Aplicações de inferência bayesiana aproximada para modelos gaussianos latentes espaço temporais

Wagner Hugo Bonat
Orientador: Paulo Justiniano Ribeiro Jr
Universidade Federal do Paraná
Programa de Pós-Graduação em Métodos Numéricos em Engenharia

15 de Fevereiro de 2010

Motivação

- Modelos estocástico têm sido amplamente utilizados tanto na comunidade científica como no mundo dos negócios em geral.
- Estudos de mercado, predição em séries financeiras, análises de componentes de solo, mapeamento de doenças, entre outros.

Motivação

- Nesta diversidade de aplicações é fácil encontrar situações de relevância prática onde os modelos tradicionais (GLM) não são adequados.
- Em geral por pelo menos uma das seguintes características:

1. efeito não linear de covariáveis,
2. observações correlacionadas no espaço,
3. observações correlacionadas no tempo,
4. heterogeneidade entre unidades não explicada por covariáveis.

Modelos estruturados aditivamente

- Nestas situações a classe dos modelos de regressão estruturados aditivamente têm sido amplamente utilizada.

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i \tag{1}$$

Modelos gaussianos latentes são um subconjunto de todos os modelos Bayesianos estruturados aditivamente, onde se supõe uma priori gaussiana para $\alpha, f^{(j)}(.)$, β_k e ϵ_i.

Inferência Bayesiana

- Inferência baseada em simulação, MCMC *Markov Chain Monte Carlo*.
- Desempenho insatisfatório quando aplicado para modelos Gaussianos latentes, basicamente pela alta dependência entre os parâmetros.
- Apesar dos avanços com MCMC ele permanece lento e complicado do ponto de vista do usuário final.

Inferência Bayesiana - Outras abordagens

- *Variational Bayes* (BISHOP, 2006).
- *Expectation-Propagation* (MINKA, 2001).
- A ferramenta mais promissora parece ser *Integrated Nested Laplace approximation*-INLA (RUE et al, 2009).

Objetivos Gerais

- Revisar os fundamentos dos modelos Gaussianos Latentes, com ênfase em modelos de interação espaço-temporal.
- Revisar o artigo de RUE et al (2009) mostrando que é possível estimar modelos com interaçaõ espaço-temporal com essa abordagem.
- Aplicar a metodologia a três conjuntos de dados.

Modelos Gaussianos Latentes

- Dados observados $\underline{y}$, $y_i|x_i \sim \pi(y_i|x_i, \theta)$
- Campo latente Gaussiano $\underline{x} \sim N(.; Q^{-1}(\underline{\theta}))$
- Hiperparâmetro $\underline{\theta}$
  1. variabilidade
  2. tamanho/força da dependência
  3. parâmetros na verossimilhança

Inferência bayesiana em MGL

- A posteriori pode ser escrita como

$$\pi(\underline{x}, \underline{\theta}|\underline{y}) \propto \pi(\underline{\theta})\pi(\underline{x}|\underline{\theta}) \prod_i \pi(y_i|x_i, \underline{\theta})$$

$$\propto \pi(\underline{\theta})|Q(\underline{\theta})|^{\frac{n}{2}} \exp\left(-\frac{1}{2}\underline{x}^T Q(\underline{\theta})\underline{x} + \sum_i \log \pi(y_i|x_i, \underline{\theta})\right)$$

- As marginais posteriori de interesse podem ser escritas como

$$\pi(x_i|\underline{y}) = \int \pi(x_i|\underline{\theta}, \underline{y})\pi(\underline{\theta}|\underline{y})d\underline{\theta}$$

$$\pi(\theta_j|\underline{y}) = \int \pi(\underline{\theta}|\underline{y})d\underline{\theta}_{-j}$$

Inferência bayesiana aproximada para MLG

- O fato chave da abordagem INLA é construir aproximações aninhadas para cada um dos componentes

$$\widetilde{\pi}(x_i|\underline{y}) = \int \widetilde{\pi}(x_i|\underline{\theta}, \underline{y})\widetilde{\pi}(\underline{\theta}|\underline{y})d\underline{\theta}$$

e

$$\widetilde{\pi}(\theta_j|\underline{y}) = \int \widetilde{\pi}(\underline{\theta}|\underline{y})d\underline{\theta}_{-j}$$

INLA - Integrated Nested Laplace approximation

- O primeiro passo é usar a seguinte identidade

$$\pi(\underline{\theta}|\underline{y}) = \frac{\pi(\underline{y}|\underline{x},\underline{\theta})\pi(\underline{x}|\underline{\theta})\pi(\underline{\theta})}{\pi(\underline{y})\pi(\underline{x}|\underline{\theta},\underline{y})} \propto \frac{\pi(\underline{y}|\underline{x},\underline{\theta})\pi(\underline{x}|\underline{\theta})\pi(\underline{\theta})}{\pi(\underline{x}|\underline{\theta},\underline{y})} \tag{2}$$

- Importante é notar que

$$\pi(\underline{x}|\underline{\theta},\underline{y}) \propto \exp\left(-\frac{1}{2}\underline{x}^T Q\underline{x} + \sum_i \log \pi(y_i|x_i,\underline{\theta})\right)$$

INLA - Integrated Nested Laplace approximation

- O núcleo do INLA é aproximar $\pi(\underline{x}|\underline{\theta}, \underline{y})$ por $\pi_G(\underline{x}|\underline{\theta}, \underline{y})$
- A aproximação usa a moda e a curvatura na moda de $\pi(\underline{x}|\underline{\theta}, \underline{y})$.
- Válida em um ponto (moda), aplicando em 2 tem-se

$$\widetilde{\pi}(\underline{\theta}|\underline{y}) \propto \left. \frac{\pi(\underline{y}|\underline{x}, \underline{\theta})\pi(\underline{x}|\underline{\theta})\pi(\underline{\theta})}{\widetilde{\pi}_G(\underline{x}|\underline{\theta}, \underline{y})} \right|_{\underline{x}=\underline{x}^*(\underline{\theta})}$$

INLA - Integrated Nested Laplace approximation

- Esta aproximação resolve três problemas no processo de inferência:
  1. Integrar fora a incerteza com respeito a $\underline{\theta}$, quando aproximando $\widetilde{\pi}(x_i|\underline{y})$.
  2. Calcular uma aproximação para a verossimilhança marginal.
  3. Marginais posteriori para os hiperparâmetros $\widetilde{\pi}(\theta_m|\underline{y})$.
- Integração numérica sobre um domínio multidimensional.

INLA - Estratégia geral

1. Selecione um conjunto de $\Theta = (\underline{\theta}_1, \ldots, \underline{\theta}_k)$
2. Para $k = 1$ até K faça
3. Calcule $\widetilde{\pi}(\underline{\theta}_k | \underline{y})$
4. Calcule $\widetilde{\pi}(x_i | \underline{\theta}_k, \underline{y})$ como uma função de x_i
5. Fim para
6. Calcule $\widetilde{\pi}(x_i | \underline{y}) = \sum_k \widetilde{\pi}(x_i | \underline{\theta}_k, \underline{y}) \widetilde{\pi}(\underline{\theta}_k | \underline{y}) \Delta_k$

INLA - Estratégia geral

- Para este algoritmo funcionar precisamos saber como obter duas quantidades

1. Como selecionar um conjunto (possívelmente pequeno) de pontos $\Theta = (\underline{\theta}_1, \ldots, \underline{\theta}_k)$.
2. Como construir uma boa aproximação para $\pi(x_i|\underline{\theta}_k, \underline{y})$

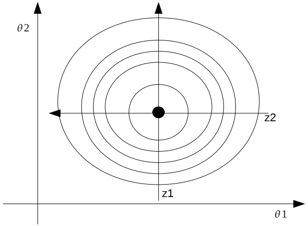
Explorando $\widetilde{\pi}(\underline{\theta}|\underline{y})$

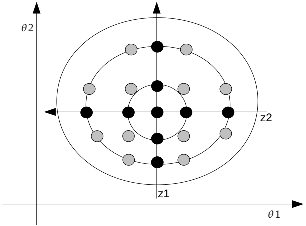
1. Encontre seu ponto de máximo de $\underline{\theta}^*$.
2. Na moda calcule a Hessiana $H > 0$. Seja $\Sigma = H^{-1}$. Para facilitar a exploração use variáveis padronizadas z ao invés de $\underline{\theta}$. Seja $\Sigma = V\delta V^T$ a decomposição em autovalores e autovetores de Σ e defina $\underline{\theta}$ através de z com

$$\underline{\theta}(z) = \underline{\theta}^* + V\delta^{1/2}z$$

3. Explore a $\log \widetilde{\pi}(\underline{\theta}|\underline{y})$ usando a z-reparametrização

Explorando $\widetilde{\pi}(\underline{\theta}|\underline{y})$





Aproximando $\pi(x_i|\underline{\theta}, \underline{y})$

- Rue et. al (2009) fazem três propostas
  1. Aproximação gaussiana $\widetilde{\pi}_G(x_i|\underline{\theta}, \underline{y})$ já explicada
  2. Aproximação de Laplace
  3. Aproximação de Laplace Simplificada

Aproximando $\pi(x_i|\underline{\theta}, \underline{y})$

$$\widetilde{\pi}_{LA}(x_i|\underline{\theta}, \underline{y}) \propto \left. \frac{\pi(\underline{y}|\underline{\theta}, \underline{x})\pi(\underline{x}|\underline{\theta})\pi(\underline{\theta})}{\widetilde{\pi}_{GG}(\underline{x}_{-i}|x_i, \underline{\theta}, \underline{y})} \right|_{\underline{x}_{-i}=\underline{x}^*_{-i}(x_i,\underline{\theta})} \tag{3}$$

- Muito cara computacionalmente
- Duas mudanças são propostas em Rue et. al (2009)
  1. $$\underline{x}^*_{-i} \approx E_{\widetilde{\pi}_G}(\underline{x}_{-i}|x_i) \tag{4}$$
  2. Somente alguns x_j próximos a x_i devem impactar na marginal de x_i.

Aproximando $\pi(x_i|\underline{\theta}, \underline{y})$

- A esperança condicional em 4 implica que

$$\frac{E_{\widetilde{\pi}_G}(x_j|x_i) - \mu_j(\underline{\theta}}{\sigma_j(\underline{\theta})} = a_{ij}(\underline{\theta})\frac{x_i - \mu_i(\underline{\theta}}{\sigma_i(\underline{\theta})} \tag{5}$$

para algum $a_{ij}(\underline{\theta})$ quando $j \neq i$. Uma regra simples é

$$R_i(\underline{\theta}) = (j : |a_{ij}(\underline{\theta})| > 0.001)$$

Aproximando $\pi(x_i|\underline{\theta}, \underline{y})$

- A expressão 3 ainda precisa ser calculada para diferentes valores de x_i.
- Para selecionar estes pontos usamos $\widetilde{\pi}_G(x_i|\underline{\theta}, \underline{y})$.

$$x_i^{(s)} = \frac{x_i - \mu_i(\underline{\theta})}{\sigma_i(\underline{\theta})}$$

Para a escolha das abscissas recorre-se a quadratura de Gauss-Hermite. Para representar a densidade $\widetilde{\pi}_{LA}(x_i|\underline{\theta}, \underline{y})$, usa-se

$$\widetilde{\pi}_{LA}(x_i|\underline{\theta}, \underline{y}) \propto N(x_i; \mu_i(\underline{\theta}, \sigma_i^2(\underline{\theta})) \times \exp(\text{cubic spline}(x_i))$$

Aspectos gerais

- Identificar possíveis impactos e alterações na qualidade da água decorrente da existência dos reservatórios.
- Covariáveis
    1. Local de coleta (montante, reservatório e jusante).
    2. Diferentes usinas hidroelétricas.
    3. Coletas realizadas trimestralmente - 06/03/2003 até 30/10/2008
- 23 datas de coletas, 8 usinas totalizando 552 observações.
- Resposta Normal após transformação logística.

Modelos ajustados

Tabela: Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados, verossimilhança marginal e critério de informação de *Akaike*.

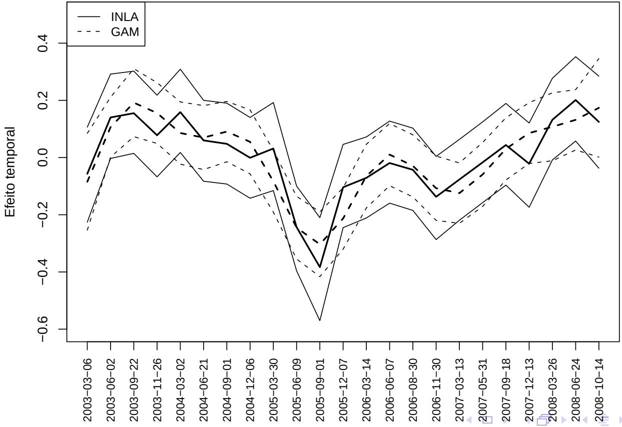
Modelos	Preditor Linear	DIC	NP	MV	AIC
1	$Y \sim 1$	$837,18$	$1,658$	$-427,32$	$837,18$
2	$Y \sim \beta_{uhe} + \beta_{loc}$	$797,86$	$10,66$	$-415,08$	$798,50$
3	$Y \sim \beta_{uhe} + \beta_{loc} + \rho_{t}$	$755,72$	$24,01$	$-402,16$	$768,68$
4	$Y \sim \beta_{uhe} + \beta_{loc} + \rho_{loc:t}$	$773,72$	$41,14$	$-563,16$	$784,85$
5	$Y \sim \beta_{uhe} + \beta_{loc} + \rho_{uhe:loc:t}$	$741,15$	$58,11$	$-440,12$	$---$

INLA x GAM

Tabela: Resultados do modelo 3 via INLA e GAM.

Parâmetros	Média Posteriori	Desvio Padrão	Estimativa	Desvio Padrão
Intercepto	$1,2725$	$0,06053$	$1,3398$	$0,06454$
Reservatorio	$0,1933$	$0,04930$	$0,1933$	$0,05016$
Jusante	$0,1557$	$0,04930$	$0,1557$	$0,05016$

INLA x GAM

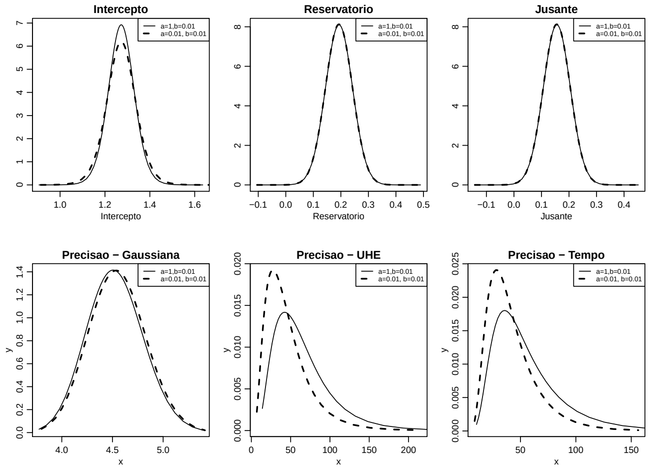

INLA
GAM
Efeito temporal
0.4
0.2
0.0
−0.2
−0.4
−0.6
2003−03−06
2003−06−02
2003−09−22
2003−11−26
2004−03−02
2004−06−21
2004−09−01
2004−12−06
2005−03−30
2005−06−09
2005−09−01
2005−12−07
2006−03−14
2006−06−07
2006−08−30
2006−11−30
2007−03−13
2007−05−31
2007−09−18
2007−12−13
2008−03−26
2008−06−24
2008−10−14

INLA x GAM

Tabela: Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados.

Abordagem	Erro quadrático	Erro absoluto	Correlação	Cobertura
INLA	$0,1921$	$0,3390$	$0,5663$	$0,6700$
GAM	$0,2679$	$0,3918$	$0,5628$	$0,5416$

Sensibilidade a priori



Aspectos gerais

- Investigar fatores associados a ocorrência de ovos de *Aedes aegypti*.
- Covariáveis
  1. Locais - Tipo de imóvel, quintal, água ligada na rede, canalizada no cômodo, fatores de risco, recipientes.
  2. Climáticas - Umidade, precipitação, temperatura máxima e mínima.
- 80 armadilhas, 124 datas de coletas, total de 2480 observações.
- Variável resposta Binomial Negativa.

Modelos ajustados

Tabela: Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal.

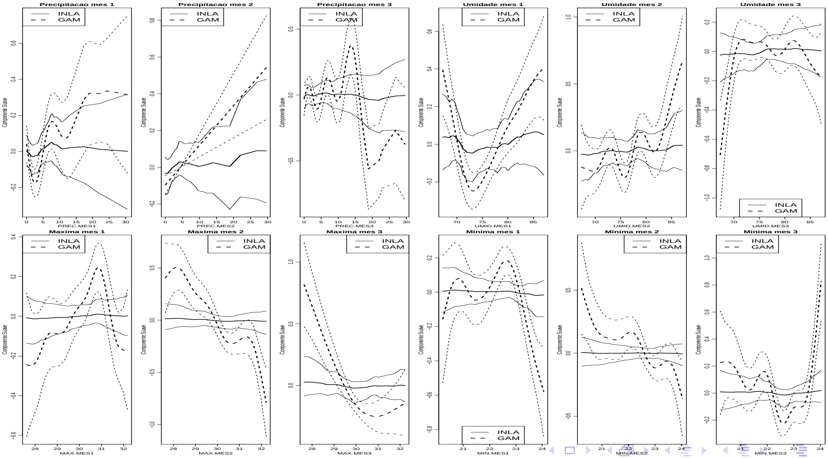
Modelos	Preditor Linear	DIC	NP	MV
1	$Y \sim 1$	$35691,77$	$1,973$	$-17853,93$
2	$Y \sim \gamma_t + \phi_i$	$35004,85$	$172,34$	$-17624,54$
3	$Y \sim \rho_t + \varphi_i$	$34969,34$	$123,04$	$-17635,08$
4	$Y \sim \rho_t + \varphi_i + \gamma_t + \phi_i$	$34968,48$	$126,34$	$-17633,14$
5	$Y \sim$ Tipo I	$34972,28$	$150,79$	$-17637,76$
5	$Y \sim$ Tipo II	$34979,30$	$156,41$	$-17990,40$
5	$Y \sim$ Tipo III	$35121,00$	$264,38$	$-23964,46$
5	$Y \sim$ Tipo IV	$35105,00$	$245,19$	$-24176,06$

INLA x GAM - Covariáveis locais

Tabela: Ajustes dos modelos para cada covariável na presença dos efeitos espaciais e temporais, abordagens INLA e GAM.

Parâmetros	Média Post.	Int. Cred.	Estimativa	Int. Conf.
Tipo de imóvel	$-0,0406$	$(-0,2314; 0,1487)$	$-0,2735$	$(-0,6282; 0,0812)$
Quintal	$0,0743$	$(-0,1332; 0,2826)$	$0,0652$	$(-0,0537; 0,1841)$
Água ligada a rede geral	$-0,0667$	$(-0,4071; 0,2744)$	$-0,1991$	$(-0,4052; 0.0070)$
Abastecimento de água	$0,0804$	$(-0,3199; 0,4799)$	$0,0926$	$(-0,2384; 0,4236)$
Água canal. no cômodo	$-0,1167$	$(-0,3448; 0,1098)$	$-0,2429$	$(-0,3762; -0,1135)$
Fatores de risco	$0,1436$	$(-0,0201; 0,3074)$	$0,0832$	$(0,0039; 0,1928)$
Rec. grandes sem tampa	$-0,0543$	$(-0,2328; 0,1237)$	$-0,1259$	$(-0,2274; -0,0243)$
Rec. grandes com tampa	$-0,0262$	$(-0,2764; 0,2219)$	$-0,0159$	$(-0,1558; 0,1240)$
Rec. pequenos sem tampa	$-0,0403$	$(-0,3710; 0,2897)$	$0,0299$	$(-0,1525; 0,2123)$
Rec. pequenos com tampa	$-0,0863$	$(-0,2989; 0,1250)$	$-0,1364$	$(-0,2528; -0,0199)$
Fre. coleta de lixo	$-0,0241$	$(-0,2265; 0,1763)$	$0,1238$	$(-0,0392; 0,2868)$
Grupos 1 - 2	$0,1509$	$(-0,0814; 0,3829)$	$0,0792$	$(-0,0436; 0,2020)$
Grupos 1 - 3	$0,1551$	$(-0,0849; 0,3949)$	$0,1201$	$(-0,0041; 0,2443)$
Grupos 1 - 4	$-0,0012$	$(-0.2325; 0,2292)$	$0,0040$	$(-0,1212; 0,1292)$

INLA x GAM - Covariáveis ambientais

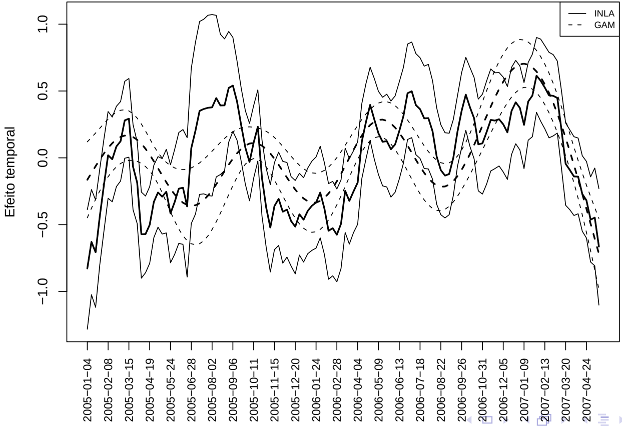


INLA x GAM - Modelo em Bonat et. al 2009

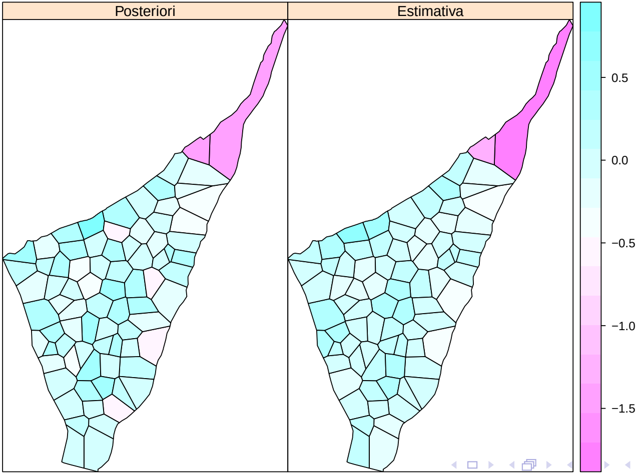
Tabela: Ajuste do modelo proposto em Bonat et. al 2009 pelas abordagens INLA e GAM.

Parâmetros	Média Posteriori	Int. Cred.	Estimativa	Int. Conf.
Intercepto	$6,4879$	$(3,8788; 9,1509)$	$5,1254$	$(3,3449; 6,9058)$
PREC.MES1	$0,0060$	$(-0,0159; 0,0275)$	$0,0277$	$(0,0161; 0,0392)$
PREC.MES2	$0,0097$	$(-0,0131; 0,0315)$	$0,0278$	$(0,0162; 0,0393)$
UMID.MES3	$0,0096$	$(-0,0251; 0,0437)$	$0,0257$	$(0,0029; 0,04843)$
Canalizada	$-0,1104$	$(-0,3430; 0,1208)$	$-0,2415$	$(-0,3751; -0,1078)$
Grande sem tampa	$-0,0452$	$(-0,2252; 0,1348)$	$-0,1119$	$(-0,2161; -0,0076)$

INLA x GAM - Efeito temporal



INLA x GAM - Efeito espacial



Medidas de erro

Tabela: Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados.

Abordagem	Erro quadrático	Erro absoluto	Correlação	Cobertura
INLA	$958264,7$	$685,23$	$0,3440$	$0,6029$
GAM	1009266	$681,21$	$0,1962$	$0,3432$

Aspectos gerais

- Entender a dinâmica da infestação das plantas.
- O modelo deve levar em consideração a estrutura espaço temporal do experimento.
- Talhão com 20 linhas e 58 plantas em cada linha.
- Total de 1160 plantas.
- Avaliadas em 30 tempos, totalizando 34800 observações.
- Variável resposta dicotômica.

Modelos ajustados

Tabela: Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados, verossimilhança marginal e critério de informação de *Akaike*.

Modelos	Preditor Linear	DIC	NP	MV
1	$Y \sim 1$	$24592,37$	$1,022$	0
2	$Y \sim \gamma_t$	$20043,72$	$29,87$	$-10090,03$
3	$Y \sim \rho_t$	$20028,79$	$20,20$	$-10038,06$
4	$Y \sim \phi_i$	$16495,41$	$1601,32$	$-8084,46$
5	$Y \sim \varphi_i$	$15164,56$	$918,60$	$-8825,41$
6	$Y \sim \gamma_t + \phi_i$	$9885,40$	$3538,92$	$-3246,39$
7	$Y \sim \rho_t + \varphi_i$	$9716,71$	$3457,86$	$-3843,93$
8	$Y \sim \rho_t + \varphi_i + \gamma_t + \phi_i$	$9601,78$	$3399,04$	$-3844,82$

INLA x GAM - Efeito temporal

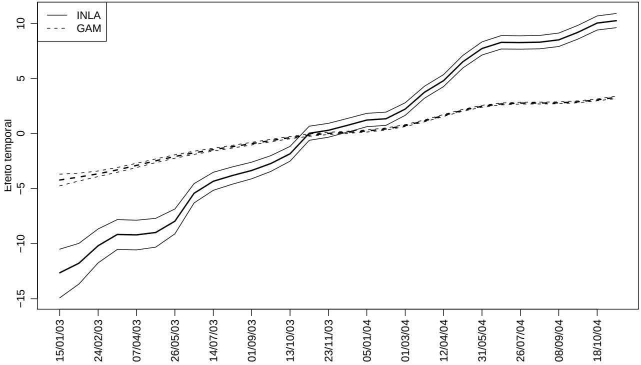


Figura: Sobreposição do efeito temporal estimado via INLA e GAM na estrutura do modelo 7.

INLA x GAM - Efeito espacial

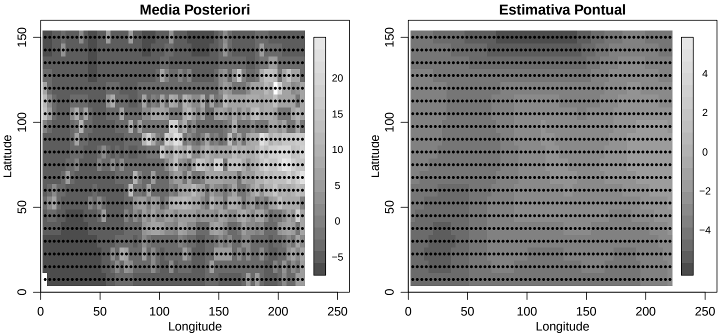


Figura: Efeito espacial estimado via INLA e GAM na estrutura do modelo 7.

INLA x GAM - Percentual observado x estimado

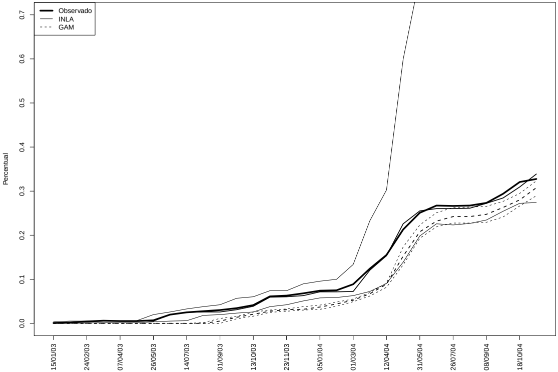


Figura: Comparação entre o percentual observado e estimado pelas abordagens INLA e GAM por data de coleta.

INLA x GAM - Percentual de acertos

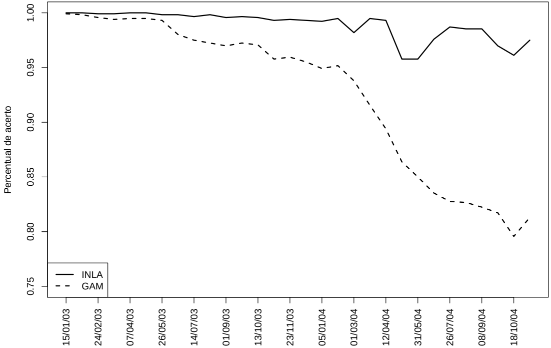


Figura: Comparação entre o percentual de acertos estimado pelas abordagens INLA e GAM por data de coleta.

Sensibilidade a priori

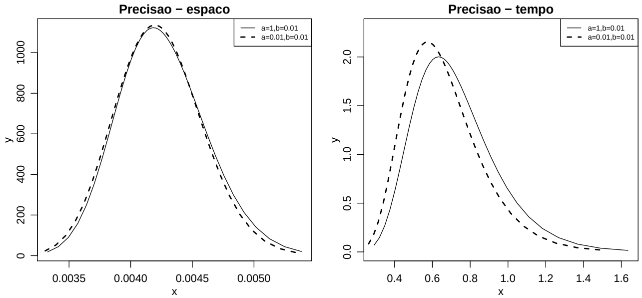


Figura: Distribuições a posteriori de acordo com a especificação de diferentes prioris para os parâmetros de precisão dos efeitos espaciais e temporais.

Gerais

- MLG é uma classe flexível de modelos
- A abordagem INLA mostrou-se eficiente
- Concordância entre INLA e GAM diferente em cada conjunto de dados
- Modelos poucos sensíveis a troca de prioris

Das comparações

- INLA mais conservador
- INLA intervalos mais realísticos
- Resultados conflitantes entre *splines* e *random walk*
- Efeitos espaciais e temporais tendem a ser super suavizados com GAM
- Diferença aumenta com a complexidade do conjunto de dados

Rue, H. ; Martino, S. ; Chopin, N.
Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.
J. R. Statistical Society B 71: 319-392 (2009)

Knorr-Held, L.
Bayesian modelling of inseparable space-time variation in disease risk.
Statistical Medicine 19: 2555-2567 (2000)

Bonat et. al,
Investigando fatores associados a ocorrência de ovos de Aedes aegypti coletados em ovitrampas em Recife/PE.
Revista Brasileira de Biometria 27: 519-537 (2009)