleos. Todos os resultados foram calculados no *cluster* de processadores SGI Altix, que opera com o Novell SUSE Linux Enterprise Server e uma unidade de conexão, que conta com 64 unidades de processamento. Cada unidade possui 64GB de RAM e 2 processadores dodecacore AMD Opteron (32 unidades com o modelo 6176 SE de 2.3 GHz e outras 32 como o modelo de 2.9 GHz de frequência). Para ter acesso à nuvem, só era possível utilizar uma unidade de processamento por simulação.

O software utilizado para a geração dos dados, estimação e análise dos resultados foi o R versão 3.2.3. O pacote `geoR` (Ribeiro e Diggle, 2001) gerou os dados analisados, calculou o EMV e o estimador proposto. O estimador RSA foi executado através de um pacote disponibilizado pelos autores Liang *et al.* (2013). O pacote `parallel` permitiu a utilização de vários *cores* de forma concomitante.

A função que o pacote `geoR` disponibiliza para calcular o EMV necessita de valores iniciais para o método de otimização. Vários testes foram realizados, em que maximizamos o EMV com valores de *input* fixos e aleatórios, sendo que a estimativa ficou muito parecida nas duas situações. Nas simulações apresentadas nesse trabalho, somente a maximização com valores iniciais escolhidos de forma aleatória será investigada.

Os dados simulados consistem de um campo aleatório gaussiano (equações 2.1 e 2.2), com observações irregularmente espaçadas observadas na região $D = [0, 100] \times [0, 100]$. A função de correlação adotada foi a exponencial $\rho(||s_i - s_j||; \phi) = \exp^{-||s_i - s_j||/\phi}$ e o modelo possui os seguintes parâmetros: $\sigma^2 = 1$, $\beta_0 = 1$ e $\beta_1 = 1$. A variável exploratória V_1 possui distribuição normal, de média 0 e desvio padrão 0,5.

Considera-se 4 diferentes cenários:

- Cenário 1: Foram simulados $Y(s_i), i = 1, \ldots, n$ com $n = 2000$ observações, $\phi = 25$ e $\tau^2 = 1$.
- Cenário 2: $n = 2000$, $\phi = 25$ e $\tau^2 = 0$.
- Cenário 3: $n = 50000$, $\phi = 25$ e $\tau^2 = 1$.
- Cenário 4: $n = 2000$, $\phi = 5$ e $\tau^2 = 1$.

Para cada cenário, o número de replicações Monte Carlo é igual a 50.

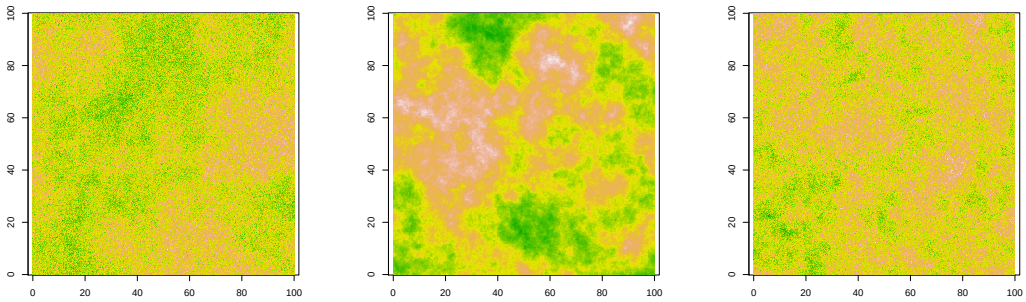

Figura 6.1: *Exemplos de simulações de superfícies dados os diferentes cenários considerados: cenário 1 e 3 (esquerda), cenário 2 (meio) e cenario 4 (direita).*

A Figura 6.1 ilustra uma superfície simulada dos diferentes cenários considerados nesse trabalho. Cada uma dessas configurações de simulação deseja entender e comparar em diferentes contextos os resultados dos estimadores propostos. O cenário 4 deseja avaliar o efeito da amostragem realizada no contexto de *increase domain*, onde o domínio aumenta junto com o número de observações. Como já foi mencionado no capítulo 2, o EMV possui ótimas propriedades para *increase domain asymptotics*, o estimador é consistente e assintoticamente normal. Por isso, os resultados das estimativas e intervalos de confiança para os parâmetros serão reportados individualmente. Os cenários 1, 2 e 3 são exemplos de *infill asymptotics*, como os parâmetros da função de covariância não são consistentes, os resultados individuais de ϕ e σ^2 serão omitidos, em vez disso, mencionaremos as estimativas da razão ϕ/σ^2.

O tamanho da amostra no cenário 3 impossibilita o cálculo do EMV. Portanto, somente os resultados dos estimadores RSA e *subsemble* serão reportados.

6.2 Algoritmos para selecionar a subamostra

Em cada cenário, calcula-se o desempenho dos estimadores *subsemble* espacial sob três diferentes esquemas de subamostragem:

- 5 centros (5C): $j = 5$ pontos centrais são sorteados e as $k = \frac{m}{5} - 1$ observações mais próximas a esses pontos são selecionadas.
- 1 centro (1C): um ponto central $j = 1$ é sorteado e as $m - 1$ observações mais próximas a esse ponto são selecionadas.
- 1 centro e pontos distantes (1CD): seleciona-se aleatoriamente um ponto central $j = 1$ mais os $k = m - 1 - 5$ vizinhos mais próximos. Em seguida, as 5 observações mais distantes ao ponto central j são escolhidas.

Um exemplo dos diferentes métodos de seleção da subamostra pode ser visto na Figura 6.2, na qual fixamos $m = 100$.

Para calcular os pesos w_i em cada subamostra foram sorteadas 50 observações agrupadas espacialmente. O método de seleção desses pontos é o 1C, $j = 1$ e $k = 50 - 1$, desse subconjunto 25 estações foram aleatoriamente atribuídas ao vetor $\mathbf{Z}^v$, as outras 25 estações fazem parte do grupo de predição $\mathbf{Z}^p$.

6.3 Características de $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$

Antes de comparar $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ com os métodos citados, as características do *subsemble* espacial para diferentes escolhas para B, m e método de seleção serão investigadas. As estimativas foram

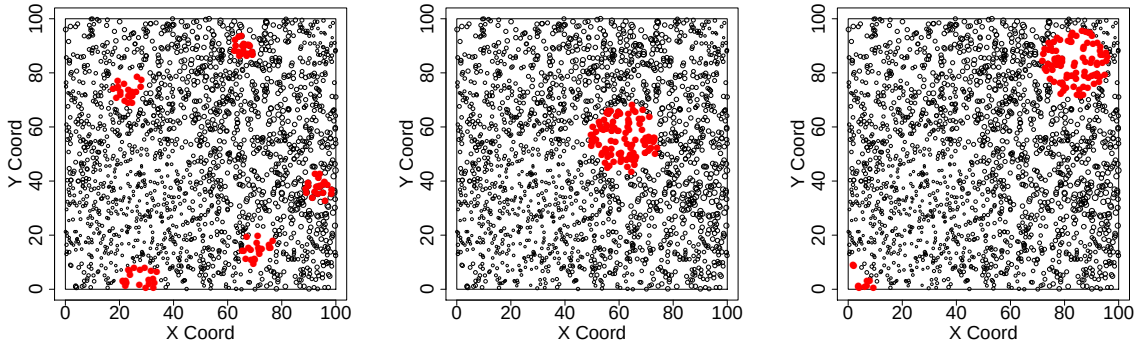

Figura 6.2: *Exemplos de métodos de seleção para subamostras de tamanho $m = 100$ (pontos vermelhos) em um campo gaussiano composto por 2000 observações: 5 centros (esquerda), 1 centro (meio) e 1 centro e pontos distantes (direita).*

calculadas com subamostras de tamanho $m = \{100, 300, 500, 700\}$. O número de subamostras B é igual a 25 e 50 e os tipos de amostragem já mencionados: 1 centro, 1 centro e pontos distantes e 5 centros.

Os tempos de processamento foram omitidos dessa seção, porque a diferença entre a geração das estimativas para os diferentes algoritmos de seleção da subamostra é inexistente. Sobre o tempo de estimação para diferentes valores de B, observa-se o comportamento esperado, sendo que a demora no cálculo das estimativas para $B = 50$ foi aproximadamente o dobro que para $B = 25$.

Para comparar os três algoritmos são utilizadas a média, o desvio-padrão e os boxplots das estimativas dos 50 conjuntos de dados simulados.

6.3.1 Cenário 1

A Tabela 6.1 apresenta os resultados das estimativas para β_0, observa-se que as médias dos valores estimados são muito parecidas com o verdadeiro valor do parâmetro. Pela Figura 6.3 o comportamento dos algoritmos de seleção do estimador *subsemble* espacial fica mais claro, o boxplot dos estimadores propostos são muito similares ao EMV.

Além disso, para $m = 100$ e $m = 300$, o algoritmo 1C apresentou menores vício e variabilidade, seguido pelo algoritmo 1CD e 5C. Para $m = 700$, os esquemas de seleção $1C$ e $5C$ apresentam menor variabilidade. Sobre os distintos valores de B, a variância das estimativas é menor quando $B = 50$, essa tendência fica mais evidente com um tamanho de subamostra menor, quando $m = 700$ a diferença torna-se imperceptível para os algoritmos 1 centro distantes e 5 centros. Sobre os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$, parece não haver diferença entre as estimativas.

Tabela 6.1: *Estimativas do parâmetro $\beta_0 = 1$ para o cenário 1, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	1.055 (0.401)	1.041 (0.401)	1.055 (0.395)	1.029 (0.383)
	50	1.061 (0.409)	1.042 (0.399)	1.034 (0.395)	1.032 (0.384)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	1.055 (0.400)	1.042 (0.399)	1.057 (0.395)	1.028 (0.382)
	50	1.058 (0.408)	1.042 (0.398)	1.035 (0.395)	1.033 (0.383)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.023 (0.330)	1.013 (0.317)	1.037 (0.352)	1.036 (0.415)
	50	1.013 (0.315)	1.014 (0.312)	1.032 (0.350)	1.037 (0.410)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.020 (0.333)	1.014 (0.314)	1.037 (0.349)	1.038 (0.415)
	50	1.016 (0.317)	1.015 (0.309)	1.035 (0.347)	1.036 (0.409)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.019 (0.404)	1.059 (0.399)	1.021 (0.405)	1.033 (0.390)
distantes	50	1.030 (0.385)	1.049 (0.388)	1.031 (0.393)	1.029 (0.387)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.014 (0.404)	1.057 (0.395)	1.020 (0.402)	1.034 (0.389)
distantes	50	1.030 (0.385)	1.047 (0.387)	1.030 (0.391)	1.030 (0.388)

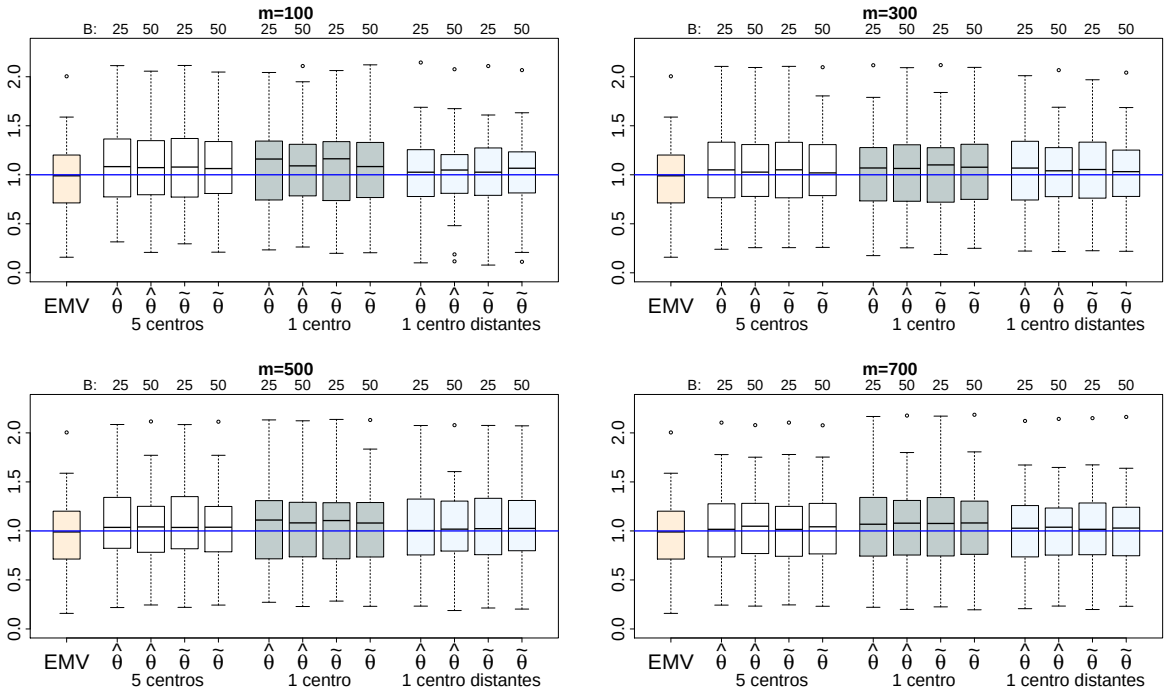

Figura 6.3: *Boxplot das estimativas de β_0 para o cenário 1, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$. A linha azul representa o verdadeiro valor do parâmetro.*

A Tabela 6.2 mostra a média e desvio-padrão das estimativa para β_1, novamente as médias das estimativas são muito parecidas com o parâmetro. Entretanto, não é possível diferenciar a qualidade dos diferentes algoritmos, essa interpretação é confirmada pela Figura 6.4. Outra característica possível de ser observada é que $B = 50$ apresentou um pouco menos de variabilidade que $B = 25$.

Tabela 6.2: *Estimativas de $\beta_1 = 1$ para o cenário 1, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	1.001 (0.078)	0.993 (0.055)	0.988 (0.056)	0.992 (0.050)
	50	0.991 (0.058)	0.991 (0.052)	0.993 (0.052)	0.991 (0.050)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	1.000 (0.078)	0.993 (0.055)	0.988 (0.056)	0.992 (0.050)
	50	0.991 (0.058)	0.991 (0.051)	0.993 (0.052)	0.991 (0.050)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.979 (0.065)	0.986 (0.057)	0.996 (0.058)	0.988 (0.057)
	50	0.984 (0.060)	0.984 (0.054)	0.992 (0.058)	0.990 (0.057)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.980 (0.064)	0.986 (0.056)	0.994 (0.057)	0.990 (0.056)
	50	0.985 (0.060)	0.984 (0.054)	0.992 (0.057)	0.990 (0.057)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.995 (0.058)	0.992 (0.059)	0.990 (0.051)	0.993 (0.054)
distantes	50	1.000 (0.054)	0.991 (0.055)	0.993 (0.050)	0.993 (0.053)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.994 (0.060)	0.991 (0.058)	0.989 (0.049)	0.993 (0.054)
distantes	50	0.999 (0.055)	0.990 (0.054)	0.992 (0.050)	0.994 (0.053)

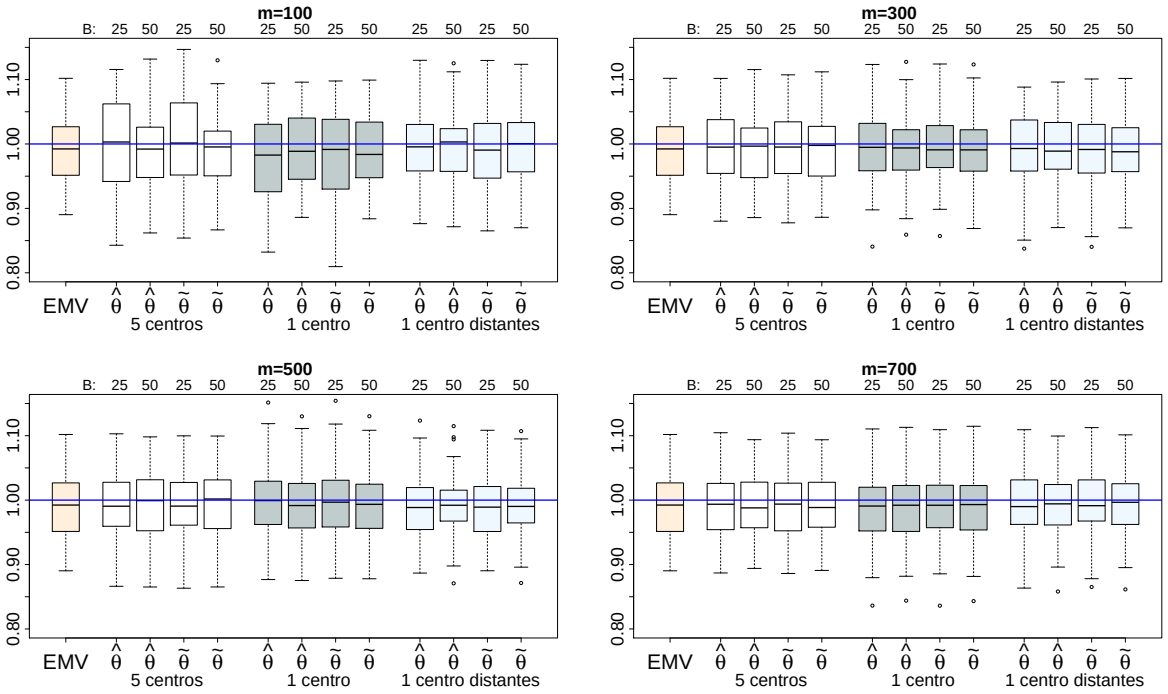

Figura 6.4: *Boxplot das estimativas do parâmetro β_1 para o cenário 1, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Entre a razão ϕ/σ^2 observa-se uma diferença entre os métodos de seleção. O algoritmo 1C apresenta os piores resultados e o algoritmo 5C apresenta menor vício e variabilidade (Tabela 6.3), os boxplots da Figura 6.5 reforçam essa interpretação. A diferença entre as estimativas de $B = 25$ e $B = 50$ é observada para subamostras $m = 100$ e $m = 300$.

Os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ apresentam diferenças para o algoritmo 5C quando $m = 100$, 1C quando $m = 500$ e 1CD para $m = 300$. É vísivel que o aumento da subamostra diminui o vício das estimativas, independentemente do algoritmo, sendo que para subamostra de tamanho 700 os estimadores *subsemble* são praticamente idênticos ao EMV.

Tabela 6.3: *Estimativas da razão $\phi/\sigma^2 = 25$ para o cenário 1, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	20.848 (6.109)	23.225 (4.947)	23.388 (5.270)	23.859 (5.363)
	50	19.941 (4.422)	23.030 (4.905)	23.440 (4.941)	23.885 (5.667)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	20.699 (5.926)	23.353 (5.133)	23.478 (5.345)	23.929 (5.396)
	50	19.979 (4.330)	23.058 (4.946)	23.461 (4.977)	23.904 (5.664)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	8.967 (5.376)	13.155 (9.252)	18.576 (10.082)	23.637 (5.619)
	50	8.997 (5.264)	13.267 (9.391)	18.243 (9.533)	23.623 (5.800)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	8.930 (5.367)	13.299 (9.389)	18.580 (10.008)	23.743 (5.636)
	50	8.999 (5.270)	13.297 (9.392)	18.254 (9.556)	23.689 (5.868)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	19.602 (6.538)	22.864 (5.829)	23.577 (5.873)	23.751 (6.152)
distantes	50	19.661 (6.185)	23.363 (5.906)	23.617 (5.983)	23.768 (5.874)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	19.182 (6.595)	22.959 (5.667)	23.531 (5.855)	23.722 (6.001)
distantes	50	19.506 (6.037)	23.367 (5.904)	23.678 (5.979)	23.781 (5.748)

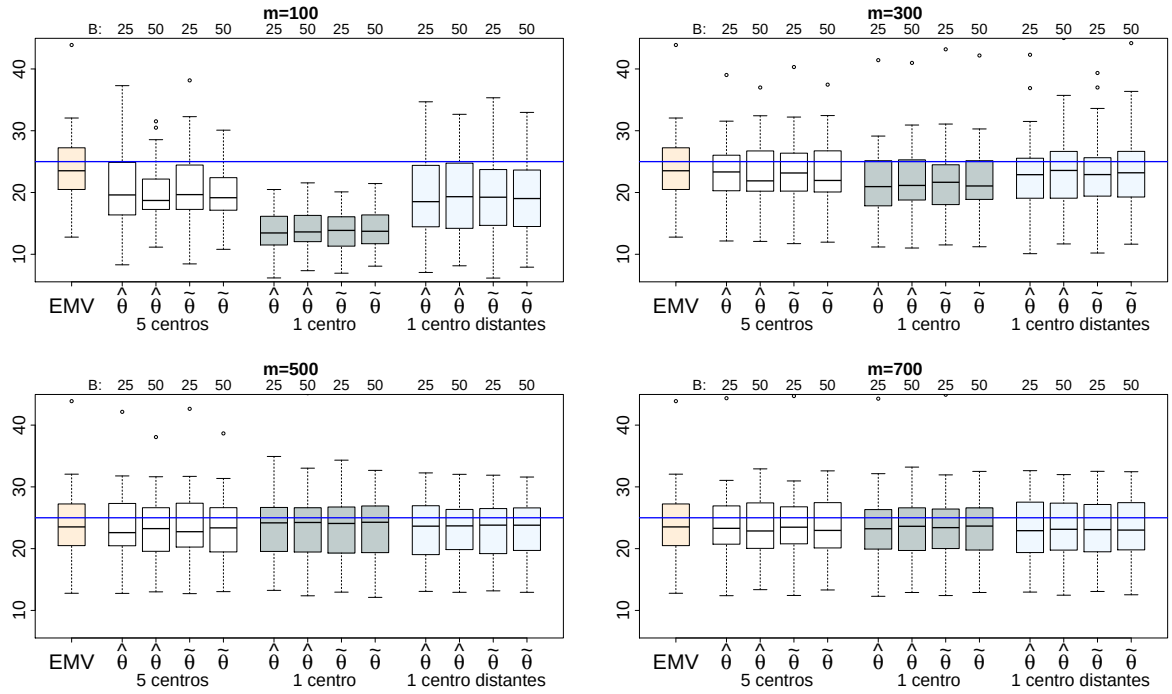

Figura 6.5: *Boxplot das estimativas de ϕ/σ^2 para o cenário 1, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Na Tabela 6.4 e Figura 6.6 temos os resultados de τ^2. A distribuição das estimativas se aproxima do EMV conforme o tamanho da subamostra aumenta. Para $m = 500$ e $m = 700$ a diferença entre os estimadores *subsemble* e o EMV é muito pequena. Sobre os diferentes métodos de seleção, 5C possui a menor variabilidade, seguido por 1CD e 1C. A diferença entre as médias estimadas para $B = 25$ e $B = 50$ está na terceira casa decimal, independentemente do tipo de seleção.

Tabela 6.4: *Estimativas do parâmetro $\tau^2 = 1$ para o cenário 1, dado diferentes algoritmos de seleção da subamostra, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.915 (0.058)	0.972 (0.043)	0.984 (0.041)	0.989 (0.039)
	50	0.907 (0.055)	0.972 (0.043)	0.983 (0.042)	0.989 (0.040)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.918 (0.057)	0.972 (0.044)	0.984 (0.042)	0.989 (0.039)
	50	0.909 (0.054)	0.971 (0.042)	0.983 (0.042)	0.989 (0.040)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.858 (0.092)	0.960 (0.062)	0.981 (0.049)	0.991 (0.040)
	50	0.859 (0.072)	0.961 (0.056)	0.980 (0.047)	0.990 (0.041)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.855 (0.092)	0.963 (0.062)	0.981 (0.049)	0.991 (0.040)
	50	0.858 (0.072)	0.963 (0.056)	0.980 (0.046)	0.990 (0.040)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.881 (0.080)	0.962 (0.051)	0.983 (0.043)	0.987 (0.041)
distantes	50	0.881 (0.068)	0.965 (0.046)	0.982 (0.041)	0.990 (0.040)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.877 (0.085)	0.963 (0.050)	0.984 (0.042)	0.987 (0.041)
distantes	50	0.881 (0.068)	0.965 (0.045)	0.983 (0.040)	0.990 (0.040)

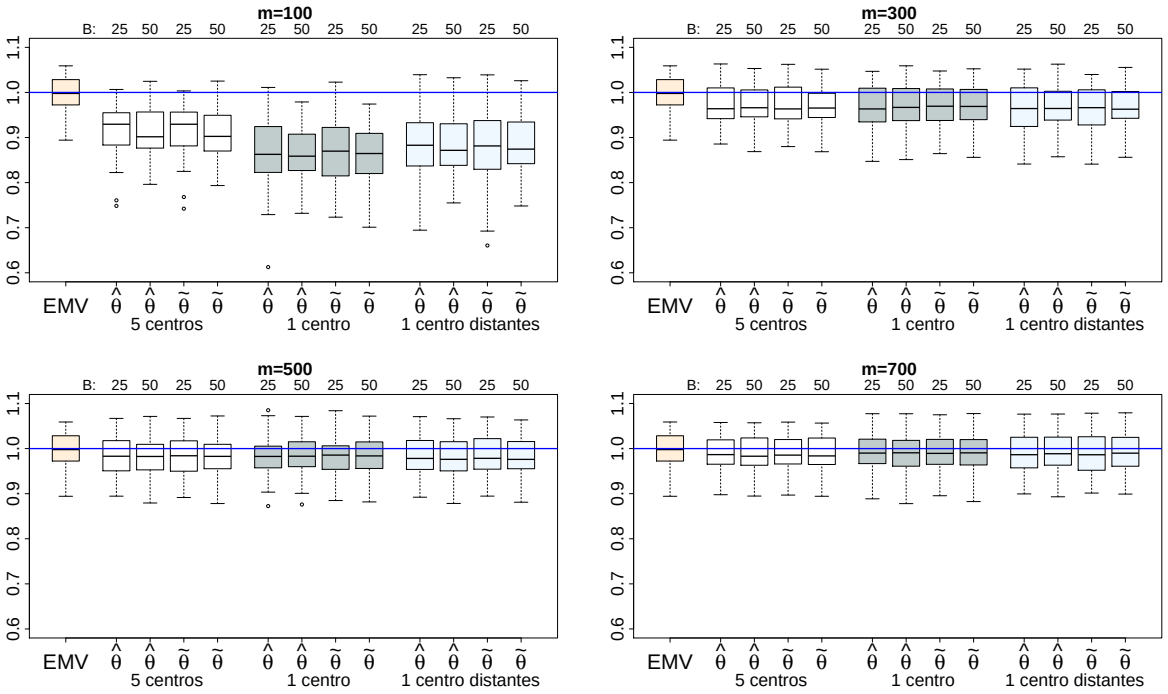

Figura 6.6: *Boxplot das estimativas de τ^2 para o cenário 1, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= $\{5C,\ 1C,\ 1CD\}$, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

6.3.2 Cenário 2

Os resultados das Tabelas 6.5, 6.6 e Figuras 6.7, 6.8 mostram que as estimativas para β_0 e β_1 são muito semelhantes, independentemente do método de seleção, tamanho da subamostra, número de repetições B e estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

Tabela 6.5: *Estimativas do parâmetro $\beta_0 = 1$ para o cenário 2, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	1.041 (0.399)	1.036 (0.394)	1.046 (0.380)	1.023 (0.374)
	50	1.053 (0.401)	1.047 (0.388)	1.039 (0.371)	1.028 (0.369)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	1.048 (0.397)	1.033 (0.394)	1.041 (0.380)	1.021 (0.371)
	50	1.050 (0.401)	1.045 (0.388)	1.039 (0.368)	1.029 (0.367)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.044 (0.403)	1.044 (0.403)	1.029 (0.402)	1.035 (0.390)
	50	1.049 (0.403)	1.040 (0.415)	1.026 (0.406)	1.030 (0.393)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.020 (0.333)	1.014 (0.314)	1.037 (0.349)	1.038 (0.415)
	50	1.016 (0.317)	1.015 (0.309)	1.035 (0.347)	1.036 (0.409)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.075 (0.403)	1.052 (0.394)	1.034 (0.374)	1.032 (0.382)
distantes	50	1.070 (0.392)	1.050 (0.388)	1.038 (0.387)	1.027 (0.381)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.069 (0.388)	1.043 (0.389)	1.036 (0.368)	1.035 (0.380)
distantes	50	1.068 (0.389)	1.044 (0.385)	1.034 (0.382)	1.031 (0.381)

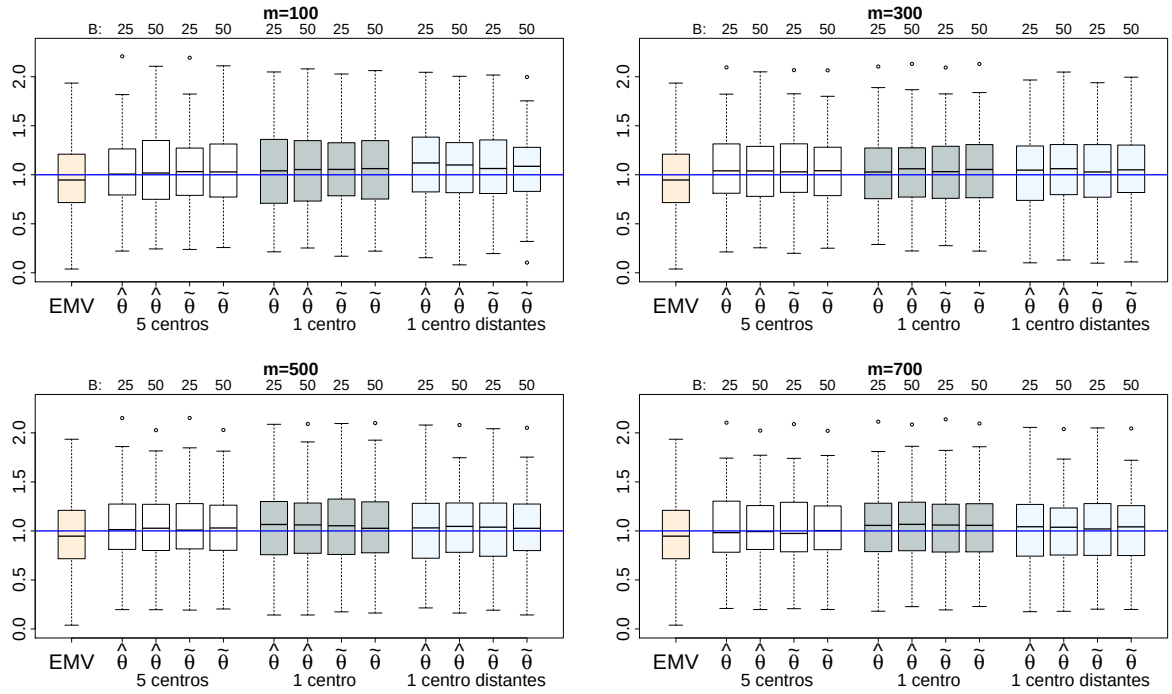

Figura 6.7: *Boxplot das estimativas de β_0 para o cenário 2, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$. A linha azul representa o verdadeiro valor do parâmetro.*

Tabela 6.6: *Estimativas do parâmetro $\beta_1 = 1$ para o cenário 2, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.999 (0.014)	0.998 (0.011)	0.998 (0.011)	0.998 (0.011)
	50	0.999 (0.011)	0.999 (0.011)	0.999 (0.011)	0.998 (0.011)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.998 (0.013)	0.998 (0.011)	0.998 (0.011)	0.999 (0.011)
	50	0.998 (0.011)	0.999 (0.011)	0.999 (0.011)	0.998 (0.011)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.998 (0.012)	0.997 (0.013)	0.999 (0.012)	0.997 (0.012)
	50	0.997 (0.012)	0.998 (0.012)	0.998 (0.012)	0.997 (0.012)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.998 (0.013)	0.998 (0.012)	0.999 (0.012)	0.997 (0.012)
	50	0.997 (0.012)	0.998 (0.012)	0.998 (0.012)	0.997 (0.012)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.000 (0.014)	0.999 (0.012)	0.998 (0.012)	0.998 (0.012)
distantes	50	0.999 (0.013)	0.999 (0.011)	0.998 (0.012)	0.998 (0.012)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.998 (0.013)	0.999 (0.012)	0.998 (0.012)	0.998 (0.012)
distantes	50	0.998 (0.013)	0.999 (0.011)	0.998 (0.012)	0.998 (0.012)

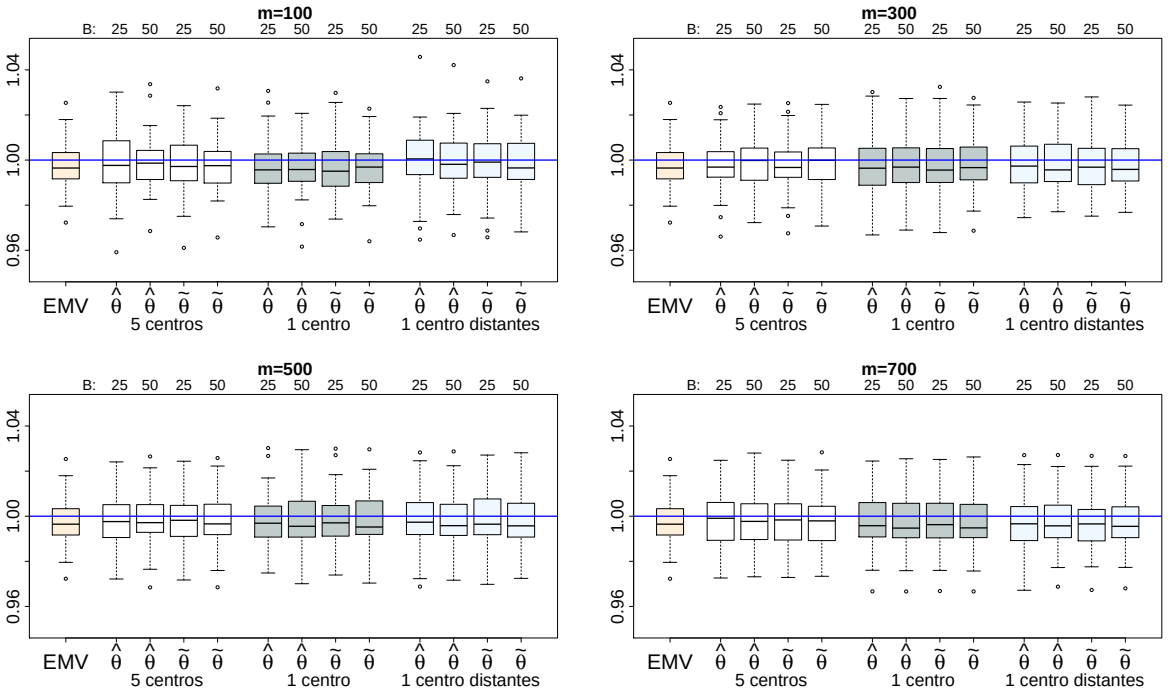

Figura 6.8: *Boxplot das estimativas de β_1 para o cenário 2, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= $\{5C,\ 1C,\ 1CD\}$, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Através da Tabela 6.7 e Figura 6.9 pode-se observar que o EMV, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ superestimam a razão ϕ/σ^2. Apesar disso, os resultados são muito próximos do parâmetro e apresentam pequenos valores para os desvios-padrões. Existe uma grande semelhança entre os diferentes métodos de seleção, não sendo possível identicar superioridade entre nenhum dos algoritmos.

Tabela 6.7: *Estimativas da razão $\phi/\sigma^2 = 25$ para o cenário 2, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	26.699 (2.129)	26.080 (2.008)	25.873 (1.846)	25.867 (1.885)
	50	26.763 (1.864)	26.062 (1.788)	25.918 (1.878)	25.826 (1.854)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	26.768 (2.170)	26.093 (1.973)	25.851 (1.865)	25.881 (1.871)
	50	26.841 (1.857)	26.110 (1.807)	25.943 (1.871)	25.830 (1.845)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	26.769 (2.166)	26.055 (1.912)	25.954 (2.024)	25.802 (1.809)
	50	26.581 (1.987)	26.083 (1.835)	25.923 (1.946)	25.848 (1.836)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	26.840 (2.227)	26.070 (1.868)	25.978 (1.983)	25.818 (1.779)
	50	26.639 (2.077)	26.120 (1.841)	25.957 (1.949)	25.851 (1.826)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	26.983 (2.291)	26.114 (2.033)	25.973 (1.957)	25.870 (1.834)
distantes	50	26.985 (2.094)	26.065 (1.957)	25.936 (1.923)	25.850 (1.839)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	27.132 (2.200)	26.087 (2.055)	26.006 (1.883)	25.877 (1.843)
distantes	50	27.083 (1.998)	26.070 (1.983)	25.951 (1.862)	25.868 (1.853)

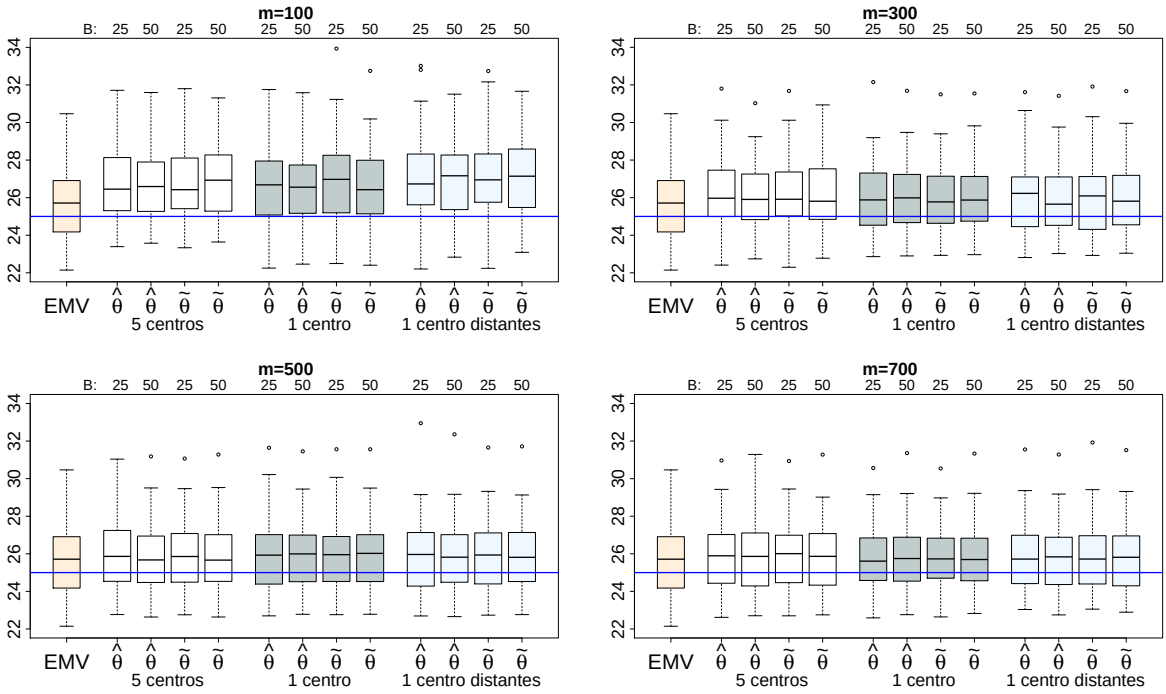

Figura 6.9: *Boxplot das estimativas de ϕ/σ^2 para o cenário 2, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Por meio da Tabela 6.8 pode-ser verificar que os estimadores *subsemble* espacial possuem um ótimo desempenho quando o efeito pepita é nulo. As estimativas e os desvios são muito próximas de zero. Pela Figura 6.10, é possivel distinguir os resultados, quanto maior o tamanho da subamostra maior a proximidade do parâmetro e dos valores estimados. Sobre os diferentes métodos de seleção, o algoritmo 1C apresenta menor variabilidade, quando o tamanho da subamostra é menor. Para $m = 700$ não é possível diferenciar os diferentes algoritmos.

Tabela 6.8: *Estimativas do parâmetro $\tau^2 = 0$ para o cenário 3, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.00167 (0.00124)	0.00102 (0.00127)	0.00098 (0.00127)	0.00078 (0.00100)
	50	0.00175 (0.00110)	0.00110 (0.00103)	0.00091 (0.00107)	0.00082 (0.00116)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.00178 (0.00126)	0.00103 (0.00122)	0.00096 (0.00125)	0.00078 (0.00099)
	50	0.00180 (0.00114)	0.00109 (0.00101)	0.00092 (0.00107)	0.00083 (0.00117)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.00173 (0.00130)	0.00104 (0.00113)	0.00090 (0.00106)	0.00074 (0.00099)
	50	0.00156 (0.00112)	0.00108 (0.00110)	0.00089 (0.00111)	0.00080 (0.00114)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.00173 (0.00133)	0.00105 (0.00110)	0.00091 (0.00109)	0.00077 (0.00104)
	50	0.00157 (0.00111)	0.00110 (0.00109)	0.00090 (0.00113)	0.00081 (0.00116)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.00182 (0.00140)	0.00115 (0.00104)	0.00083 (0.00110)	0.00077 (0.00106)
distantes	50	0.00180 (0.00125)	0.00111 (0.00104)	0.00086 (0.00108)	0.00078 (0.00107)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.00185 (0.00135)	0.00111 (0.00104)	0.00083 (0.00106)	0.00080 (0.00109)
distantes	50	0.00184 (0.00120)	0.00109 (0.00103)	0.00086 (0.00105)	0.00081 (0.00111)

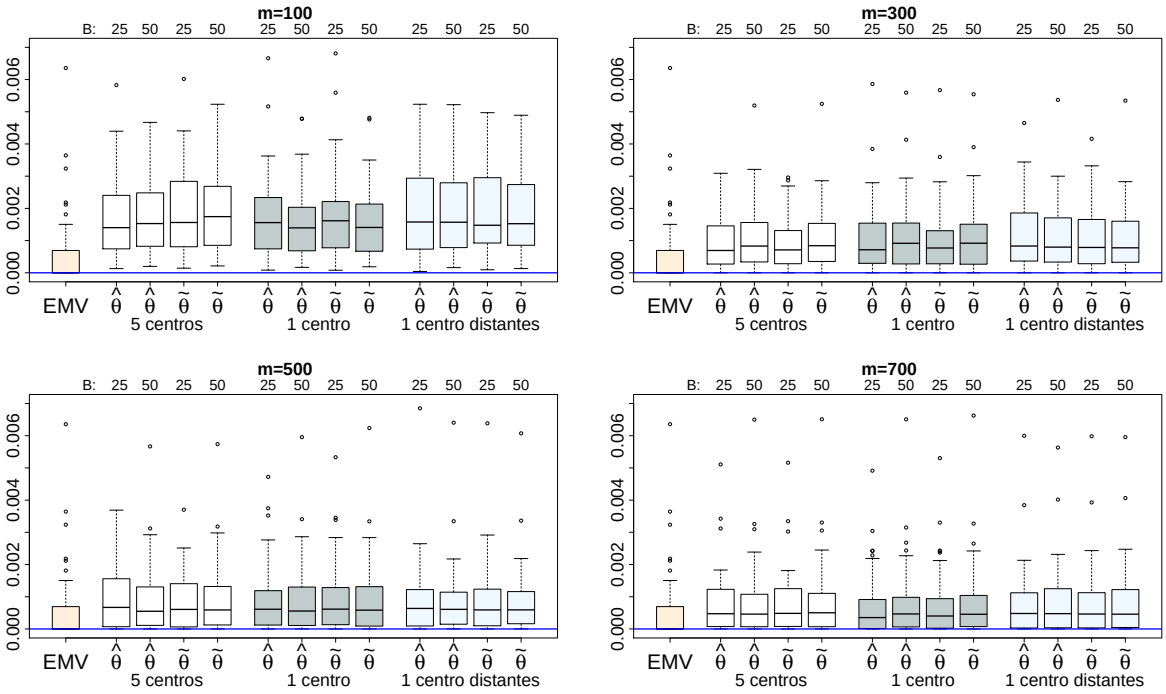

Figura 6.10: *Boxplot das estimativas de τ^2 para o cenário 2, dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= $\{5C, 1C, 1CD\}$, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

6.3.3 Cenário 3

A Tabela 6.9 e a Figura 6.11 apresentam os resultados para β_0. Todas as estimativas são muito próximas, independentemente do tipo de algoritmo de seleção, valor de B e tipo de estimador.

Uma característica desses resultados é que não há muita diferença na qualidade das estimativas, dado diferentes tamanhos de subamostra e tipo de estimador. Por exemplo, para $m = 100$ e $m = 700$, os diferentes estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ são muito parecidos.

Tabela 6.9: *Estimativas do parâmetro $\beta_0 = 1$ para o cenário 3, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.902 (0.458)	0.893 (0.454)	0.889 (0.456)	0.914 (0.473)
	50	0.891 (0.467)	0.905 (0.463)	0.896 (0.459)	0.900 (0.437)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.900 (0.456)	0.894 (0.454)	0.887 (0.455)	0.915 (0.474)
	50	0.894 (0.468)	0.906 (0.462)	0.894 (0.458)	0.901 (0.437)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.877 (0.498)	0.889 (0.489)	0.874 (0.470)	0.920 (0.505)
	50	0.894 (0.478)	0.898 (0.487)	0.880 (0.487)	0.919 (0.477)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.879 (0.504)	0.882 (0.493)	0.867 (0.470)	0.924 (0.496)
	50	0.891 (0.482)	0.889 (0.494)	0.871 (0.484)	0.920 (0.478)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.953 (0.464)	0.962 (0.441)	0.949 (0.444)	0.982 (0.461)
distantes	50	0.965 (0.445)	0.961 (0.436)	0.964 (0.440)	0.969 (0.446)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.958 (0.464)	0.957 (0.433)	0.951 (0.447)	0.988 (0.457)
distantes	50	0.967 (0.445)	0.955 (0.433)	0.962 (0.439)	0.977 (0.445)

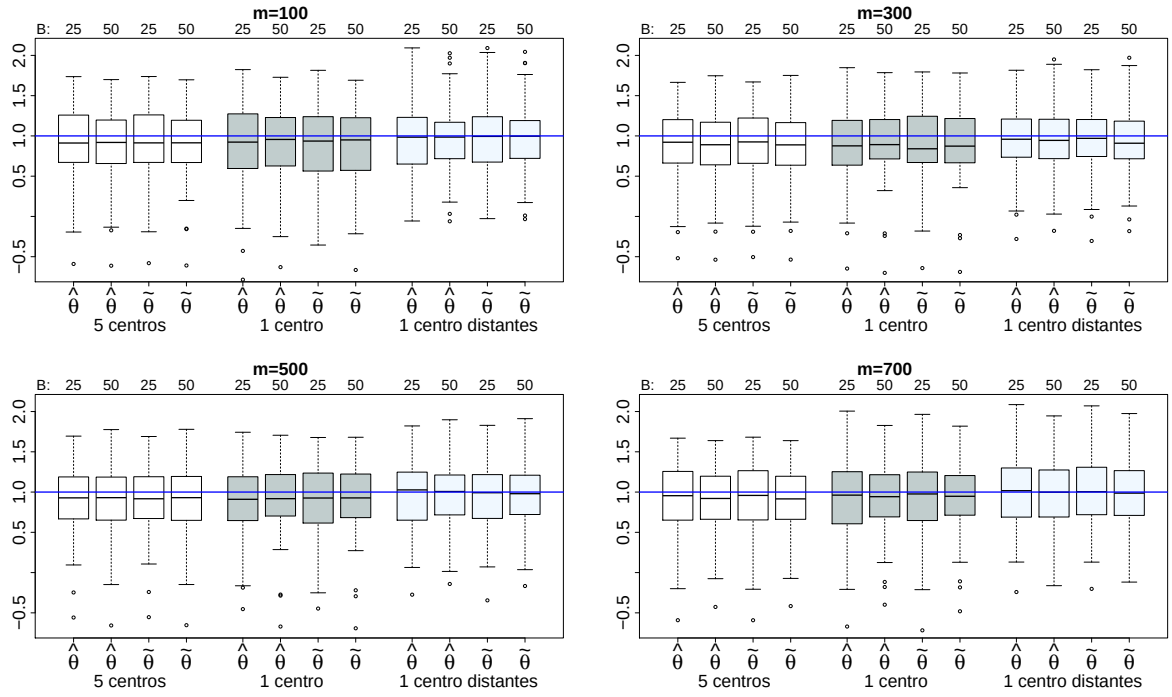

Figura 6.11: *Boxplot das estimativas de β_0 para o cenário 3 dado diferentes estimadores $\{\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$. A linha azul representa o verdadeiro valor do parâmetro.*

Através da Tabela 6.10, pode-se perceber que os desvios-padrões das estimativas de β_1 diminuem, conforme o tamanho de m aumenta. Além disso, $B = 50$ apresenta menor variabilidade que $B = 25$. Pela Figura 6.12, nota-se que os algoritmos possuem resultados muito semelhantes e que não há diferença entre os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

Tabela 6.10: *Estimativas do parâmetro $\beta_1 = 1$ para o cenário 3, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.990 (0.046)	0.996 (0.023)	1.003 (0.020)	0.996 (0.016)
	50	0.999 (0.036)	1.000 (0.023)	1.000 (0.012)	1.001 (0.014)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.989 (0.046)	0.995 (0.022)	1.003 (0.020)	0.995 (0.016)
	50	0.998 (0.035)	1.000 (0.023)	1.000 (0.012)	1.001 (0.014)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.001 (0.049)	1.001 (0.027)	1.001 (0.021)	1.001 (0.019)
	50	0.997 (0.035)	0.999 (0.022)	0.999 (0.014)	0.999 (0.014)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.001 (0.047)	0.999 (0.025)	1.001 (0.019)	1.001 (0.018)
	50	0.997 (0.034)	0.998 (0.020)	0.999 (0.013)	0.999 (0.014)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.013 (0.057)	0.992 (0.032)	0.999 (0.021)	1.001 (0.015)
distantes	50	1.012 (0.043)	0.993 (0.022)	0.999 (0.019)	1.000 (0.011)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.010 (0.053)	0.994 (0.031)	0.999 (0.020)	1.001 (0.015)
distantes	50	1.011 (0.044)	0.994 (0.021)	0.998 (0.019)	1.000 (0.011)

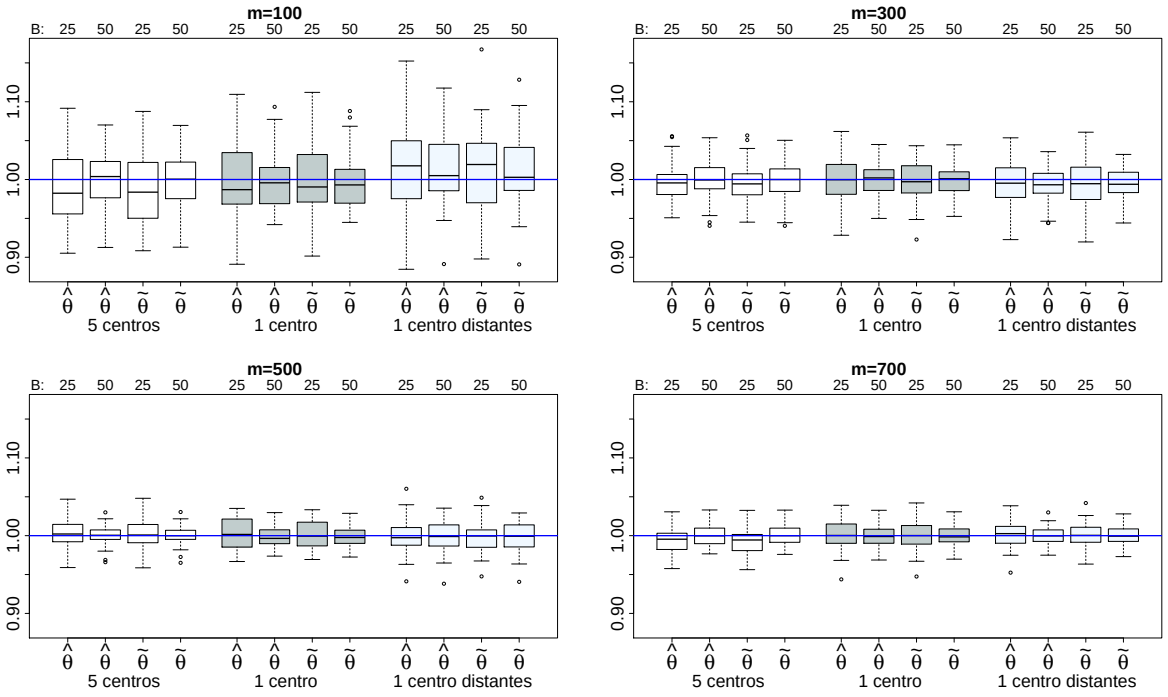

Figura 6.12: *Boxplot das estimativas de β_1 para o cenário 3 dado diferentes estimadores $\{\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Para a razão ϕ/σ^2, observa-se diferença entre os métodos de seleção (Figura 6.13 e Tabela 6.11). O algoritmo 1C apresenta os piores resultados, com o valor do parâmetro sendo muito subestimado, essa característica fica evidente para menores valores de m. Os métodos de seleção 5C e 1CD apresentam resultados muito similares, entretanto, é possível observar um menor vício e variabilidade do método 5C. Analisando as estimativas usando $B = 25$, nota-se uma maior variabilidade do que aquelas utilizando $B = 50$, esse comportamento é independentemente do m, estimador e método de seleção. Na prática, não parece haver diferença entre os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

Tabela 6.11: *Estimativas da razão $\phi/\sigma^2 = 25$ para o cenário 3, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	20.424 (4.8350)	25.164 (4.421)	25.060 (4.000)	25.324 (4.313)
	50	20.481 (3.531)	25.040 (3.9150)	25.876 (3.488)	25.519 (3.205)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	20.448 (4.732)	25.119 (4.416)	24.995 (4.024)	25.305 (4.179)
	50	20.323 (3.621)	25.011 (3.866)	25.823 (3.448)	25.478 (3.193)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	4.752 (1.103)	14.557 (2.504)	19.000 (2.765)	21.115 (3.222)
	50	4.692 (0.900)	14.339 (2.064)	18.857 (2.322)	21.093 (2.561)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	4.713 (1.193)	14.425 (2.495)	19.013 (2.679)	21.115 (3.013)
	50	4.683 (0.932)	14.247 (1.971)	18.869 (2.308)	21.081 (2.478)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	26.690 (6.838)	26.111 (7.006)	23.712 (4.193)	25.697 (5.090)
distantes	50	26.615 (5.992)	25.902 (5.826)	24.719 (3.765)	25.447 (4.657)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	26.444 (6.125)	25.706 (6.620)	23.941 (4.582)	25.826 (5.19)
distantes	50	26.472 (5.836)	25.626 (5.460)	24.916 (3.782)	25.484 (4.562)

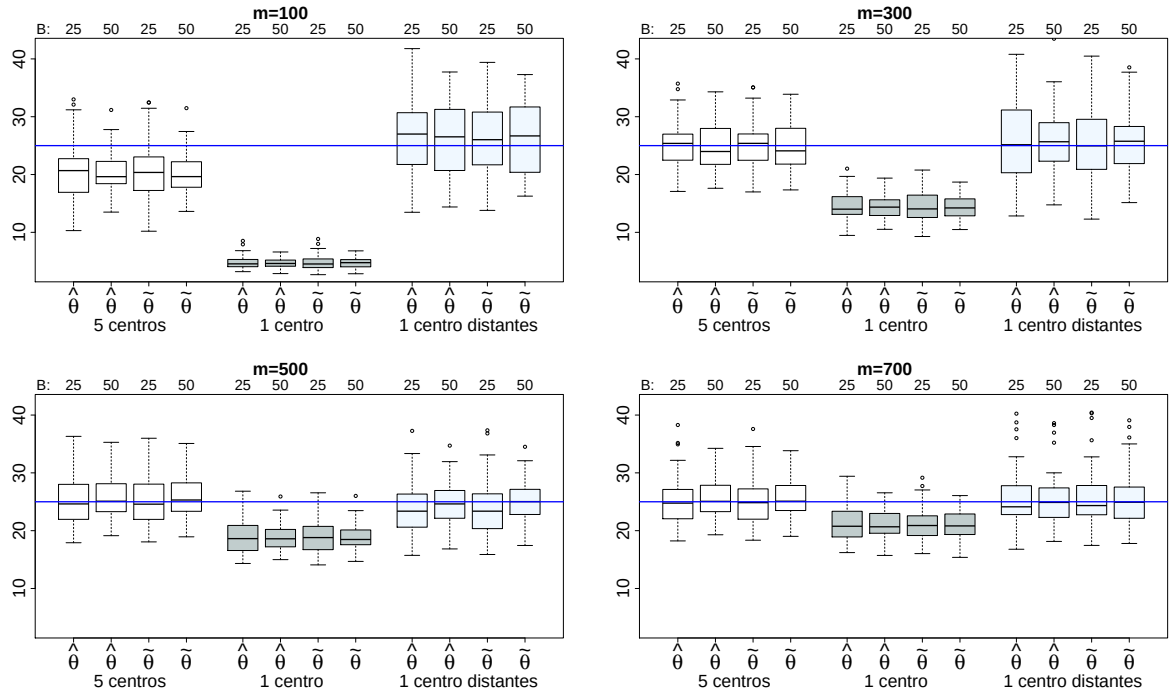

Figura 6.13: *Boxplot das estimativas de ϕ/σ^2 para o cenário 3 dado diferentes estimadores $\{\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

A interpretação dos resultados das estimativas de τ^2 são parecidas com o da razão ϕ/σ^2, observa-se que o tamanho da subamostra possui clara influência nas estimativas. Além disso, o algoritmo 1C possui maior vício e variabilidade, sendo essa característica mais evidente para $m = 100$ e $m = 300$. Novamente, as estimativas considerando $B = 50$ possuem menor variabilidade que $B = 25$. Para o método de seleção 1CD é possível observar uma pequena diferença entre os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

Tabela 6.12: *Estimativas do parâmetro $\tau^2 = 1$ para o cenário 3, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.930 (0.038)	0.981 (0.021)	0.991 (0.016)	0.990 (0.011)
	50	0.940 (0.024)	0.983 (0.017)	0.990 (0.014)	0.994 (0.012)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.929 (0.039)	0.981 (0.021)	0.991 (0.017)	0.990 (0.011)
	50	0.939 (0.024)	0.983 (0.017)	0.989 (0.014)	0.994 (0.013)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.825 (0.047)	0.958 (0.024)	0.984 (0.0160	0.986 (0.014)
	50	0.830 (0.043)	0.957 (0.020)	0.982 (0.012)	0.986 (0.012)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.825 (0.051)	0.957 (0.023)	0.983 (0.015)	0.986 (0.013)
	50	0.828 (0.045)	0.957 (0.019)	0.981 (0.013)	0.985 (0.011)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.922 (0.054)	0.970 (0.028)	0.984 (0.018)	0.989 (0.015)
distantes	50	0.916 (0.044)	0.969 (0.016)	0.985 (0.015)	0.990 (0.013)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.922 (0.052)	0.970 (0.027)	0.984 (0.017)	0.990 (0.014)
distantes	50	0.916 (0.043)	0.969 (0.017)	0.985 (0.014)	0.990 (0.012)

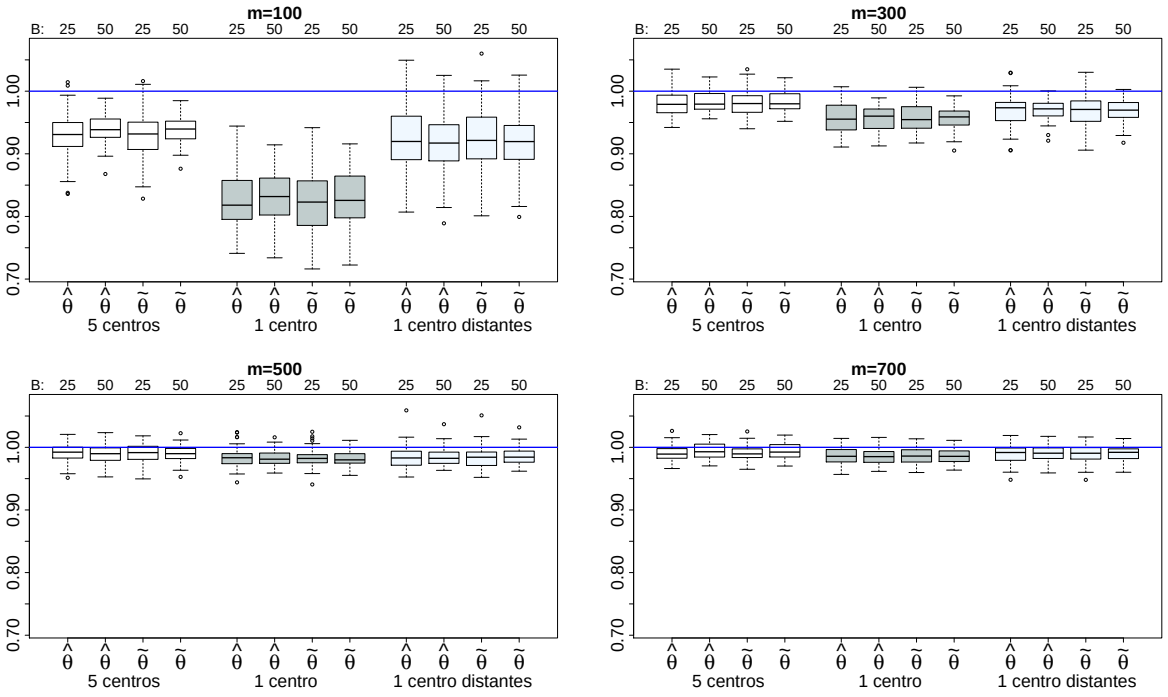

Figura 6.14: *Boxplot das estimativas de τ^2 para o cenário 3 dado diferentes estimadores $\{\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

6.3.4 Cenário 4

Os resultados das Tabelas 6.13, 6.14 e Figuras 6.15, 6.16 mostram que as estimativas para β_0 e β_1 se assemelham ao verdadeiro valor do parâmetro, independentemente do método de seleção, tamanho da subamostra, número de repetições B e estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

Tabela 6.13: *Estimativas do parâmetro $\beta_0 = 1$ para o cenário 4, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.989 (0.128)	0.989 (0.127)	0.989 (0.114)	0.977 (0.112)
	50	0.988 (0.118)	0.981 (0.118)	0.989 (0.113)	0.986 (0.114)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.991 (0.130)	0.988 (0.125)	0.988 (0.113)	0.977 (0.112)
	50	0.988 (0.117)	0.982 (0.118)	0.989 (0.113)	0.986 (0.114)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.978 (0.130)	0.979 (0.136)	0.985 (0.114)	0.982 (0.123)
	50	0.976 (0.119)	0.976 (0.129)	0.978 (0.120)	0.977 (0.123)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.986 (0.138)	0.977 (0.133)	0.985 (0.113)	0.983 (0.122)
	50	0.980 (0.123)	0.976 (0.127)	0.978 (0.118)	0.978 (0.122)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.987 (0.150)	1.010 (0.135)	1.020 (0.123)	1.020 (0.115)
distantes	50	0.993 (0.142)	1.009 (0.125)	1.021 (0.120)	1.017 (0.113)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.992 (0.149)	1.009 (0.138)	1.018 (0.121)	1.021 (0.116)
distantes	50	0.994 (0.144)	1.008 (0.127)	1.020 (0.118)	1.018 (0.113)

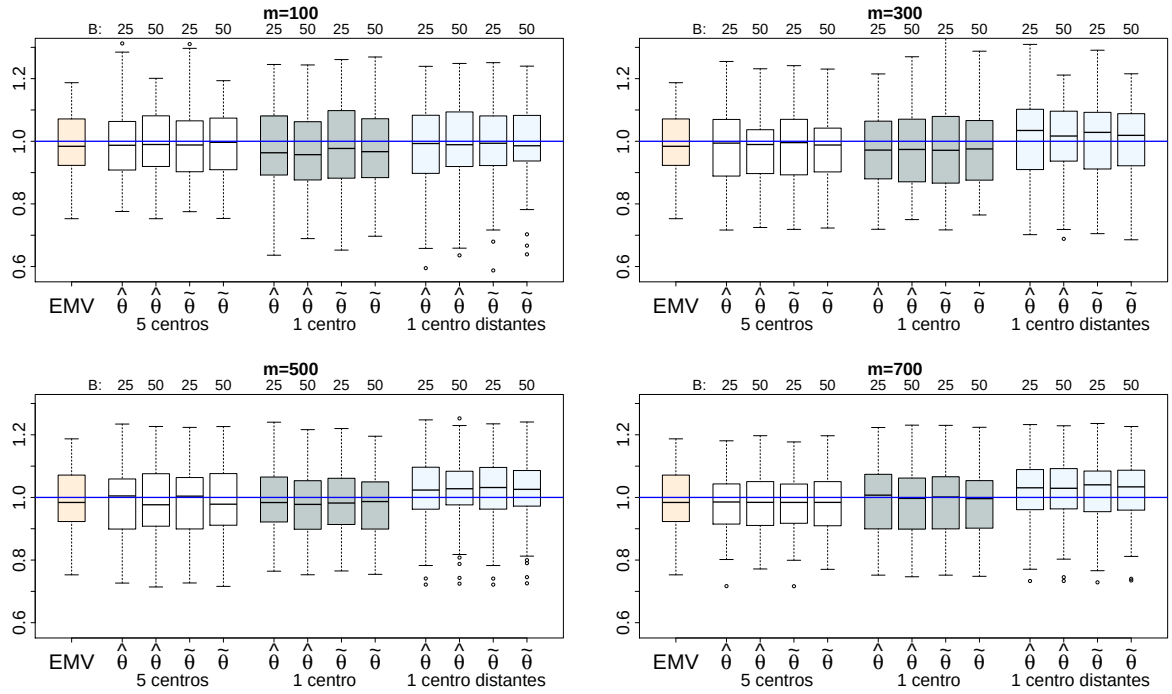

Figura 6.15: *Boxplot das estimativas de β_0 para o cenário 4 dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$. A linha azul representa o verdadeiro valor do parâmetro.*

Tabela 6.14: *Estimativas do parâmetro $\beta_1 = 1$ para o cenário 4, dado diferentes algoritmos de seleção, m e B..*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.985 (0.064)	0.987 (0.051)	0.978 (0.051)	0.983 (0.049)
	50	0.986 (0.058)	0.984 (0.053)	0.980 (0.046)	0.984 (0.047)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.986 (0.064)	0.987 (0.051)	0.978 (0.051)	0.983 (0.049)
	50	0.987 (0.058)	0.983 (0.053)	0.980 (0.046)	0.983 (0.047)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.987 (0.057)	0.979 (0.055)	0.980 (0.049)	0.984 (0.053)
	50	0.993 (0.055)	0.983 (0.052)	0.981 (0.047)	0.984 (0.050)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.985 (0.058)	0.980 (0.053)	0.980 (0.048)	0.984 (0.052)
	50	0.992 (0.054)	0.983 (0.051)	0.981 (0.047)	0.984 (0.050)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.996 (0.072)	0.999 (0.058)	0.995 (0.056)	0.992 (0.052)
distantes	50	0.998 (0.063)	0.997 (0.054)	0.994 (0.053)	0.995 (0.050)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.988 (0.067)	0.998 (0.057)	0.995 (0.055)	0.993 (0.052)
distantes	50	0.994 (0.061)	0.997 (0.055)	0.994 (0.053)	0.995 (0.050)

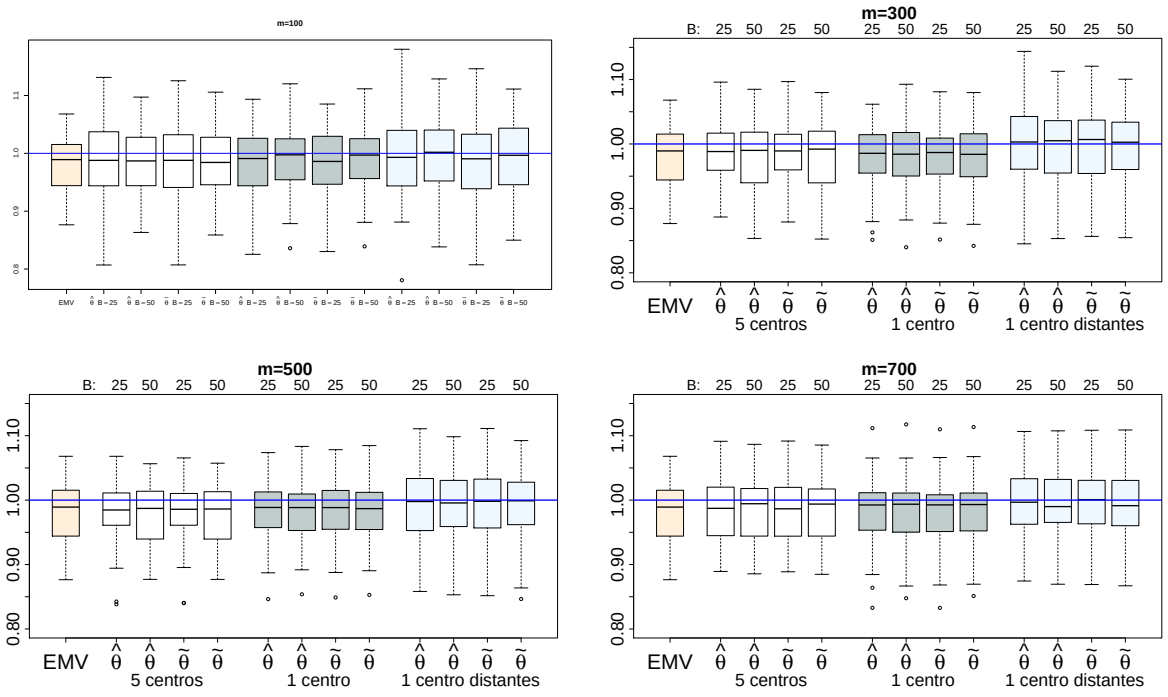

Figura 6.16: *Boxplot das estimativas de β_1 para o cenário 4 dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= $\{5C, 1C, 1CD\}$, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Por meio da Tabela 6.15 e Figura 6.17, nota-se que as estimativas para ϕ são extremamente acuradas. Quando m aumenta, as estimativas ficam mais próximas do verdadeiro valor do parâmetro, além disso, a variabilidade diminui. Para $m = 500$ e $m = 700$, o boxplot das estimativas de $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ são similares ao boxplot do EMV.

Ao analisar os diferentes métodos de seleção, é possível observar que o 1C apresenta maior vício quando $m = 100$. Para os demais valores da subamostra, não é clara a superioridade de nenhum dos algoritmos.

Tabela 6.15: *Estimativas do parâmetro $\phi = 5$ para o cenário 4, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	5.338 (1.228)	4.605 (0.662)	4.821 (0.693)	4.841 (0.705)
	50	5.309 (0.934)	4.676 (0.620)	4.781 (0.642)	4.850 (0.678)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	5.345 (1.191)	4.611 (0.650)	4.825 (0.697)	4.841 (0.706)
	50	5.329 (0.980)	4.685 (0.623)	4.785 (0.638)	4.853 (0.678)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	4.084 (0.784)	4.485 (0.611)	4.744 (0.684)	4.769 (0.715)
	50	4.096 (0.755)	4.541 (0.607)	4.752 (0.671)	4.763 (0.639)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	4.122 (0.766)	4.499 (0.589)	4.746 (0.661)	4.770 (0.710)
	50	4.130 (0.714)	4.555 (0.602)	4.758 (0.654)	4.764 (0.640)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	5.331 (2.190)	4.820 (1.199)	4.834 (0.738)	4.744 (0.621)
distantes	50	5.277 (2.321)	4.829 (1.043)	4.808 (0.692)	4.786 (0.627)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	5.379 (2.163)	4.817 (1.161)	4.818 (0.722)	4.753 (0.626)
distantes	50	5.299 (2.153)	4.819 (0.999)	4.811 (0.686)	4.788 (0.630)

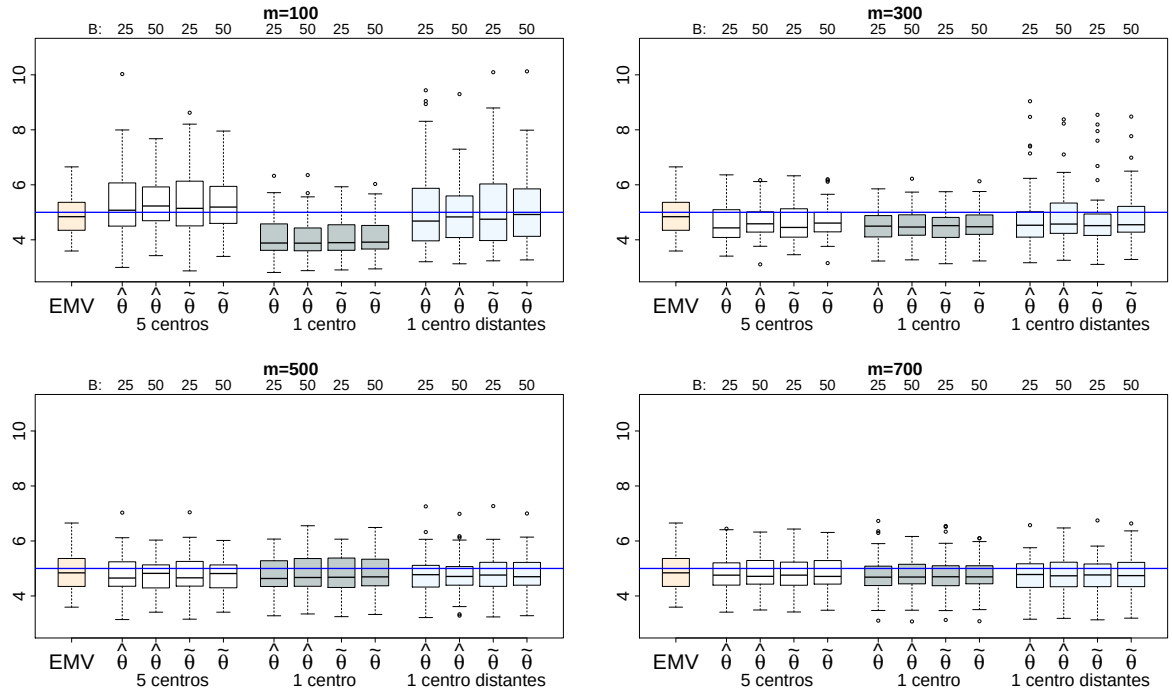

Figura 6.17: *Boxplot das estimativas de ϕ para o cenário 4 dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Através dos resultados da Tabela 6.16 e Figura 6.18 é possível observar que as estimativas com uma subamostra de somente 100 observações estimaram muito acuradamente o parâmetro σ^2. Quando $m = 500$ e $m = 700$, os resultados são muito semelhantes àqueles estimados pelo EMV.

Analisando os resultados condicionando aos diferentes métodos de seleção, número de repetições B e estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$, nota-se que todas as estimativas são muito semelhantes.

Tabela 6.16: *Estimativas do parâmetro $\sigma^2 = 1$ para o cenário 4, dado diferentes algoritmos de seleção, m e B..*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	1.056 (0.126)	0.994 (0.122)	0.992 (0.114)	0.990 (0.124)
	50	1.045 (0.109)	0.983 (0.117)	0.987 (0.119)	0.988 (0.122)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	1.061 (0.128)	0.996 (0.122)	0.992 (0.114)	0.991 (0.124)
	50	1.048 (0.108)	0.984 (0.116)	0.987 (0.118)	0.988 (0.122)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.033 (0.128)	0.986 (0.115)	0.997 (0.113)	0.989 (0.114)
	50	1.031 (0.117)	0.975 (0.117)	0.994 (0.110)	0.987 (0.109)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.033 (0.123)	0.986 (0.115	0.993 (0.114	0.991 (0.115)
	50	1.031 (0.116)	0.978 (0.115)	0.993 (0.110)	0.988 (0.109)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	1.073 (0.161)	1.008 (0.149)	0.989 (0.123)	0.979 (0.109)
distantes	50	1.084 (0.162)	1.007 (0.140)	0.993 (0.118)	0.982 (0.109)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	1.070 (0.157)	1.004 (0.148)	0.990 (0.120)	0.979 (0.111)
distantes	50	1.082 (0.157)	1.004 (0.140)	0.995 (0.117)	0.983 (0.110)

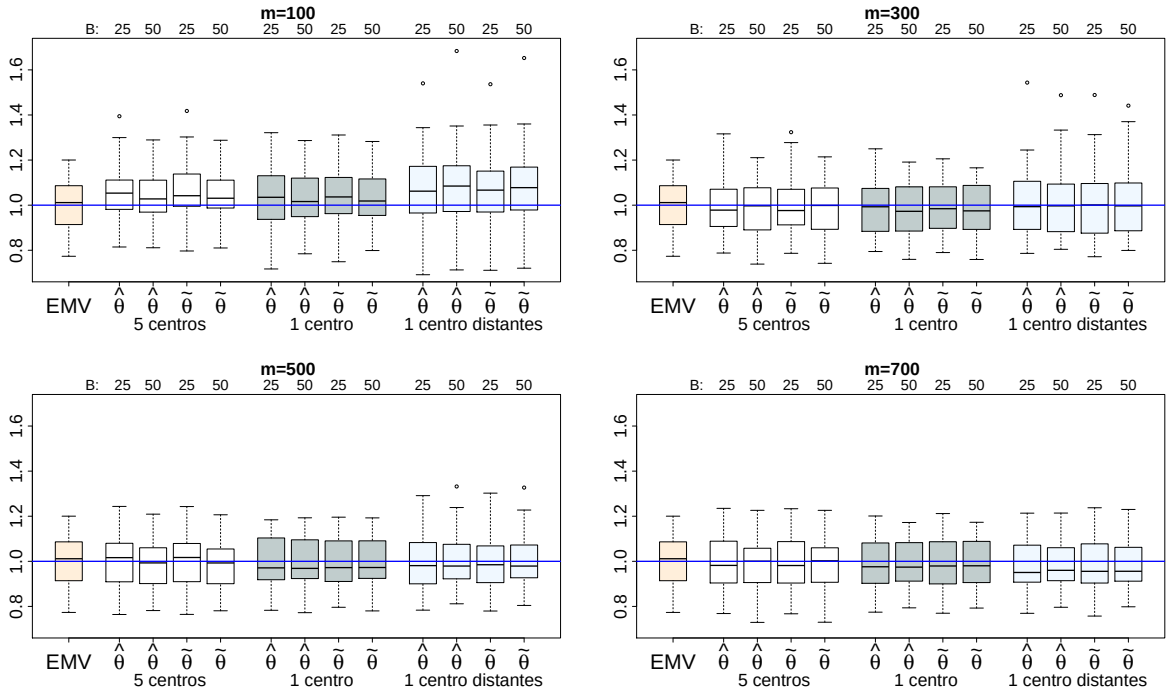

Figura 6.18: *Boxplot das estimativas de σ^2 para o cenário 4 dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= $\{5C, 1C, 1CD\}$, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Entre a razão ϕ/σ^2 observa-se uma diferença entre os métodos de seleção. Para uma dimensão de subamostra igual a 100, o algoritmo 1C apresenta os piores resultados, enquanto o método de seleção 5C apresenta menor vício e variabilidade (Tabela 6.3 e Figura 6.5). Para $m = 300$ e $m = 500$ não é possível visualizar diferença entre os diferentes algoritmos. Há redução de variabilidade quando aumentamos o valor de B, sendo essa característica mais evidente para $m = 100$ e $m = 300$.

O aumento da subamostra diminui a variabilidade das estimativas, para subamostra de tamanho 700, os estimadores *subsemble* são muito similares ao EMV.

Tabela 6.17: *Estimativas da razão $\phi/\sigma^2 = 5$ para o cenário 4, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	5.153 (1.417)	4.686 (0.802)	4.905 (0.803)	4.944 (0.828)
	50	5.131 (1.017)	4.820 (0.857)	4.895 (0.770)	4.971 (0.867)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	5.138 (1.403)	4.685 (0.781)	4.910 (0.809	4.943 (0.827)
	50	5.136 (1.053)	4.825 (0.855)	4.896 (0.767)	4.972 (0.868)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	4.012 (0.893)	4.592 (0.711)	4.805 (0.819)	4.856 (0.758)
	50	4.019 (0.829)	4.703 (0.707	4.830 (0.828	4.863 (0.728)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	4.036 (0.826)	4.602 (0.682)	4.828 (0.800	4.852 (0.755)
	50	4.051 (0.799)	4.702 (0.696)	4.840 (0.797)	4.858 (0.725)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	5.011 (1.844)	4.827 (1.147)	4.938 (0.833)	4.885 (0.731)
distantes	50	4.857 (1.547)	4.831 (0.947)	4.879 (0.757)	4.909 (0.714)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	5.063 (1.861)	4.844 (1.148)	4.911 (0.816)	4.892 (0.718)
distantes	50	4.892 (1.525)	4.844 (0.941)	4.875 (0.750)	4.906 (0.705)

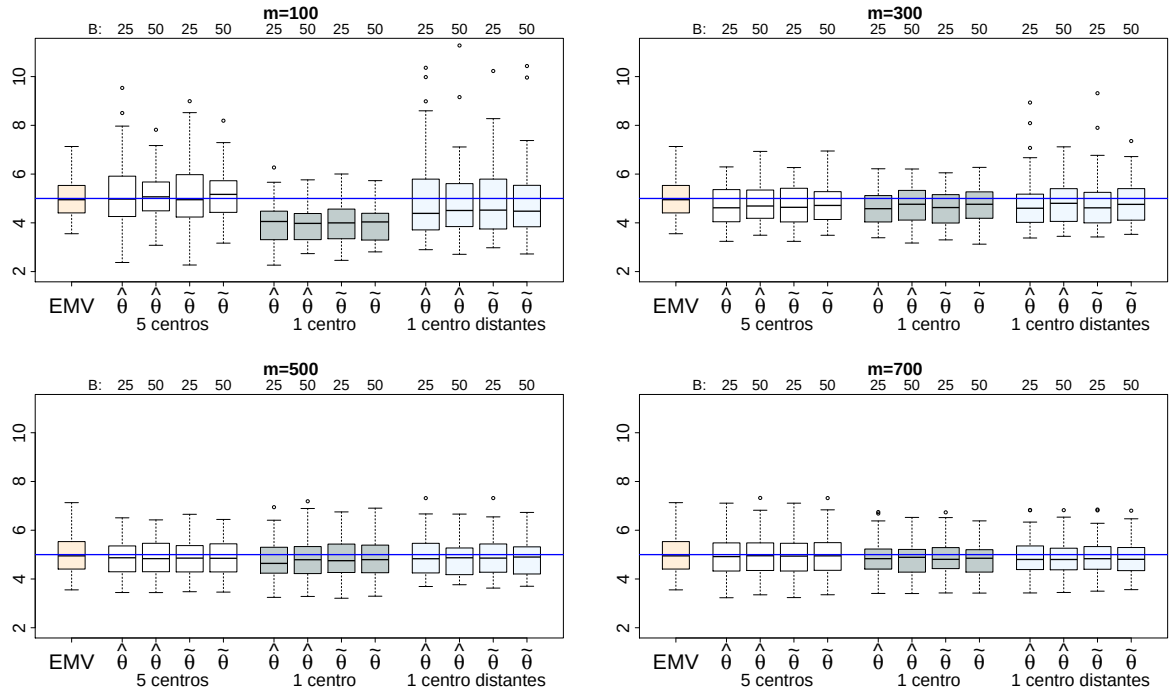

Figura 6.19: *Boxplot das estimativas de ϕ/σ^2 para o cenário 4 dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

Na Tabela 6.18 e Figura 6.20 temos os resultados de τ^2. A distribuição das estimativas se aproxima do EMV conforme o tamanho da subamostra aumenta. Para $m = 500$ e $m = 700$ a diferença entre os estimadores *subsemble* e o EMV é muito pequena. Sobre os diferentes métodos de seleção, 1CD possui a menor variabilidade, seguido por 1C e 5C. Quando $m = 700$, não é possível observar diferença entre as estimativas de $B = 25$ e $B = 50$.

Tabela 6.18: *Estimativas do parâmetro $\tau^2 = 1$ para o cenário 4, dado diferentes algoritmos de seleção, m e B.*

Estimador	B	$m = 100$	$m = 300$	$m = 500$	$m = 700$
$\hat{\boldsymbol{\theta}}$ 5 centros	25	0.852 (0.087)	0.953 (0.068)	0.979 (0.065)	0.988 (0.066)
	50	0.864 (0.087)	0.962 (0.072)	0.978 (0.062)	0.989 (0.067)
$\tilde{\boldsymbol{\theta}}$ 5 centros	25	0.851 (0.083)	0.954 (0.067)	0.979 (0.065)	0.988 (0.066)
	50	0.864 (0.085)	0.962 (0.072)	0.978 (0.062)	0.989 (0.067)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.837 (0.096)	0.949 (0.067)	0.976 (0.070)	0.987 (0.059)
	50	0.837 (0.089)	0.956 (0.065)	0.976 (0.066)	0.988 (0.060)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.841 (0.089)	0.951 (0.068)	0.978 (0.068)	0.986 (0.059)
	50	0.840 (0.085)	0.957 (0.066)	0.978 (0.065)	0.988 (0.059)
$\hat{\boldsymbol{\theta}}$ 1 centro	25	0.843 (0.078)	0.943 (0.076)	0.968 (0.062)	0.976 (0.058)
distantes	50	0.839 (0.075)	0.943 (0.068)	0.966 (0.058)	0.976 (0.055)
$\tilde{\boldsymbol{\theta}}$ 1 centro	25	0.844 (0.076)	0.943 (0.074)	0.965 (0.061)	0.976 (0.058)
distantes	50	0.840 (0.072)	0.943 (0.068)	0.965 (0.058)	0.976 (0.055)

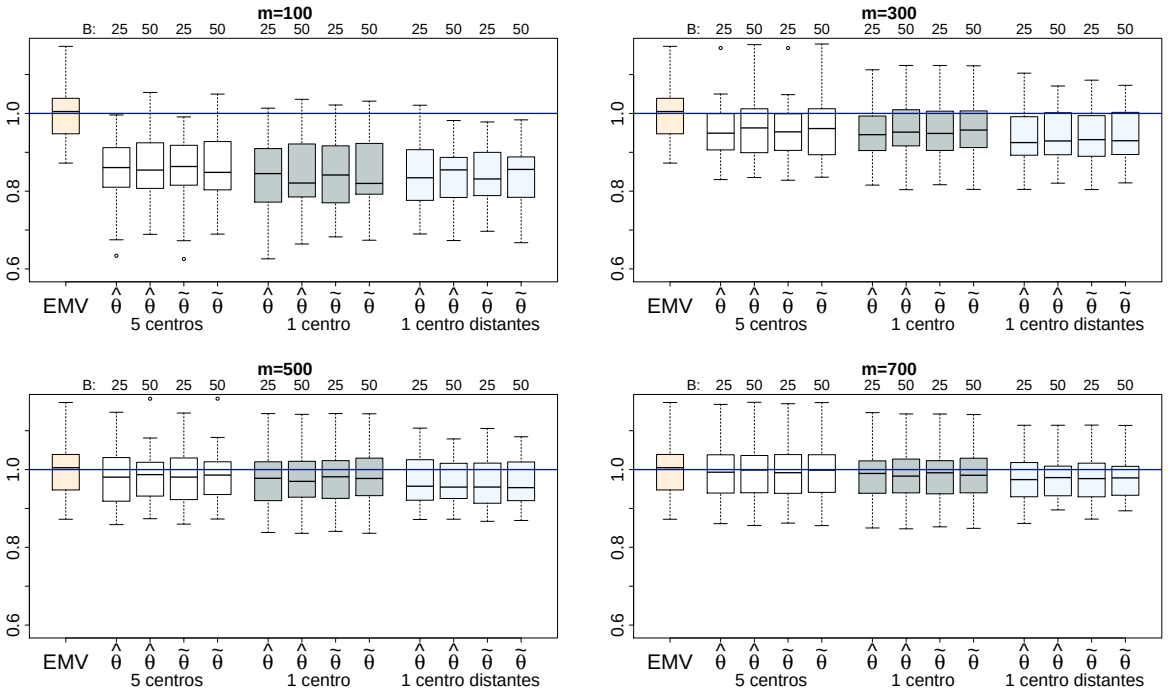

Figura 6.20: *Boxplot das estimativas de τ^2 para o cenário 4 dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$, métodos de seleção= {5C, 1C, 1CD}, tamanho da subamostra $m = \{100, 300, 500, 700\}$ e valores de $B = \{25, 50\}$.*

6.3.5 Comentários

Como era esperado, quanto maior a subamostra m melhor é a estimativa, sendo essa característica mais evidente para os parâmetros da função de covariância. Os diferentes valores de B influenciam pouco no vício e variância dos estimadores *subsemble* espacial, quando $B = 50$, possuem um pouco menos de variabilidade que para $B = 25$. Entretanto, é importante salientar que $B = 50$ necessita do dobro de tempo para o cálculo dos resultados.

Estimativas dos parâmetros β_0 e β_1 são muito parecidas com o EMV e, consequentemente, com os verdadeiros valores dos parâmetros. Essa interpretação é independente do cenário, algoritmo de seleção da subamostra e estimador $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$. Em relação a ϕ, σ^2, ϕ/σ^2 e τ^2, observa-se uma diferença entre os métodos de seleção, o algoritmo 5C e 1CD possuem um desempenho similar e superior ao 1C. Apesar disso, quando o tamanho da subamostra aumenta, essa diferença entre os três algoritmos diminui, tendendo a zero.

Em geral, pode-se dizer que os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ têm um desempenho semelhante nos cenários analisados.

6.4 Estudos Comparativos

Nesta seção compara-se os resultados dos estimadores *subsemble* com o EMV e RSA. Dado os resultados da seção 6.3 optou-se por mencionar somente os resultados do algoritmo 5 centros, visto que este algoritmo apresentou as melhores estimativas e uma tendência mais robusta dado o tamanho da subamostra. Entre os valores de *input* decidiu-se por $B = 25$, pois o ganho de acurácia para $B = 50$ não foi tão acentuado dado o aumento de tempo de processamento. O tamanho das subamostras utilizados nas estimativas dos estimadores RSA, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ são $m = \{100, 300, 500, 700\}$.

As tabelas desta seção irão exibir novamente a média e desvio-padrão das estimativas, para facilitar a análise dos resultados um índice superior aos estimadores indicará o tamanho da subamostra, por exemplo, RSA^{500} é o estimador RSA para uma subamostra de tamanho 500.

Com o objetivo de comparar a aderência da distribuição amostral dos estimadores RSA e

subsemble ao EMV, as densidades das estimativas de ϕ/σ^2 serão ilustradas para os cenários 1, 2 e 3. Para o cenário 4 a densidade de ϕ será mostrada individualmente. Os demais parâmetros terão suas densidade omitidas, pois acredita-se que elas não acrescentarão informação relevante sobre os diferentes estimadores.

Com a finalidade de verificar a qualidade da aproximação das equações 4.5 e 4.6, foram calculados os intervalos de confiança dos estimadores propostos e do EMV. Visto que o cenário 4 pode ser interpretado como um experimento observado no *increasing domain asymptotics*, é conhecido que os EMV são consistentes e sua distribuição assintótica é normal, tornando possível a obtenção de intervalos de confiança assintóticos para os parâmetros do modelo. Os cenários 1, 2 e 3 ainda não possuem comprovações matemáticas dessa propriedade, apesar disso é possível observar uma tendência à normalidade nos gráficos das densidades dos estimadores. Com o propósito de uma comparação mais completa com o EMV, os IC para o caso de *infill asymptotics* também serão calculados, embora tal aproximação não seja comprovada.

Além das estimativas, o tempo de processamento (em minutos) também será mensurado. Para os estimadores EMV e RSA somente as estimativas com 4 núcleos serão reportadas, pois em estudos simulados constatou-se que não houve diminuição no tempo de processamento quando mais núcleos eram utilizados, para os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ a duração das operações para *hardware* de 4 e 24 *cores* serão apresentadas. Como a obtenção dos estimadores propostos foram realizadas de forma conjunta, o tempo de processamento será mencionado somente para $\hat{\boldsymbol{\theta}}$.

6.4.1 Cenário 1

Através da Figura 6.21 e Tabela 6.19 pode-se observar que os estimativas de $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, RSA e o EMV para β_0 e β_1 possuem resultados muito similares, em que as médias são próximas ao verdadeiro valor do parâmetro. Adicionalmente, quando o tamanho da subamostra aumenta, a variância dos estimadores *subsemble* diminuem, se aproximando da variância do EMV. Para a razão ϕ/σ^2 e τ^2, todos os estimadores subestimam o verdadeiro valor do parâmetro, todavia, este vício diminui quando o tamanho da subamostra m aumenta. Novamente, os estimadores *subsemble* espacial apresentam variabilidade e vício inferiores ao RSA.

Tabela 6.19: *Média e desvio-padrão dos valores estimados do EMV, RSA, $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ para o cenário 1.*

Método	m	β_0	β_1	ϕ/σ^2	τ^2	4CPU(m)	24CPU(m)
$\boldsymbol{\theta}$	-	1.000	1.000	25.000	1.000	-	
MLE	-	0.990 (0.370)	0.990 (0.049)	24.055 (5.393)	0.998 (0.037)	33.720	-
	100	1.055 (0.401)	1.001 (0.078)	20.848 (6.109)	0.915 (0.058)	0.111	0.032
$\hat{\boldsymbol{\theta}}$	300	1.041 (0.401)	0.993 (0.055)	23.225 (4.947)	0.972 (0.043)	1.723	0.432
	500	1.055 (0.395)	0.988 (0.056)	23.388 (5.270)	0.984 (0.041)	9.295	1.529
	700	1.029 (0.383)	0.992 (0.050)	23.859 (5.363)	0.989 (0.039)	23.731	3.831
	100	1.055 (0.400)	0.979 (0.065)	20.699 (5.926)	0.918 (0.057)	-	-
$\tilde{\boldsymbol{\theta}}$	300	1.042 (0.399)	0.986 (0.057)	23.353 (5.133)	0.972 (0.044)	-	-
	500	1.057 (0.395)	0.996 (0.058)	23.478 (5.345)	0.984 (0.042)	-	-
	700	1.028 (0.382)	0.988 (0.057)	23.929 (5.396)	0.989 (0.039)	-	-
	100	1.037 (0.382)	0.992 (0.054)	18.727 (5.907)	0.937 (0.071)	0.163	-
RSA	300	1.020 (0.376)	0.990 (0.056)	21.822 (5.639)	0.980 (0.052)	2.960	-
	500	1.011 (0.375)	0.995 (0.055)	22.579 (5.304)	0.989 (0.047)	14.753	-
	700	1.004 (0.372)	0.991 (0.063)	23.052 (5.403)	0.995 (0.052)	40.784	-

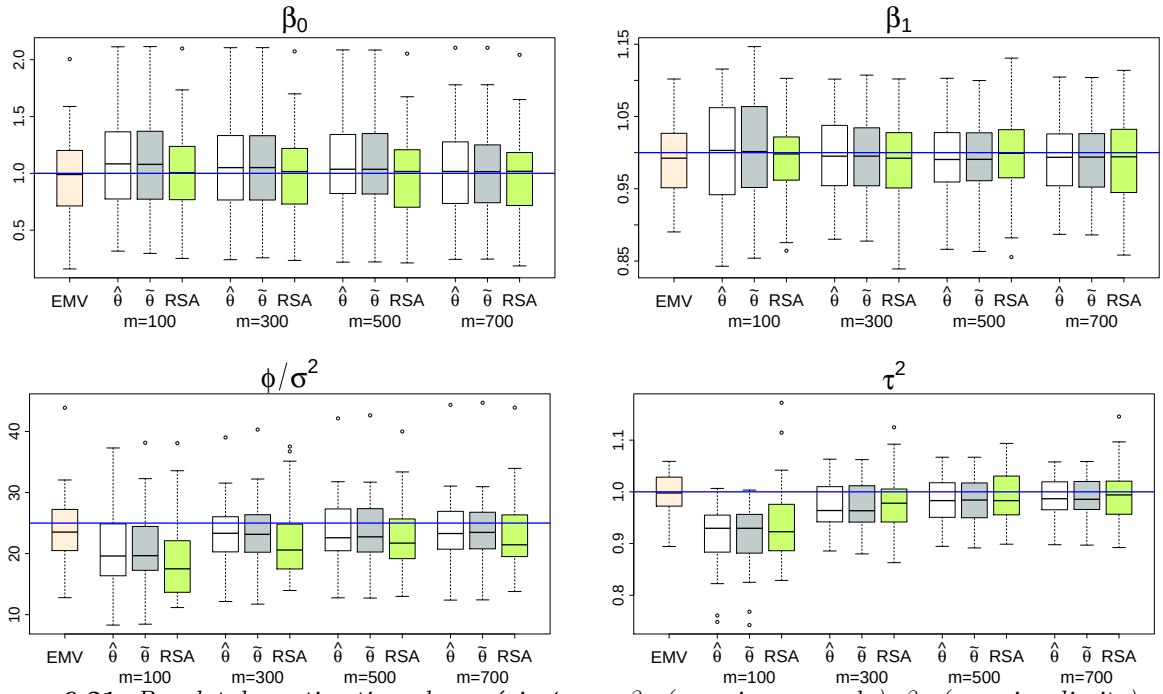

Figura 6.21: *Boxplot das estimativas do cenário 1 para β_0 (superior esquerdo), β_1 (superior direito), ϕ/σ^2 (inferior esquerdo) e τ^2 (inferior direito) dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}, RSA\}$ e tamanho da subamostra $m = \{100, 300, 500, 700\}$. A linha azul representa o verdadeiro valor do parâmetro.*

Calculou-se os intervalos de confiança (IC) do EMV e dos estimadores propostos, a Tabela 6.20 mostra que a diferença entre a amplitude dos intervalos é substancial, por exemplo, para ϕ/σ^2 e $m = 100$ o erro padrão do EMV é pelo menos 6,8 vezes menor que dos estimadores *subsemble* $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$. Todavia, quando o tamanho da subamostra aumenta, os erros padrões diminuem, ficando mais próximos à média dos erros padrões do EMV. Quando analisa-se a cobertura do intevalo, o EMV apresentou cobertura inferior ao nível de confiança para β_0, β_1, ϕ/σ^2, para os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ a cobertura foi igual ou superior a 95% para β_1, ϕ/σ^2 e τ^2.

Para comparar a distribuição amostral dos estimadores RSA e *subsemble* com o EMV, a Figura 6.22 apresentada as densidades das estimativas de ϕ/σ^2. É possível observar comportamento simétrico, com mediana próxima ao verdadeiro valor do parâmetro (linha azul). Note que, quando $m = 700$, quase não é possível visualizar diferença entre os métodos. Para menores valores de m, os estimadores propostos possuem um comportamento mais semelhante ao EMV.

Tabela 6.20: *Porcentagem de cobertura (% Cob) para intervalos com 95% de confiança e média do erro padrão (EP) para as estimativas do cenário 1.*

		β_0		β_1		ϕ/σ^2		τ^2	
Estimador	m	% Cob	EP	% Cob	EP	% Cob	EP	% Cob	EP
MLE	-	84%	0.339	92%	0.048	86%	4.528	96%	0.040
	100	88%	0.411	100%	0.217	100%	30.612	100%	0.218
$\hat{\boldsymbol{\theta}}$	300	88%	0.395	100%	0.124	100%	15.523	100%	0.110
	500	86%	0.400	100%	0.096	98%	11.164	100%	0.082
	700	86%	0.389	100%	0.081	96%	9.142	100%	0.068
	100	90%	0.410	100%	0.217	100%	39.836	100%	0.233
$\tilde{\boldsymbol{\theta}}$	300	88%	0.402	100%	0.124	100%	15.632	100%	0.108
	500	86%	0.400	100%	0.096	98%	11.326	100%	0.082
	700	86%	0.384	100%	0.081	96%	9.134	100%	0.068

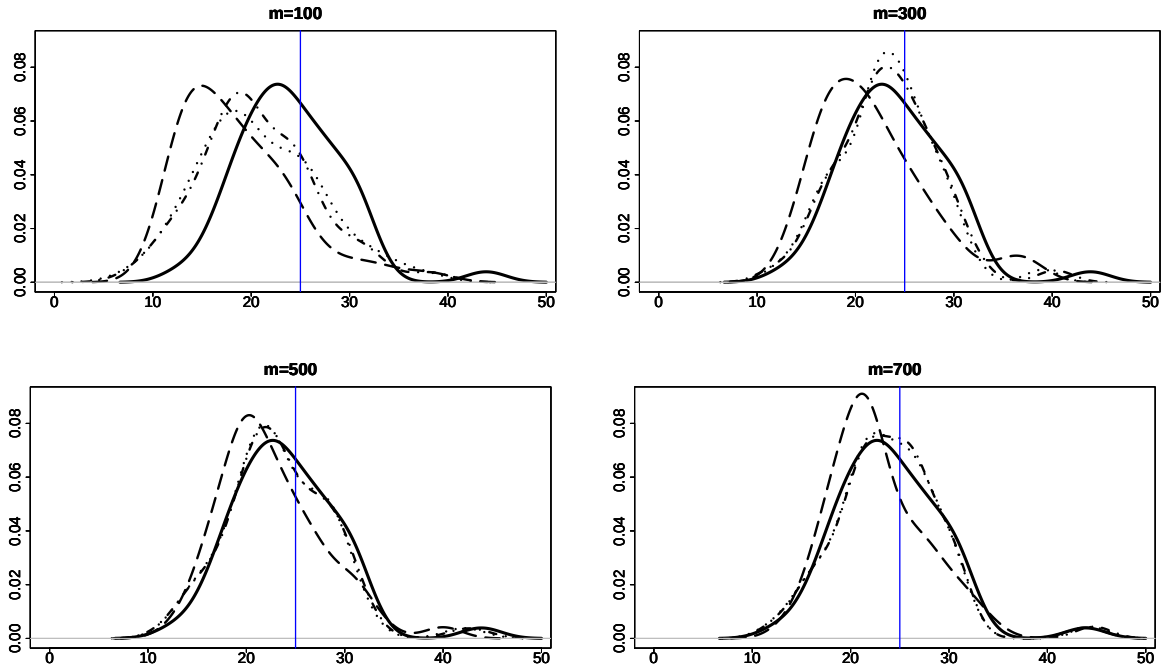

Figura 6.22: *Densidades das estimativas para o cenário 1 de ϕ/σ^2 do EMV(———), $\hat{\boldsymbol{\theta}}$ (·····), $\tilde{\boldsymbol{\theta}}$ (·–·–·) e RSA (– – ––) para subamostras de tamanho $m = 100$ (superior esquerdo), $m = 300$ (superior direito), $m = 500$ (inferior esquerdo) e $m = 700$ (inferior direito).*

6.4.2 Cenário 2

Por meio da Tabela 6.21 e Figura 6.23 pode-se observar que para o parâmetro β_0, os estimadores $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$ e EMV possuem resultados muito semelhantes. Independentemente do tamanho da subamostra, as estimativas apresentam um comportamento simétrico, com mediana muito próxima ao verdadeiro valor do parâmetro (linha azul). Além disso, quando o tamanho da subamostra aumenta, a variabilidade dos estimadores *subsemble* diminuem, e se aproximam dos desvios do EMV. O estimador RSA possui os piores resultados, principalmente para $m = 500$ e $m = 700$. Pode-se observar que as estimativas para β_1 apresentam médias próximas ao verdadeiro valor do parâmetro, independentemente do estimador. Entretanto, quando analisamos a variabilidade, o estimador RSA apresenta os piores resultados, enquanto para $m \leq 100$ os estimadores *subsemble* possuem o mesmo desvio-padrão do EMV.

Tabela 6.21: *Comparação das estimativas do EMV, RSA, $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ para o cenário 2.*

Método	m	β_0	β_1	ϕ/σ^2	τ^2	4CPU(m)	24CPU(m)
$\boldsymbol{\theta}$	-	1.000	1.000	25.000	0	-	-
MLE	-	0.981 (0.358)	0.998 (0.011)	25.608 (1.873)	0.001 (0.001)	33.923	-
	100	1.041 (0.399)	0.999 (0.014)	26.699 (2.129)	0.002 (0.001)	0.122	0.040
$\hat{\boldsymbol{\theta}}$	300	1.036 (0.394)	0.998 (0.011)	26.080 (2.008)	0.001 (0.001)	2.038	0.512
	500	1.046 (0.380)	0.998 (0.011)	25.873 (1.846)	0.001 (0.001)	8.831	1.822
	700	1.023 (0.374)	0.998 (0.011)	25.867 (1.885)	0.001 (0.001)	21.596	3.969
	100	1.048 (0.397)	0.998 (0.013)	26.768 (2.170)	0.002 (0.001)	-	-
$\tilde{\boldsymbol{\theta}}$	300	1.033 (0.394)	0.998 (0.011)	26.093 (1.973)	0.001 (0.001)	-	-
	500	1.041 (0.380)	0.998 (0.011)	25.851 (1.865)	0.001 (0.001)	-	-
	700	1.021 (0.371)	0.999 (0.011)	25.881 (1.871)	0.001 (0.001)	-	-
	100	0.962 (0.416)	0.994 (0.054)	28.818 (8.421)	0.043 (0.128)	0.182	-
RSA	300	0.866 (0.818)	0.989 (0.044)	26.042 (2.514)	0.005 (0.002)	3.826	-
	500	1.075 (1.212)	0.999 (0.028)	13.688 (13.290)	0.205 (0.229)	10.678	-
	700	1.212 (1.223)	0.996 (0.037)	9.192 (12.951)	0.260 (0.210)	33.042	-

Para a razão ϕ/σ^2 e τ^2, nota-se que os resultados do EMV, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$, RSA^{100} e RSA^{300}, são muito

similares ao verdadeiro valor do parâmetro. Esse comportamento não é observado para RSA^{500} e RSA^{700}, pois os valores das estimativas apresentam vício e alta variabilidade.

A Figura 6.24 ilustra as densidades das estimativas de ϕ/σ^2 do EMV, *subsemble* e RSA. É possível observar que os estimadores *subsemble* espacial possuem densidades mais parecidas com o EMV. Além disso, quase não é possível visualizar diferença entre o $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

Quando o efeito pepita é nulo, as distribuições das estimativas para τ^2 apresentam assimetria, apesar disso, os intervalos de confiança serão calculados. Através da Tabela 6.22, observa-se que conforme aumenta o tamanho da amostra, a média do erro padrão dos estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ aproximam-se do EMV.

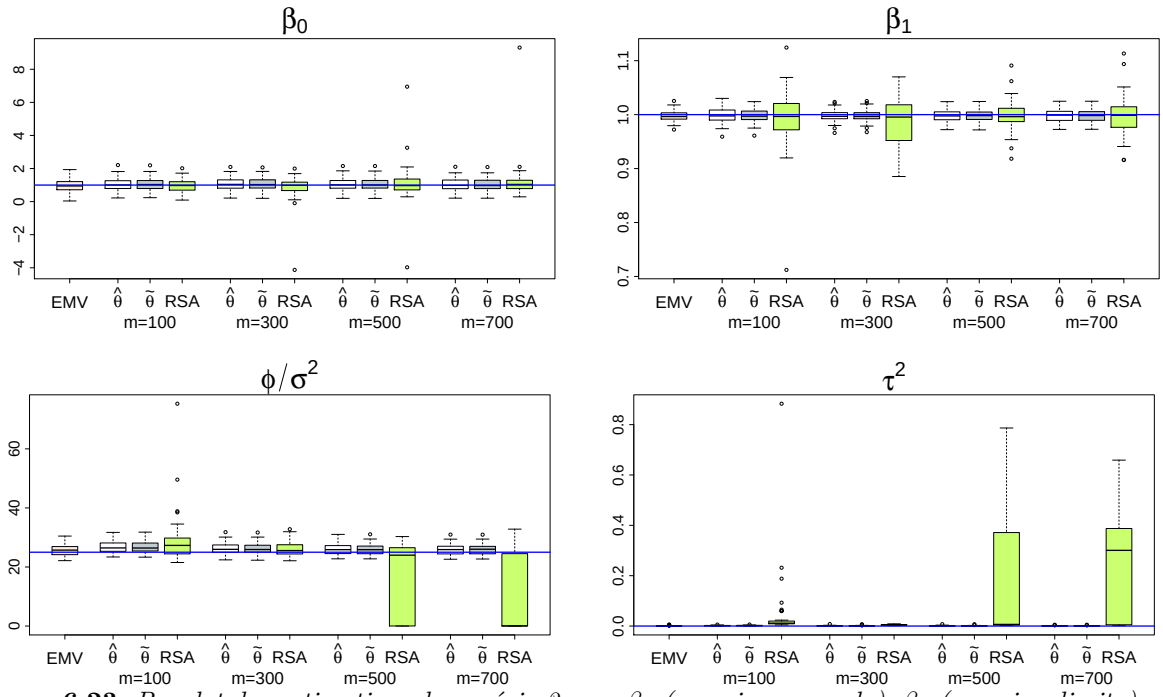

Figura 6.23: *Boxplot das estimativas do cenário 2 para β_0 (superior esquerdo), β_1 (superior direito), ϕ/σ^2 (inferior esquerdo) e τ^2 (inferior direito) dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}, RSA\}$ e tamanho da subamostra $m = \{100, 300, 500, 700\}$. A linha azul representa o verdadeiro valor do parâmetro.*

Tabela 6.22: *Porcentagem de cobertura (% Cob) para intervalos com 95% de confiança e média do erro padrão (EP) para as estimativas do cenário 2.*

		β_0		β_1		ϕ/σ^2		τ^2	
Estimador	m	% Cob	EP	% Cob	EP	% Cob	EP	% Cob	EP
MLE	-	88%	0.323	90%	0.009	70%	1.052	94%	0.001
$\hat{\boldsymbol{\theta}}$	100	92%	0.421	100%	0.043	100%	7.121	100%	0.006
	300	92%	0.406	100%	0.024	100%	3.427	100%	0.003
	500	92%	0.399	100%	0.018	98%	2.509	100%	0.002
	700	90%	0.388	100%	0.015	94%	2.008	100%	0.002
$\tilde{\boldsymbol{\theta}}$	100	94%	0.429	100%	0.043	100%	7.306	98%	0.002
	300	90%	0.415	100%	0.024	100%	3.483	100%	0.002
	500	92%	0.403	100%	0.018	98%	2.502	100%	0.002
	700	90%	0.394	100%	0.015	94%	2.008	98%	0.002

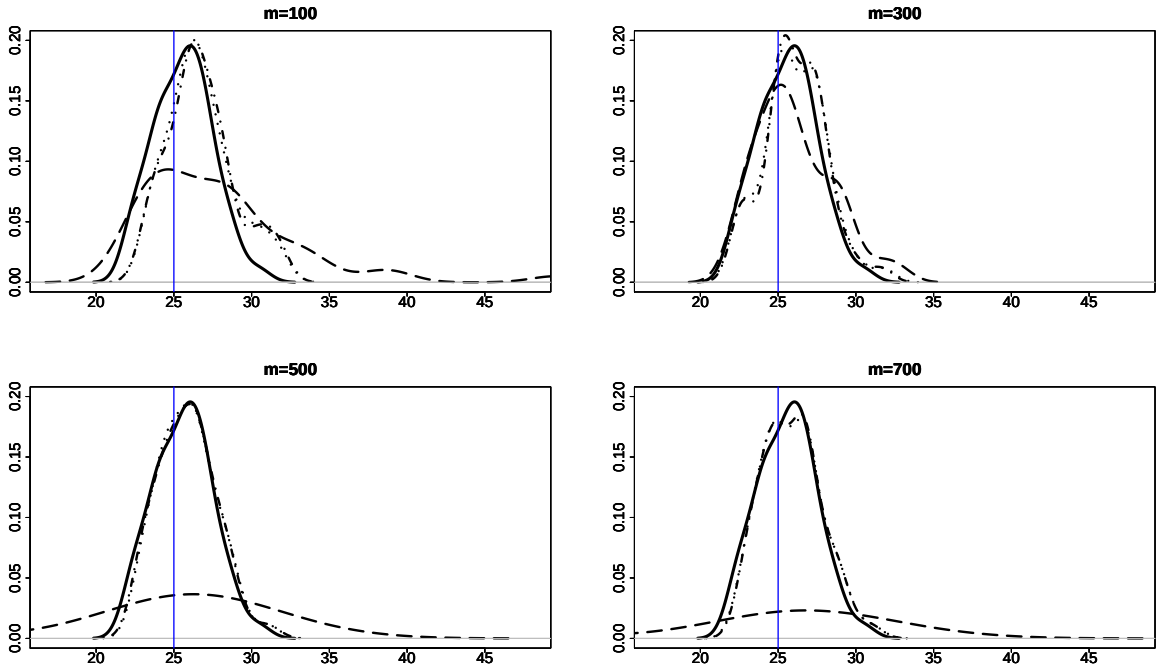

Figura 6.24: *Densidades das estimativas para o cenário 2 de ϕ/σ^2 do EMV(———), $\hat{\boldsymbol{\theta}}$ (· · · · ·), $\tilde{\boldsymbol{\theta}}$ (· − · − ·) e RSA (− − −−) para subamostras de tamanho $m = 100$ (superior esquerdo), $m = 300$ (superior direito), $m = 500$ (inferior esquerdo) e $m = 700$ (inferior direito).*

6.4.3 Cenário 3

A Figura 6.25 e Tabela 6.23 mostram os resultados do cenário 3, este cenário difere do cenário 1 somente pelo tamanho do banco de dados. Os resultados são qualitativamente similares àqueles do cenário 1, a diferença é que o maior tamanho da amostra ressaltou características já observadas. Mais um vez, nota-se que os estimadores RSA e *subsemble* possuem um comportamento semelhante quando estimam o parâmetro β_0. Todos apresentam pouco vício e variabilidade, também, a qualidade das estimativas não é afetada pelo valor de m. Para o parâmetro β_1, a variância das estimativas de $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ são influenciadas pela dimensão da subamostra, apresentando menor variabilidade quando mais dados são considerados na estimação.

Tabela 6.23: *Comparação das estimativas do RSA, $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ para o cenário 3.*

Método	m	β_0	β_1	ϕ/σ^2	τ^2	4CPU(m)	24CPU(m)
$\boldsymbol{\theta}$	-	1.000	1.000	25.000	1	-	
$\hat{\boldsymbol{\theta}}$	100	0.902 (0.458)	0.990 (0.046)	20.424 (4.835)	0.930 (0.038)	0.165	0.047
	300	0.893 (0.454)	0.996 (0.023)	25.164 (4.421)	0.981 (0.021)	2.143	0.414
	500	0.889 (0.456)	1.003 (0.020)	25.060 (4.000)	0.991 (0.016)	10.690	1.604
	700	0.914 (0.473)	0.996 (0.016)	25.324 (4.313)	0.990 (0.011)	25.827	4.119
$\tilde{\boldsymbol{\theta}}$	100	0.900 (0.456)	0.989 (0.046)	20.448 (4.732)	0.929 (0.039)	-	-
	300	0.894 (0.454)	0.995 (0.022)	25.119 (4.416)	0.981 (0.021)	-	-
	500	0.887 (0.455)	1.003 (0.020)	24.995 (4.024)	0.991 (0.017)	-	-
	700	0.915 (0.474)	0.995 (0.016)	25.305(4.179)	0.990 (0.011)	-	-
RSA	100	0.898 (0.441)	0.999 (0.028)	19.036 (6.001)	0.930 (0.068)	0.256	-
	300	0.905 (0.423)	0.998 (0.031)	21.895 (5.504)	0.981 (0.051)	3.244	-
	500	0.907 (0.416)	0.999 (0.027)	22.824 (4.920)	0.983 (0.039)	15.525	-
	700	0.909 (0.412)	1.007 (0.031)	23.340 (4.687)	0.989 (0.038)	43.244	-

A razão ϕ/σ^2 é melhor estimada por $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$, com essa característica mais evidente quando $m \geq 300$, o RSA apresentou maior vício e variabilidade. Com respeito ao efeito pepita τ^2, observa-se que todos os estimadores subestimam o verdadeiro valor do parâmetro. Mas, esse vício tende a

diminuir, conforme aumenta o tamanho da subamostra. A variância também diminui para maiores tamanho de subamostra, especialmente para os estimadores *subsemble* espacial.

A Figura 6.26 apresenta as densidades das estimativas de ϕ/σ^2, nota-se que os estimadores *subsemble* apresentam um comportamento simétrico para todos os tamanhos de subamostra, enquanto o estimador RSA possui assimetria, que diminui com o aumento do número de observações consideradas na estimação. Como não foi possível calcular o EMV, a Tabela 6.24 possui somente os resultados para $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$, é possível observar que o erro padrão dos estimadores diminui conforme o tamanho da subamostra aumenta. Sobre a cobertura do intervalo, as estimativas de β_0 possuem cobertura inferior àquela especificada pelo nível de confiança, para os demais parâmetros, a porcentagem é de 100%.

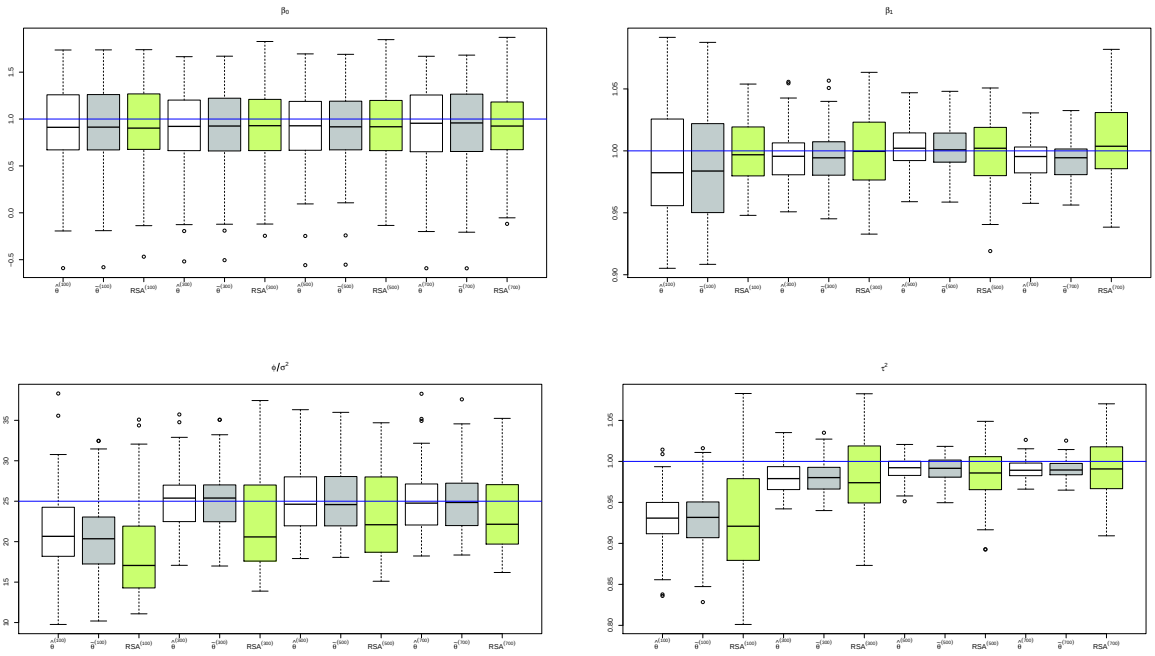

Figura 6.25: *Boxplot das estimativas do cenário 3 para β_0 (superior esquerdo), β_1 (superior direito), ϕ/σ^2 (inferior esquerdo) e τ^2 (inferior direito) dado diferentes estimadores $\{\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}, RSA\}$ e tamanho da subamostra $m = \{100, 300, 500, 700\}$. A linha azul representa o verdadeiro valor do parâmetro.*

Tabela 6.24: *Porcentagem de cobertura (% Cob) para intervalos com 95% de confiança e média do erro padrão (EP) para as estimativas do cenário 3.*

		β_0		β_1		ϕ/σ^2		τ^2	
Estimador	m	% Cob	EP	% Cob	EP	% Cob	EP	% Cob	EP
$\hat{\boldsymbol{\theta}}$	100	92%	0.414	100%	0.208	100%	43.754	100%	0.172
	300	90%	0.415	100%	0.119	100%	29.213	100%	0.091
	500	88%	0.420	100%	0.092	100%	21.873	100%	0.070
	700	86%	0.410	100%	0.078	100%	17.799	100%	0.058
$\tilde{\boldsymbol{\theta}}$	100	92%	0.412	100%	0.208	100%	43.992	100%	0.070
	300	90%	0.416	100%	0.119	100%	29.089	100%	0.070
	500	88%	0.421	100%	0.092	100%	21.630	100%	0.070
	700	90%	0.409	100%	0.078	100%	17.885	100%	0.058

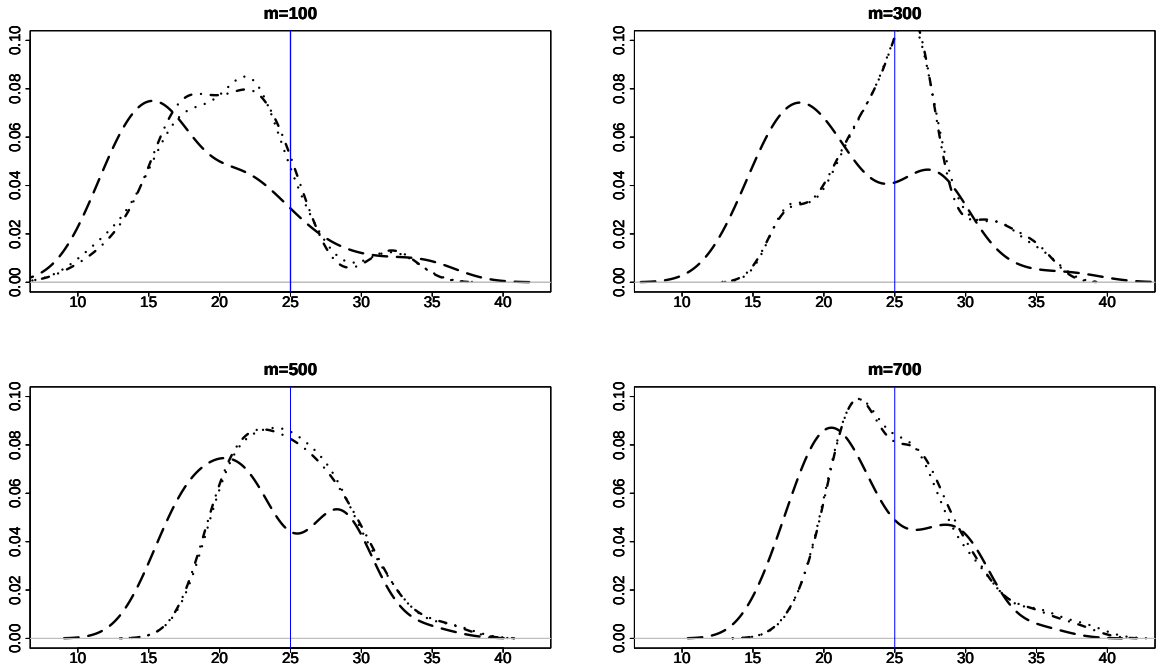

Figura 6.26: *Densidades das estimativas para o cenário 3 de* ϕ/σ^2 *do* $\hat{\boldsymbol{\theta}}$ *(·····),* $\tilde{\boldsymbol{\theta}}$ *(·−·−·) e RSA (−−−−) para subamostra* $m = 100$ *(superior esquerdo),* $m = 300$ *(superior direito),* $m = 500$ *(inferior esquerdo) e* $m = 700$ *(inferior direito).*

6.4.4 Cenário 4

Tabela 6.25: *Comparação das estimativas do EMV, RSA,* $\tilde{\boldsymbol{\theta}}$ *e* $\hat{\boldsymbol{\theta}}$ *para o cenário 4.*

Método	m	β_0	β_1	ϕ	σ^2	ϕ/σ^2	τ^2	4CPU(m)	24CPU(m)
$\boldsymbol{\theta}$	-	1	1	5	1	5	1 -	-	
MLE	-	0.990	0.983	4.932	1.001	4.976	0.999	36.050	-
		(0.103)	(0.045)	(0.703)	(0.115)	(0.814)	(0.063)		
	100	0.989	0.985	5.338	1.056	5.153	0.852	0.139	0.046
		(0.128)	(0.064)	(1.228)	(0.126)	(1.417)	(0.087)		
$\hat{\boldsymbol{\theta}}$	300	0.989	0.987	4.605	0.994	4.686	0.953	1.645	0.598
		(0.127)	(0.051)	(0.662)	(0.122)	(0.802)	(0.068)		
	500	0.989	0.978	4.821	0.992	4.905	0.979	6.153	2.150
		(0.114)	(0.051)	(0.693)	(0.114)	(0.803)	(0.065)		
	700	0.977	0.983	4.841	0.990	4.944	0.988	15.609	4.549
		(0.112)	(0.049)	(0.705)	(0.124)	(0.828)	(0.066)		
	100	0.991	0.986	5.345	1.061	5.138	0.851	-	-
		(0.130)	(0.064)	(1.191)	(0.128)	(1.403)	(0.083)		
$\tilde{\boldsymbol{\theta}}$	300	0.988	0.987	4.611	0.996	4.685	0.954	-	-
		(0.125)	(0.051)	(0.650)	(0.122)	(0.781)	(0.067)		
	500	0.988	0.978	4.825	0.992	4.910	0.979	-	-
		(0.113)	(0.051)	(0.697)	(0.114)	(0.809)	(0.065)		
	700	0.977	0.983	4.841	0.991	4.943	0.988	-	-
		(0.112)	(0.049)	(0.706)	(0.124)	(0.827)	(0.066)		
	100	0.991	0.984	4.671	1.022	4.699	0.976	0.164	-
		(0.113)	(0.058)	(0.988)	(0.988)	(1.351)	(0.110)		
RSA	300	0.987	0.981	4.746	1.015	4.761	0.982	2.942	-
		(0.114)	(0.057)	(0.865)	(0.865)	(1.111)	(0.085)		
	500	0.989	0.980	4.773	1.004	4.816	0.989	14.755	-
		(0.111)	(0.056)	(0.779)	(0.779)	(0.948)	(0.085)		
	700	0.987	0.984	4.828	1.005	4.861	0.992	41.883	-
		(0.107)	(0.057)	(0.744)	(0.744)	(0.917)	(0.080)		

Os resultados da Tabela 6.25 e Figura 6.27 mostram que as estimativas para β_0, β_1, ϕ, σ^2 e ϕ/σ^2 se assemelham ao verdadeiro valor do parâmetro, independentemente do estimador e tamanho da subamostra. Além disso, é possível observar que os desvios-padrões das estimativas do RSA e estimadores *subsemble* possuem valores muito parecidos com o EMV. Essa interpretação é confirmada pela Figura 6.28, em que para $m = 700$ não é possível diferenciar as densidades dos valores estimados para ϕ.

Para o parâmetro τ^2, os resultados para subamostras de tamanho $m = 100$ e $m = 300$ subestimaram o verdadeiro valor do parâmetro. Quando $m = 500$ e 700, os estimadores RSA, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ possuem resultados muito similares ao EMV, não sendo possível identificar a superioridade de nenhuma metodologia.

Através da Tabela 6.26, observa-se que conforme o valor m aumenta, o erro padrão dos estimadores *subsemble* aproximam-se do EMV. Sobre a cobertura dos intervalos, o EMV apresentou cobertura inferior ao nível de confiança para σ^2, ϕ/σ^2 e τ^2. Os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ possuem cobertura de 100% para todos os parâmetros e valores de m.

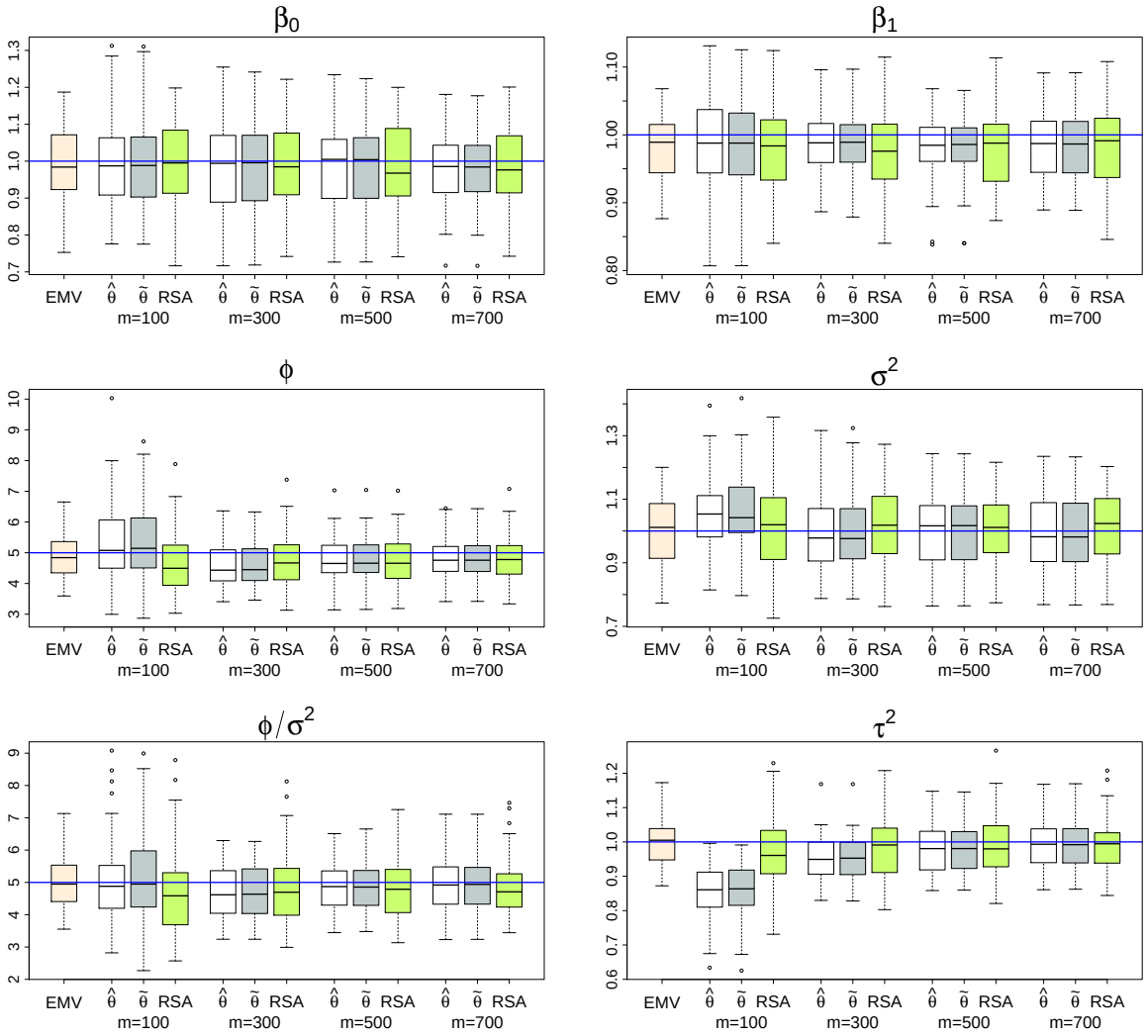

Figura 6.27: *Boxplot das estimativas do cenário 4 para β_0 (superior esquerdo), β_1 (superior direito), ϕ (centro esquerdo), σ^2 (centro direito), ϕ/σ^2 (inferior esquerdo) e τ^2 (inferior direito) dado diferentes estimadores $\{EMV, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}, RSA\}$ e tamanho da subamostra $m = \{100, 300, 500, 700\}$. A linha azul representa o verdadeiro valor do parâmetro.*

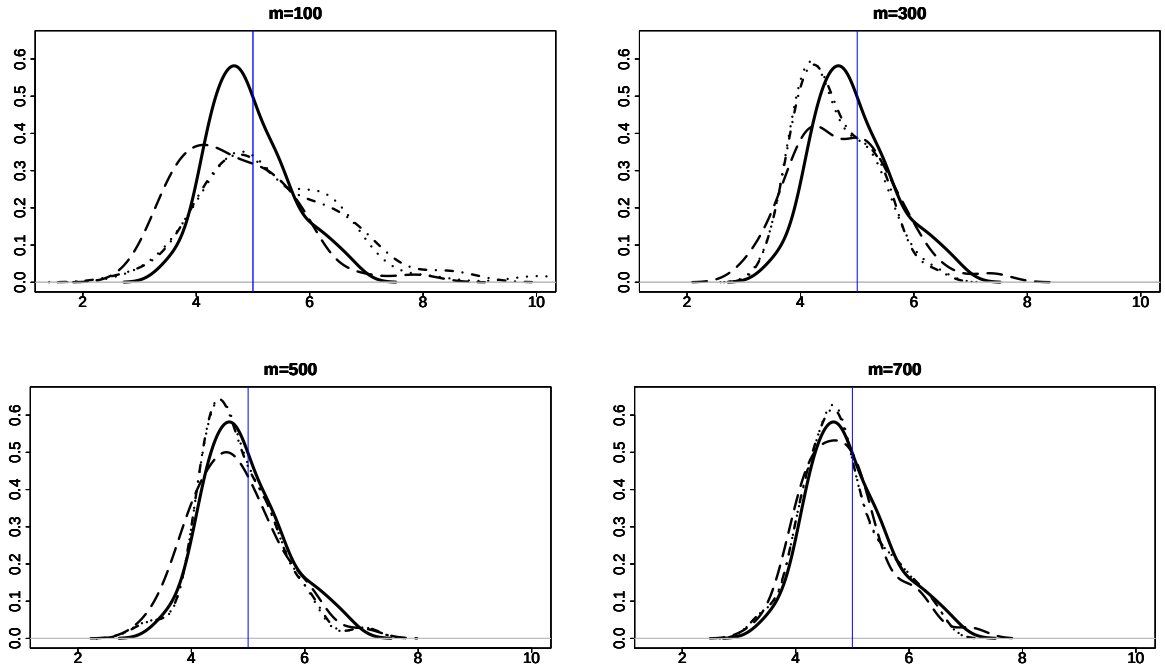

Figura 6.28: *Densidades para o cenário 4 de ϕ dos EMV(———), $\hat{\boldsymbol{\theta}}$ (· · · · ·), $\tilde{\boldsymbol{\theta}}$ (· − · − ·) e RSA (− − − −) para subamostra $m = 100$ (superior esquerdo), $m = 300$ (superior direito), $m = 500$ (inferior esquerdo) e $m = 700$ (inferior direito).*

Tabela 6.26: *Porcentagem de cobertura (% Cob) para intervalos com 95% de confiança e média do erro padrão (EP) para as estimativas do cenário 4.*

		β_0		β_1		ϕ		σ^2		ϕ/σ^2		τ^2	
Método	m	% Cob	EP	% Cob	EP	% Cob	EP	% Cob	EP	% Cob	EP	Cob	EP
MLE	-	96%	0.115	98%	0.053	96%	0.777	92%	0.110	94%	0.843	92%	0.061
	100	100%	0.295	100%	0.241	100%	8.917	100%	0.598	100%	9.565	100%	0.458
$\hat{\boldsymbol{\theta}}$	300	100%	0.219	100%	0.138	100%	2.192	100%	0.273	100%	2.479	100%	0.171
	500	100%	0.186	100%	0.107	100%	1.609	100%	0.210	100%	1.852	100%	0.128
	700	100%	0.168	100%	0.090	100%	1.336	100%	0.180	100%	1.510	100 %	0.106
	100	100%	0.296	100%	0.241	100%	8.966	100%	0.583	100%	9.598	80%	0.127
$\tilde{\boldsymbol{\theta}}$	300	100%	0.219	100%	0.138	100%	2.197	100%	0.273	100%	2.479	100%	0.127
	500	100%	0.187	100%	0.107	100%	1.612	100%	0.211	100%	1.847	100%	0.127
	700	100%	0.168	100%	0.090	100%	1.336	100%	0.191	100%	1.508	100%	0.106

6.4.5 Comentários

Para todos os cenários, o tempo necessário para a estimação dos estimadores *subsemble* foram inferiores à do estimador RSA, essa diferença torna-se mais evidente quando utiliza-se 24 *cores*. Esses resultados indicam que os estimadores propostos são mais rápidos que o estimador RSA, caraterística muito importante quando lidamos com grandes conjuntos de dados.

Sobre a qualidade das estimativas, pode-se concluir que os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ apresentaram, de modo geral, estimativas mais precisas para todos os cenários, com comportamento mais semelhante ao EMV. Essa característica é reforçada pelas densidades das estimativas de ϕ/σ^2 (cenários 1, 2, 3) e ϕ (cenário 4) dos parâmetros da função de covariância.

Para o cenário 2, o estimador RSA não apresentou bons resultados quando $m = 500$ e $m = 700$. Acredita-se que essa tendência pode ser devida à escolha dos valores iniciais de otimização do algoritmo (a_t e b_t).

Os intervalos de confiança dos estimadores propostos apresentaram cobertura superior ao nível de confiança. Os estimadores $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ também estimaram erros padrões com valores superiores ao EMV, sendo essa diferença mais acentuada para menores valores de m.

Capítulo 7

Análise de Dados Reais

Existem inúmeros exemplos de bancos de dados georreferenciados, entre as várias opções considera-se um conjunto de dados composto por estações de monitoramento do clima, localizadas nos Estados Unidos, os dados estão disponíveis no site do NOAA e são muito conhecidos na literatura, Kaufman *et al.* (2008), Furrer *et al.* (2006) e Liang *et al.* (2013) são alguns exemplos de artigos que utilizaram essas observações para ilustrar a eficácia de suas metodologias.

Os dados que serão utilizados na análise possuem 11.918 observações, que datam de abril de 1948. A variável de interesse é o total mensal de valores incomuns de precipitação, que são definidas como o total mensal de precipitação padronizadas pela média e desvio padrão de cada estação. Assim como Liang *et al.* (2013), os dados foram divididos aleatoriamente em 2 partes, uma parte é composta por 11.000 observações e é utilizada para estimar os parâmetros do modelo (Figura 7.1). O outro pedaço, composto por 918 estações, quantifica a qualidade de predição, através do erro quadrático médio (Figura 7.2).

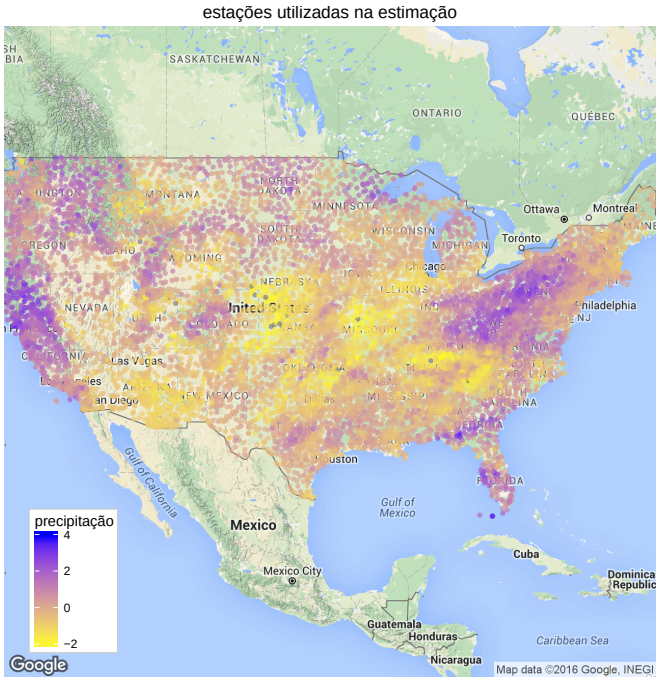

Figura 7.1: *Localização das 11.000 estações de monitoramento utilizadas para estimar os parâmetros do modelo.*

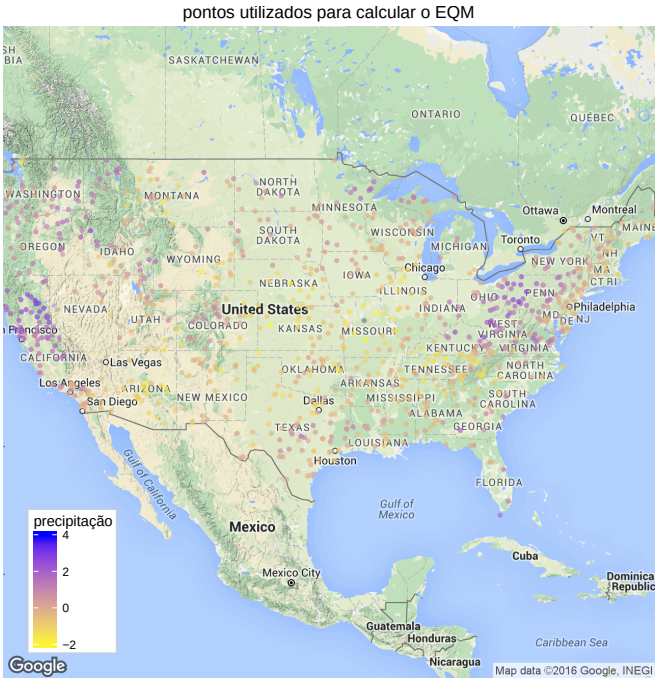

Figura 7.2: *Localização das 918 estações de monitoramento utilizadas no cálculo do erro quadrático médio.*

As funções de covariância utilizadas nas estimativas foram a exponencial e Matérn, com mesmo tamanho de subamostras $m = \{100, 300, 500, 700\}$ e $B = 25$. Como o modelo gerador dos dados é desconhecido, a qualidade das estimativas dos estimadores baseados em subamostras serão comparadas com o EMV.

Os valores dos parâmetros estimados utilizando a função de covariância exponencial são apresentados na Tabela (7.1). Os resultados indicam que, tanto o estimador RSA quanto os estimadores $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ estimam de modo satisfatório a razão dos parâmetros ϕ/σ^2 do modelo. Se consideramos ϕ e σ^2 individualmente, os estimadores $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ possuem valores muito semelhantes ao EMV.

Quando comparamos o tempo de processamento dos resultados, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ demandam menos tempo que o RSA. Para *hardware* composto por 4 núcleos, a diferença é de alguns segundos para subamostras $m = 100$ e $m = 300$, e de minutos para $m = 500$ e $m = 700$. Se o número de *cores* aumenta para 24, a diferença torna-se muito maior, por exemplo, para $m = 700$ os estimadores propostos são 8 vezes mais rápidos que o estimador RSA e 841 vezes mais rápidos que o EMV.

Para calcular os valores preditos $\hat{\mathbf{Y}}$ para as 918 estações é necessário inverter uma matriz de tamanho 11000×11000, o que demanda muita memória e tempo de processamento. Como alter-

Tabela 7.1: *Comparação das estimativas do EMV, RSA, $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ para os dados de anomalias de precipitação. A função de covariância é a exponencial.*

Método	m	β_0	ϕ	σ^2	ϕ/σ^2	τ^2	4CPU(m)	24CPU(m)
MLE	-	0.256	9.999	2.966	3.371	0.049	5634.772	-
	100	0.089	7.130	2.411	2.957	0.045	0.168	0.069
$\hat{\boldsymbol{\theta}}$	300	0.084	9.968	3.239	3.077	0.041	2.648	0.638
	500	0.343	12.046	3.652	3.298	0.053	9.515	2.057
	700	0.334	17.962	5.585	3.209	0.050	45.824	6.683
	100	0.192	7.672	2.700	2.841	0.044	-	-
$\tilde{\boldsymbol{\theta}}$	300	0.063	8.931	3.048	2.93	0.040	-	-
	500	0.265	11.528	3.483	3.309	0.053	-	-
	700	0.214	17.415	5.333	3.216	0.050	-	-
	100	0.149	3.290	0.876	3.756	0.081	0.164	-
RSA	300	0.159	3.160	0.803	3.935	0.058	2.937	-
	500	0.147	2.901	0.821	3.533	0.057	15.473	-
	700	0.150	2.848	3.440	0.828	0.055	53.331	-

Tabela 7.2: *Comparação do EQM de predição do EMV, RSA, $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ para as estações de monitoramento localizadas na Figura 7.2. A função de covariância utilizada é a exponencial.*

		nº vizinhos			
Método	m	25	50	100	150
EMV	-	0.0979	0.0976	0.0977	0.0978
$\hat{\boldsymbol{\theta}}$	100	0.0976	0.0972	0.0974	0.0974
	300	0.0989	0.0985	0.0988	0.0988
	500	0.098	0.0977	0.0979	0.0979
	700	0.098	0.0977	0.0979	0.0979
$\tilde{\boldsymbol{\theta}}$	100	0.0975	0.0972	0.0973	0.0973
	300	0.0975	0.0971	0.0973	0.0973
	500	0.0980	0.0977	0.0979	0.0979
	700	0.0980	0.0976	0.0978	0.0979
RSA	100	0.1002	0.0998	0.1002	0.1002
	300	0.0989	0.0985	0.0988	0.0988
	500	0.0984	0.0981	0.0983	0.0983
	700	0.0982	0.0979	0.0981	0.0981

nativa, Liang *et al.* (2013) consideram somente as observações mais próximas dos pontos a serem preditos, diminuindo o tamanho da matriz de covariância necessária para calcular a krigagem. Ao contrário de Liang *et al.* (2013), em vez de considerar um raio de distância entre os pontos, utilizou-se a quantidade de vizinhos para calcular a matriz de covariância do estimador BLUP, permitindo conhecer *a priori* sua dimensão. Para cada uma das 918 estações, a krigagem foi realizada considerando somente os 25, 50, 100 e 150 vizinhos mais próximos de cada uma das 918 estações da Figura 7.2.

Pela Tabela 7.2 é possível notar que a diferença entre a quantidade de vizinhos considerados no cálculo do preditor não é determinante na qualidade das estimativas de $\hat{\mathbf{Y}}$. Entretando, um comportamento incomum pode ser observado: os menores valores do EQM são observados quando o tamanho da matriz de covariância utilizada no cálculo do BLUP é 50×50, segundo Liang *et al.* (2013), a causa pode ser a má especificação da função de covariância ou a não estacionariedade dos dados.

Os valores dos EQM mantêm-se muito parecidos para cada estimador, porém é possível apontar que o estimador $\tilde{\boldsymbol{\theta}}$ apresentou os menores valores para o EQM, seguido do $\hat{\boldsymbol{\theta}}$, EMV e RSA, respectivamente. Logo, pode-se afirmar que o estimador *subsemble* espacial é eficaz quando desejamos fazer predição, além de possuir um desempenho superior ao RSA.

A Figura 7.3 mostra as superfícies de predição do EMV, $\hat{\boldsymbol{\theta}}^{700}$, $\tilde{\boldsymbol{\theta}}^{700}$ e RSA^{700}, respectivamente. A superfície de predição foi calculada em um grid regular com 10.430 observações e o número de vizinhos utilizados para gerar a matriz de covariância da predição foi igual à 25. Comparando as imagens, não é visível a diferença entre os estimativas, podemos afirmar que as predições baseadas em subamostras são muito parecidas com a superfície gerada pelo EMV.

A fim de mostrar a flexibilidade do estimador *subsemble* espacial, as estimativas para a função de covariância Matérn também foram calculadas $\left(\rho(||\mathbf{s}_i - \mathbf{s}_j||; \kappa, \phi) = 2^{\kappa-1}\Gamma(\kappa)^{-1}\left(\frac{||s_i - s_j||}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{||s_i - s_j||}{\phi}\right)\right)$. Como a função Matérn possui mais parâmetros, a maximização da função de verossimilhança torna-se mais complexa, dificultando a obtenção de $\hat{\boldsymbol{\theta}}_i$ para subamostras menores. Por esse motivo, não foi possível calcular o estimador *subsemble* para $m = 100$. Consequentemente, para geração dos resultados são consideradas subamostras de tamanho 300, 500 e 700. O número de repetições B assume novamente o valor 25.

Pela Tabela 7.3, nota-se que os estimadores *subsemble* apresentam resultados muito similares ao EMV. Para o parâmetro de suavização κ, o valor de m exerce maior influência nos resultados, parece ser necessário um maior tamanho de subamostra para captar o comportamento do parâmetro de suavização, ademais, é possível observar que $\tilde{\boldsymbol{\theta}}$ possui estimativas de κ mais próximas ao EMV. Em relação aos demais parâmetros, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$ são muito semelhantes.

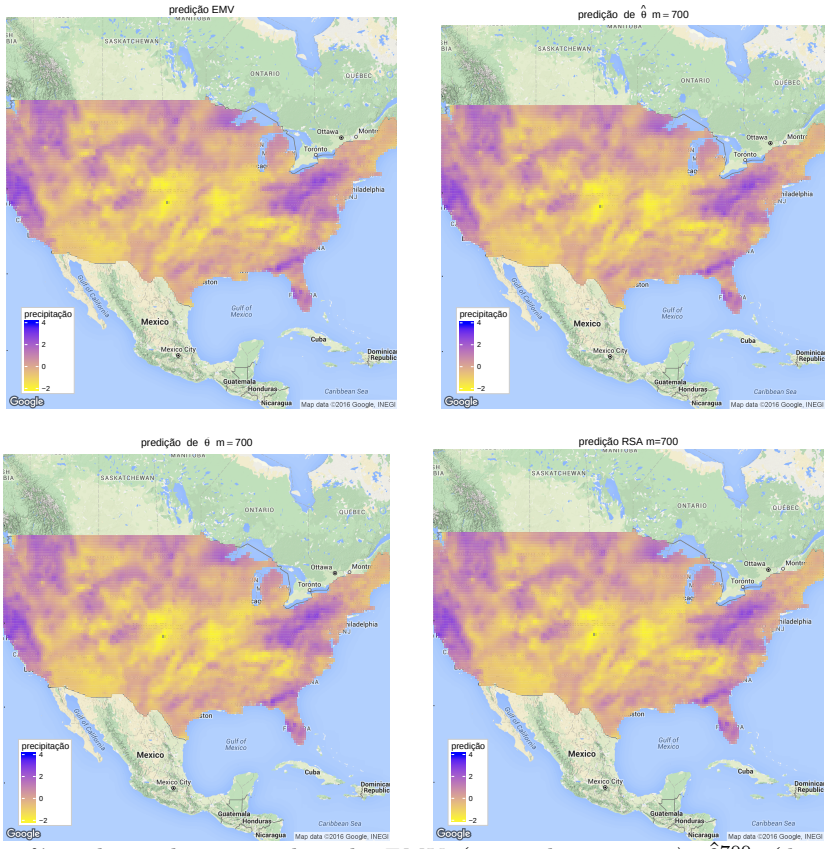

Figura 7.3: *Superfície de predição gerada pelo EMV (esquerda superior),* $\hat{\boldsymbol{\theta}}^{700}$ *(direita superior),* $\tilde{\boldsymbol{\theta}}^{700}$ *(esquerda inferior) e* RSA^{700} *(direita inferior). A função de covariância é a exponencial.*

Para obtenção dos resultados o EMV precisou de 21944 minutos, o que equivale a 15 dias e 5 horas de processamento. Os estimadores $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ foram muito mais rápidos quando 24 núcleos eram utilizados: para $m = 300$ um tempo de 1,205 minutos e para $m = 700$ um tempo de 14,27 minutos. Observe que, os estimadores *subsemble* podem proporcionar uma grande economia de tempo, mantendo a qualidade das estimativas, visto que seus resultados são muito parecidos com o EMV.

Tabela 7.3: *Comparação das estimativas do EMV,* $\tilde{\boldsymbol{\theta}}$ *e* $\hat{\boldsymbol{\theta}}$ *para os dados de anomalias de precipitação. A função de covariância é a Matérn.*

Método	m	β_0	ϕ	σ^2	τ^2	κ	4CPU(m)	24CPU(m)
EMV	-	0.121	2.267	0.772	0.050	0.532	21944.68	-
$\hat{\boldsymbol{\theta}}$	300	0.097	2.795	0.721	0.059	1.050	7.208	1.205
	500	0.149	2.350	0.761	0.048	0.629	4.932	21.831
	700	0.088	2.324	0.753	0.045	0.769	21.23	14.270
$\tilde{\boldsymbol{\theta}}$	300	0.090	2.823	0.705	0.054	0.843	-	-
	500	0.148	2.592	0.795	0.048	0.608	-	-
	700	0.080	2.339	0.781	0.047	0.684	-	-

Por meio da Tabela 7.4 é possível avaliar a qualidade de predição dos estimadores, considerando que a função de covariância é da família Matérn. Novamente, pode-se observar que a maior quantidade de vizinhos não aumenta a qualidade da predição. Comparando-se os resultados, nota-se que não há muita diferença entre os estimadores, em que o EMV apresentou os menores valores para o EQM, seguido pelo $\hat{\boldsymbol{\theta}}^{500}$ e $\tilde{\boldsymbol{\theta}}^{500}$.

As superfícies de predição calculadas por meio da função de covariância Matérn foram omitidas, visto que os resultados das Tabelas 7.2 e 7.4 foram muito parecidos e não seria possível visualizar diferenças entre as figuras das superfícies. Além disso, é importante ressaltar que pela Tabela 7.3 as

Tabela 7.4: *Comparação do EQM de predição do EMV, $\tilde{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ para as estações de monitoramento localizadas na Figura 7.2. A função de covariância utilizada é a Matérn.*

		nº vizinhos			
Método	m	25	50	100	150
MLE	-	0.0977	0.0974	0.0976	0.0976
$\hat{\boldsymbol{\theta}}$	300	0.1175	0.118	0.1204	0.1205
	500	0.0985	0.0981	0.0984	0.0984
	700	0.1005	0.1002	0.1007	0.1007
$\tilde{\boldsymbol{\theta}}$	300	0.1076	0.1074	0.1086	0.1086
	500	0.0986	0.0981	0.0984	0.0984
	700	0.099	0.0986	0.0989	0.0990

estimativas para o parâmetro de suavização κ do EMV, $\hat{\boldsymbol{\theta}}^{500}$, $\tilde{\boldsymbol{\theta}}^{700}$, $\tilde{\boldsymbol{\theta}}^{500}$, $\tilde{\boldsymbol{\theta}}^{700}$ são próximas de 0,5. Quando $\kappa = 0,5$ a função de covariância é a exponencial, um caso particular da família Matérn.

Capítulo 8

Conclusões

Neste trabalho foram propostos dois estimadores para análise de grandes bancos de dados no contexto de geoestatística. Os estimadores *subsemble* espacial são exemplos da estratégia *divide and conquer*, no qual o problema de estimação pode ser dividido em três passos: selecionar pequenos subconjuntos do banco de dados; analisar cada subconjunto separadamente e agregar os resultados para gerar as estimativas. A grande diferença dessa nova metodologia é que cada pedaço em que o banco de dados é dividido pode ser analisado separadamente, o que possibilita a utilização de programação paralela. Esse tipo de recurso reduz o tempo necessário para o cálculo das estimativas, além de exigir menos memória RAM. Outra vantagem da técnica proposta é a facilidade de implementação, pois os valores necessários para calcular os estimadores são de fácil interpretabilidade.

Foi realizado um estudo Monte Carlo para comparar os estimadores propostos com uma metodologia indicada para análise de massivas observações e para avaliar a qualidade dos estimadores em relação ao EMV. Os resultados indicam que as distribuições amostrais dos estimadores *subsemble* espacial e do EMV são similares. Como esperado a qualidade das estimativas depende do tamanho da subamostra, da força de correlação espacial e da presença ou ausência do efeito pepita. Maiores tamanho de subamostra produziram estimativas com menores vício e variabilidade. Ao contrário, quanto maior a força de correlação espacial, piores eram os resultados. Quando o efeito pepita é diferente de zero, há maior dificuldade na estimação dos parâmetros da função de covariância, sendo que essa característica é compartilhada pelos EMV, $\hat{\boldsymbol{\theta}}$ e $\tilde{\boldsymbol{\theta}}$.

No estudo simulado, comparou-se os estimadores RSA e *subsemble*. O estimador proposto apresentou melhores estimativas para os parâmetros da função de covariância. Além disso, o RSA necessita de muito mais tempo de processamento para gerar os resultados. Para a mesma configuração de *hardware*, os estimadores *subsemble* foram mais rápidos, quando o multiprocessamento era utilizado essa diferença aumentava consideravelmente. Por exemplo, na aplicação a dados de precipitação, o tempo para obtenção dos resultados considerando o estimador RSA foi 8 vezes maior.

Através da matriz de informação de Fisher assintótica, mensurou-se a variabilidade dos estimadores *subsemble* espacial, tornando possível o cálculo de intervalos de confiança. Também foram obtidas propriedades teóricas, em que condicionado a algumas suposições, o estimador é consistente.

Os estimadores desenvolvidos podem ser estendidos de várias formas. Uma opção seria a aplicação no contexto de análise espaço-temporal. Outra possibilidade é considerar diferentes estimadores na análise de cada subamostra, nesse trabalho, somente o EMV foi utilizado.

Referências Bibliográficas

Andrieu *et al.***(2005)** Christophe Andrieu, Éric Moulines e Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization*, 44(1): 283–312. Citado na pág. 2, 7

Banerjee *et al.***(2008)** Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finley e Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848. Citado na pág. 1

Borovskikh(1996) Yuri Vasilevich Borovskikh. *U-statistics in Banach Spaces*. VSP, Utrecht. Citado na pág. 17

Bühlmann *et al.***(2016)** Peter Bühlmann, Petros Drineas, Michael Kane e Mark van der Lann. *Handbook of Big Data*. Chapman and Hall/CRC, New York. Citado na pág. 2

Castrillón-Candás *et al.***(2015)** Julio E. Castrillón-Candás, Marc G. Genton e Rio Yokota. Multi-level restricted maximum likelihood covariance estimation and kriging for large non-gridded spatial datasets. *Spatial Statistics*, (in press). Citado na pág. 1

Cressie(2015) Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, New York. Citado na pág. 3

Datta *et al.***(2016)** Abhirup Datta, Sudipto Banerjee, Andrew O. Finley e Alan E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812. Citado na pág. 1

Efron() Bradley Efron. *Bootstrap methods: another look at the jackknife*. Springer, New York. Citado na pág. 14, 20

Finley *et al.***(2009)** Andrew O. Finley, Huiyan Sang, Sudipto Banerjee e Alan E. Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884. Citado na pág. 1

Fuentes(2007) Montserrat Fuentes. Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102(477):321–331. Citado na pág. 1, 4

Furrer *et al.***(2006)** Reinhard Furrer, Marc G. Genton e Douglas Nychka. Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523. Citado na pág. 1, 53

Guha *et al.***(2012)** Saptarshi Guha, Ryan Hafen, Jeremiah Rounds, Jin Xia, Jianfu Li, Bowei Xi e William S. Cleveland. Large complex data: divide and recombine (d&r) with rhipe. *Stat*, 1(1): 53–67. Citado na pág. 11

Guyon(1982) Xavier Guyon. Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105. Citado na pág. 4

Hall *et al.***(1995)** Peter Hall, Joel L. Horowitz e Bing-Yi Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574. Citado na pág. 20

Kaufman *et al.***(2008)** Cari G. Kaufman, Mark J. Schervish e Douglas W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555. Citado na pág. 1, 5, 53

Kleiner *et al.***(2014)** Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar e Michael I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816. Citado na pág. 2

Kunsch(1989) Hans R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 3(17):1217–1241. Citado na pág. 20

Lahiri(1996) Soumendra N. Lahiri. On inconsistency of estimators based on spatial data under infill asymptotics. *Sankhyā: The Indian Journal of Statistics, Series A*, 58(3):403–417. Citado na pág. 18

Lahiri e Zhu(2006) Soumendra N. Lahiri e Jun Zhu. Resampling methods for spatial regression models under a class of stochastic designs. *The Annals of Statistics*, 34(4):1774–1813. Citado na pág. xi, 21, 22

Lahiri *et al.***(1999)** Soumendra N. Lahiri, Mark S. Kaiser, Noel Cressie e Nan-Jung Hsu. Prediction of spatial cumulative distribution functions using subsampling. *Journal of the American Statistical Association*, 94(445):86–97. Citado na pág. 20

Lahiri(2013) Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, New York. Citado na pág. 20

Lee(1990) A. J. Lee. *U-Statistics: Theory and Practice*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, New York. Citado na pág. 20

Liang *et al.***(2013)** Faming Liang, Yichen Cheng, Qifan Song, Jincheol Park e Ping Yang. A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association*, 108(501):325–339. Citado na pág. 2, 5, 6, 10, 17, 18, 23, 53, 55

Lindgren *et al.***(2011)** Finn Lindgren, Håvard Rue e Johan Lindstrom. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73(4):423–498. Citado na pág. 1, 6

Liu e Singh(1992) Regina Y. Liu e Kesar Singh. Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, 225:248. Citado na pág. 20

Loh(2005) Wei-Liem Loh. Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields. *The Annals of Statistics*, 33(5):2344–2394. Citado na pág. 4

Mardia e Marshall(1984) Kanti V. Mardia e R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146. Citado na pág. 4, 17, 22

Nordman e Lahiri(2004) Daniel J. Nordman e Soumendra N. Lahiri. On optimal spatial subsample size for variance estimation. *Annals of statistics*, 32(5):1981–2027. Citado na pág. 12

Nordman *et al.***(2007)** Daniel J. Nordman, Soumendra N. Lahiri e Brooke L. Fridley. Optimal block size for variance estimation by a spatial block bootstrap method. *Sankhyā: The Indian Journal of Statistics*, 69(3):468–493. Citado na pág. 12, 20

Politis *et al.***(1999a)** Dimitris Politis, Joseph P. Romano e Michael Wolf. Weak convergence of dependent empirical measures with application to subsampling in function spaces. *Journal of statistical planning and inference*, 79(2):179–190. Citado na pág. 20

Politis e Romano(1994a) Dimitris N. Politis e Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313. Citado na pág. 20

Politis e Romano(1994b) Dimitris N. Politis e Joseph P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050. Citado na pág. 17, 20, 22

Politis e Sherman(2001) Dimitris N. Politis e Michael Sherman. Moment estimation for statistics from marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):261–275. Citado na pág. 12, 22

Politis *et al.***(1999b)** Dimitris N. Politis, Joseph P. Romano e Michael Wolf. *Subsampling*. Springer, New York. Citado na pág. 12, 17, 20, 22

Ribeiro e Diggle(2001) Paulo J. Ribeiro e Peter J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):14–18. URL http://CRAN.R-project.org/doc/Rnews/. ISSN 1609-3631. Citado na pág. 23

Robbins e Monro(1951) Herbert Robbins e Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, 22(3):400–407. Citado na pág. 2, 6, 7, 10

Rosenblatt(2012) Murray Rosenblatt. *Stationary sequences and random fields*. Springer Science & Business Media, New York. Citado na pág. 21

Rue e Tjelmeland(2002) Håvard Rue e Hååkon Tjelmeland. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian journal of Statistics*, 29(1):31–49. Citado na pág. 1

Sapp *et al.***(2014)** Stephanie Sapp, Mark J. van der Laan e John Canny. Subsemble: an ensemble method for combining subset-specific algorithm fits. *Journal of applied statistics*, 41(6):1247–1259. Citado na pág. 2

Schifano *et al.***(2016)** Elizabeth D. Schifano, Jing Wu, Chun Wang, Jun Yan e Ming-Hui Chen. Online Updating of Statistical Inference in the Big Data Setting. *Technometrics*, 58(3):393–403. Citado na pág. 2

Sherman(1996) Michael Sherman. Variance estimation for statistics computed from spatial lattice data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):509–523. Citado na pág. 12, 20

Sherman e Carlstein(1994) Michael Sherman e Edward Carlstein. Nonparametric estimation of the moments of a general statistic computed from spatial data. *Journal of the American Statistical Association*, 89(426):496–500. Citado na pág. 20

Stein(2008) Michael L. Stein. A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, 37(1):3–10. Citado na pág. 1

Stein *et al.***(2004)** Michael L. Stein, Zhiyi Chi e Leah J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66 (2):275–296. Citado na pág. 1, 6

Stein *et al.***(1989)** Michael L. Stein *et al.* Design and Analysis of Computer Experiments: Comment. *Statistical Science*, 4(4):432–433. Citado na pág. 4

Sun e Stein(2016) Ying Sun e Michael L. Stein. Statistically and computationally efficient estimating equations for large spatial datasets. *Journal of Computational and Graphical Statistics*, 25(1):187–208. Citado na pág. 1

Sun *et al.***(2012)** Ying Sun, Bo Li e Marc G Genton. Geostatistics for large datasets. Em *Advances and challenges in space-time modelling of natural events*, páginas 55–77. Springer. Citado na pág. 2

Tanenbaum(2009) Andrew Tanenbaum. *Modern operating systems*. Pearson Education, Inc.„ Amsterdam. Citado na pág. 11

Van Der Vaart(1996) Aad Van Der Vaart. Maximum likelihood estimation under a spatial sampling scheme. *The Annals of Statistics*, 24(5):2049–2057. Citado na pág. 4

Vecchia(1988) Aldo V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):297–312. Citado na pág. 1, 5, 6

Whittle(1954) P. Whittle. On stationary processes in the plane. *Biometrika*, 41(3):434–449. Citado na pág. 4

Ying(1993) Zhiliang Ying. Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, 21(3):1567–1590. Citado na pág. 4

Zhang(2004) Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261. Citado na pág. 4, 17

Zhao e Chen(1990) Lincheng Zhao e Xiru Chen. Normal approximation for finite-population U-statistics. *Acta mathematicae applicatae Sinica*, 6(3):263–272. Citado na pág. 18