PROPRIEDADES DOS ESTIMADORES DOS PARÂMETROS DO MODELO AUTOLOGÍSTICO ESPACIAL

Denise N. Viola (LCE - ESALQ/USP - DES/UFBa), Clarice G. B. Demétrio (LCE - ESALQ/USP), Bryan F. Manly (Western EcoSystems Technology), Paulo J. Ribeiro Júnior (DES-UFPR)

Resumo

Neste artigo é feito um estudo de simulação para verificar o comportamento dos estimadores de pseudo-verossimilhança dos parâmetros do modelo autologístico, considerando diferentes estruturas de covariáveis e de vizinhança, três intensidades de infestação de uma praga e cinco valores para o parâmetro de correlação entre os vizinhos. Uma aplicação dos modelos considerados no estudo de simulação é feita a um conjunto de dados provenientes de um experimento com pimenta do sino, utilizado por Gumpertz, Graham e Ristino (1997). Mostra-se que o método de estimação por pseudo-verossimilhança pode ser usado, com certa cautela, quando o interesse está na contribuição das covariáveis, mas não deve ser usado quando o interesse está na estimação da correlação espacial.

Palavras chaves: Modelo autologístico, dependência espacial, dados binários, distribuição Bernoulli

Abstract

Properties of the spatial autologistic model parameters - A simulation study
In this paper a simulation study is performed in order to verify the effect of different covariate and neighbouring structures, three pest infestation levels and five different spatial correlation coefficient values on the pseudo-likelihood estimators of the autologistic model parameters. An application of the metodology is presented using the bell pepper data from Gumpertz, Graham e Ristino (1997). It is shown that the pseudo-likelihood method can be used when the researcher is interested on the covariate contributions but should not be used when he is interested on the estimation of the spatial correlation.
Keywords: autologistic model, spatial correlation, binary data, Bernoulli distribution

0.1 Introdução

Variáveis respostas binárias, isto é, do tipo sucesso/fracasso são muito comuns na Experimentação Agronômica. Assim, por exemplo, em Fitopatologia, o pesquisador pode estar interessado no estudo da presença ou ausência de uma determinada doença e sua distribuição espacial. Nesse tipo de estudo, espera-se, em geral, que as observações sejam correlacionadas no espaço e/ou no tempo.

O modelo padrão para a análise de respostas binárias tem sido o modelo logístico de regressão que tem como uma de suas pressuposições a independência das observações. Assim sendo, no caso de existir dependência espacial ou temporal esse modelo não é indicado e modelos que levam em consideração a correlação existente entre as observações têm sido propostos na literatura.

Com a finalidade de incorporar a informação da vizinhança como covariável foram propostos os modelos autologísticos (BESAG, 1972, AUGUSTIN; MUGGLESTONE; BUCKLAND, 1996, GUMPERTZ; GRAHAM; RISTANO, 1997). As áreas de aplicação são diversas e incluem estudos sobre fauna aquática de macro invertebrados em 76 lagoas inglesas (SANDERSON; EYRE; RUSHTON, 2005), comportamento de clientes em relação a políticas de seguro (MOON; RUSSEL 2005), rupturas causadas por besouros em pinus nos Estados Unidos (WU et al., 2005), mapeamento de pobreza em países em desenvolvimento (PETRUCCI; SALVATI; SEGHIERI 2004), distribuição espacial de renas na Suécia (TETERUKOVSKIY; EDEMIRS, 2003), distribuição de vegetação em florestas, considerando covariáveis climáticas (HE; ZHOU; ZHU 2003), distribuição da epidemia do *Phytophthora* em pimenta do sino considerando efeitos de variáveis do solo (GUMPERTZ; GRAHAM; RISTAINO, 1997), distribuição de espécies de plantas considerando covariáveis climáticas (WU; HUFFER, 1997), distribuição espacial de alces em uma região da Escócia (AUGUSTIN; MUGGLESTONE; BUCKLAND, 1996), análise genética de características familiares (ABEL; GLOLMARD; MALLET, 1993), dentre outros. Uma comparação entre o modelo autologístico e o modelo logístico regressivo é apresentada em Abel, Golmard, Mallet, (1993).

Entretanto, esses tipos de modelos ainda necessitam de estudos mais detalhados em relação às propriedades dos métodos de estimação e dos estimadores propostos. O método de pseudo-verossimilhança é relativamente simples quando comparado com métodos alternativos, porém suas propriedades não têm sido extensivamente estudadas (PETRUCCI; SALVATI; SEGHIERI, 2004). Com o grande desenvolvimento dos recursos computacionais o uso de simulações tem sido muito comum para o estudo de propriedades estatísticas de interesse. Esses estudos são baseados em informações reais e utilizados como repetições de um experimento, sendo indicado para respostas contínuas e discretas.

Neste artigo é feito um estudo de simulação para verificar o comportamento dos estimadores de pseudo-verossimilhança dos parâmetros do modelo autologístico, considerando diferentes estruturas de covariáveis e de vizinhança, três intensidades de infestação de uma praga e cinco valores para o parâmetro de correlação entre os vizinhos. Uma aplicação dos modelos considerados no estudo de simulação é feita a um conjunto de dados provenientes de um experimento com pimenta do sino, utilizado por Gumpertz, Graham e Ristino (1997).

O restante do artigo está organizado como se segue. A Seção 0.2 descreve o modelo logístico enquanto que a Seção 0.3, o modelo autologístico com a estimação dos parâmetros descrita na

0.4. Na Seção 0.5 é feita a descrição da simulação cujos resultados são apresentados e discutidos na Seção 0.6. Uma aplicação do modelo autologístico é mostrada na Seção 0.6.1. Finalmente, algumas considerações finais aparecem na Seção 0.7.

0.2 Modelo logístico

Modelos lineares generalizados (MLG) envolvem três componentes, a saber um componente sistemático, um aleatório e uma função de ligação. O componente sistemático é definido durante o planejamento do experimento e as variáveis explicativas entram na forma de soma linear dos efeitos, isto é, com preditor linear $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, em que $\mathbf{X}$ é a matriz do modelo, $\boldsymbol{\beta}$ é o vetor de parâmetros. O componente aleatório é estabelecido após definidas as medidas que serão realizadas, em que o conjunto de variáveis aleatórias $Y_i, i = 1, 2, \ldots, n$ são mutuamente independentes com distribuição pertencente à família exponencial na forma canônica e $E(Y_i) = \mu_i$. A função de ligação relaciona o componente aleatório ao componenete sistemático, ou seja, a média da distribuição ao preditor linear. Logo, na seleção de modelos a serem ajustados a um conjunto de dados, é importante escolher a distribuição da variável resposta, a matriz do modelo e a função de ligação (DEMÉTRIO, 2001). Um caso particular dos MLG é o modelo de regressão logística que pode ser usado para a análise de variáveis binárias independentes.

Sejam Y_i, $i = 1, 2, \ldots, n$, variáveis aleatórias com distribuição de Bernoulli com probabilidade de sucesso π_i, sendo que y_i assume os valores zero (fracasso) ou um (sucesso). Tem-se que $E(Y_i) = \pi_i$ e $Var(Y_i) = \pi_i(1 - \pi_i)$. Então, um modelo linear generalizado permite que as probabilidades de sucesso π_i sejam modeladas em termos de p variáveis explanatórias x_{ik}, $k = 1, \ldots, p$, através de

$$g(\pi_i) = \boldsymbol{\beta}^T \mathbf{x}_i$$

em que g é uma função de ligação adequada e $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos. Considerando-se a função de ligação logística, tem-se

$$\text{logit}(\pi_i) \;=\; \log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

e, portanto,

$$\theta_i = P(Y = 1|x) = \frac{\exp\{\beta_0 + \sum_{k=1}^{p} \beta_j x_{ik}\}}{1 + \exp\{\beta_0 + \sum_{k=1}^{p} \beta_j x_{ik}\}}.$$

0.3 Modelo autologístico

O modelo autologístico teve impulso na área de Estatística Espacial com os artigos de Besag (1972, 1974) e é uma generalização do modelo logístico, considerando dependência espacial entre as respostas. A autocorrelação pode ser induzida por funções das respostas dos vizinhos como covariáveis no modelo e podem ser consideradas diversas estruturas de vizinhança chamadas de primeira, segunda e terceira ordens, respectivamente, com quatro, oito e doze vizinhos, conforme Figura 1. De uma forma geral, tem como preditor linear

Figura 1 – Representação esquemática de estrutura de vizinhança sobre um látice regular (primeira, segunda e terceira ordens).

$$\text{logit}(\pi_i) \quad = \quad \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \gamma^T \mathbf{z}_i$$

em que π_i é a probabilidade de sucesso de um evento para o i-ésimo indivíduo, $i = 1, \ldots, n$, β_k é o k-ésimo parâmetro associado à covariável x_{ik}, $\boldsymbol{\gamma}$ é o vetor parâmetros associado ao vetor $\mathbf{z}_i$ das informações da vizinhança. Portanto, a probabilidade de um sucesso é dada por

$$P(Y_i = 1|\text{vizinhos}) = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \gamma^T \mathbf{z}_i\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \boldsymbol{\gamma}^T \mathbf{z}_i\}}.$$

A covariável $\mathbf{z}_i = z_i$ foi definida, por Augustin, Mugglestone e Buckland (1996), como o peso médio ponderado dos k_i vizinhos do i-ésimo indivíduo, isto é,

$$z_i = \frac{\sum_{j=1}^{k_i} w_{ij} y_j}{\sum_{j=1}^{k_i} w_{ij}}$$

sendo que $w_{ij} = \frac{1}{h_{ij}}$, em que h_{ij} é a distância euclidiana entre as observações i e j. Considerando-se a distância entre indivíduos de uma unidade, no caso de estrutura de vizinhança de primeira ordem, tem-se que

$$z_i = \frac{1}{4}(y_{i_1} + y_{i_2} + y_{i_3} + y_{i_4}). \tag{1}$$

Uma forma alternativa para definir a covariável $\mathbf{z}_i$, usada por Gumpertz, Grahan e Ristino (1997) considera a estrutura de vizinhança separadamente, isto é, efeitos de linhas, colunas e diagonais (diagonal A e diagonal B), fornecendo assim, informações de efeitos direcionais. Logo, para estrutura de vizinhança de primeira ordem,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i_1} + \ldots + \beta_p x_{i_p} + \gamma_1 L_i + \gamma_2 C_i,$$

em que $\boldsymbol{\gamma}^T = (\gamma_1\ \gamma_2)$, $\mathbf{z}_i^T = (L_i\ C_i)$, γ_1 e γ_2 são os parâmetros associados à informação dos vizinhos nas linhas $L_i = \dfrac{y_{i_1} + y_{i_2}}{2}$ e colunas $C_i = \dfrac{y_{i_3} + y_{i_4}}{2}$.

0.4 Estimação

A estimação dos parâmetros do modelo logístico de regressão, considerando que as observações são independentes, é feita, geralmente, pelo método da máxima verossimilhança, ML.

No caso da modelagem de observações correlacionadas espacialmente com o uso do modelo de regressão autologístico não é possível escrever a função de verossimilhança de forma fechada, sendo, portanto, em geral, desconhecida uma expressão analítica para a constante de normalização. Diversos métodos aproximados foram, então, propostos para estimação dos parâmetros desse modelo, tais como máxima pseudo-verossimlhança, MPL, e “coding”, COD, (BESAG, 1972), “bootstrap” (BESAG, 1977), equações de estimação (BESAG, 1986), máxima verossimilhança com simulação Monte Carlo, MCL, (GEYER, 1991, GEYER, 1992, GEYER, 1994, WU; HUFFER, 1997, HUFFER; WU, 1998, GRIFFITH, 2002, SHERMAN, APANOSOVICH, CARROLL, 2006), máxima verossimilhança com simulação Monte Carlo via cadeias de Markov, MCMC (GU; KONG, 1998, GU; ZHU, 2001, WARD; GLEDITSCH, 2002), máxima pseudo-verossimlhança generalizada, MGPL, (HUANG; OGATA, 2002), dentre outros. Um método computacional estatisticamente eficiente foi desenvolvido para o cálculo da constante de normalização por Pettitt, Friel, Reeves (2003).

A estimativa da máxima pseudo-verossimilhança para um vetor de parâmetros desconhecidos $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \boldsymbol{\gamma})^T$ é definida como um vetor $\hat{\boldsymbol{\theta}}$ que maximize a função de pseudo-verossimilhança (PETRUCCI; SALVATI; SEGHIERI, 2004)

$$L(\boldsymbol{\pi}) = \prod_{i=1}^{n} P(Y_i = 1|\text{vizinhos}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{(1-y_i)} \tag{2}$$

ou o seu logaritmo

$$\ell(\boldsymbol{\pi}) = \sum_{i=1}^{n} y_i \log \pi_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i).$$

Para o modelo autologístico, essa aproximação é computacionalmente simples, pois necessita apenas de algum “software” que tenha regressão logística. Entretanto, os erros padrões das estimativas dos parâmetros são calculados, como se as observações fossem independentes (PETRUCCI; SALVATI; SEGHIERI, 2004). Uma proposta feita por Gumpertz, Grahan e Ristino (1997) é usar o método de reamostragem “bootstrap” paramétrico com amostrador de Gibbs que gera estimativas consistentes com boa precisão. A fim de preservar a dependência espacial dos dados originais, gera-se um determinado número de látices, pelo amostrador de Gibbs, usando-se como parâmetros as estimativas dos parâmetros obtidas a partir de um conjunto de observações. A seguir, é usado o método “bootstrap” para a obtenção de uma seqüência de estimativas MPL e, então, os erros padrões das estimativas dos parâmetros.

O método de pseudo-verossimilhança, de acordo com Ward; Gleiditsch (2002) é fácil de implementar, mais eficiente do que COD e mostra propriedades assintóticas razoáveis. Entretanto, tende a ser ineficiente quando há forte correlação espacial.

0.5 Um estudo de simulação

A fim de verificar o efeito causado por diferentes estruturas de covariáveis e dependência espacial sobre os estimadores de pseudo-verossimilhança dos parâmetros do modelo autologístico, é proposto um estudo de simulação, considerando-se um látice de 20 por 20 com distância de uma unidade entre pontos. Foram feitas 1.000 simulações, usando-se os pacotes estatísticos GeoR (RIBEIRO JR.; DIGGLE, 2001) e Rcitrus (KRAINSKI; RIBEIRO JR., 2006), do R. Para cada simulação foram seguidos os passos que se seguem.

Inicialmente, foram gerados valores para duas covariáveis X_1 e X_2, para três situações, a saber independendentes, com dependência espacial e não correlacionadas e com dependência espacial e correlacionadas. Os valores para X_1 e X_2 foram gerados a partir de uma distribuição normal de média 0 e variância 1. No segundo caso, também foi suposta distribuição normal de média 0 e variância 1, para ambas as variáveis e para a dependência espacial foram usados, respectivamente, para gerar X_1 e X_2, os valores 5 e 7 unidades para o parâmetro referente a alcance. Na terceira situação, X_1 foi gerada com distribuição normal de média 0 e variância 1 e parâmetro de alcance igual a 6 unidades enquanto que $X_2 = 0,9X_1 + 0,3\epsilon$, com ϵ gerado a partir de uma distribuição normal de média zero e variância 1. Isso faz com que a correlação entre X_1 e X_2 esteja em torno de $0,9$.

Em uma segunda etapa, foram calculados valores para p, usando-se

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}.$$

A fim de se terem porcentagens de infestação em torno de 10% (baixa), 30% (média) e 50% (alta), os valores usados para os parâmetros (β_0, β_1, β_2) foram, respectivamente, (0, 1, 1), (-1, 0,25, 0,25) e (-3, -1, -1).

Em uma terceira etapa, foram calculados valores para a covariável z, usando-se a expressão (1) com os valores de y_{i_r}, $r = 1, \ldots, 8$, substituídos por valores de p obtidos na segunda etapa.

Em uma quarta etapa, foram recalculados os valores de p, considerando-se

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma z)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma z)}$$

em que o parâmetro γ assumiu os valores 0,00, 0,25, 0,50, 0,75 e 1,00.

Em uma quinta etapa, foram calculados novos valores para z e valores de π dados por

$$\pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma z)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma z)}.$$

Em uma sexta etapa, foram gerados valores para a variável resposta Y a partir de uma distribuição Bernoulli com probabilidade de sucesso π.

A cada conjunto de dados gerados nas 90 situações, foram ajustados três modelos, considerando

M1: logit$(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma z$

M2: logit$(\pi) = \beta_0 + \gamma z$

M3: logit$(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Note que M1 é o modelo usado na geração dos dados, enquanto que M2 e M3 estão sendo usados para verificar o efeito do uso de modelos incompletos. Modelos semelhantes foram ajustados aos dados, considerando estrutura espacial de segunda ordem, mostrando resultados semelhantes aos de primeira ordem.

Foram, então, obtidas médias e erros padrões e médias dos erros padrões fornecidos pelo ajuste das 1000 estimativas de cada parâmetro.

0.6 Resultados e discussão do estudo de simulação

As médias das estimativas de cada parâmetro (Est), os erros padrões (EPC) e as médias dos erros padrões fornecidos pelo ajuste das 1000 simulações (EP) estão apresentados nas Tabelas 1 a 3.

De uma forma geral, nota-se que as médias das estimativas dos parâmetros β_1 e β_2 têm valores não muito distantes dos valores verdadeiros, mas com variações dependendo da correlação espacial γ verdadeira, e da forma como as covariáveis foram geradas. Ainda, nota-se que os erros padrões de suas estimativas são muito próximos da média dos erros padrões fornecidos pelo modelo. Observa-se, ainda uma influência pequena nas médias das estimativas dos parâmetros β_1 e β_2

obtidas pelos diferentes modelos mostrando uma robustez no processo de estimação. Entretanto, as médias das estimativas do parâmetro γ têm uma disparidade muito grande em relação ao valor γ com o qual foram gerados os dados. De forma semelhante existem disparidades entre ECP e EP. A seguir são feitos comentários mais específicos.

Observa-se que quando as covariáveis foram geradas sem correlação e sem dependência espacial, de uma forma geral, as médias das estimativas dos parâmetros β_1 e β_2 são muito próximas dos valores verdadeiros para todos os casos. Observa-se ainda, que a média das estimativas de γ aumentam à medida em que aumenta a correlação entre os vizinhos. Verifica-se ainda que os erros padrões das estimativas têm valores muito próximos da média das estimativas dos erros padrões dados pelo modelo, embora aumentem à medida que aumenta o valor da correlação espacial usada na geração dos dados.

Quando as covariáveis foram geradas sem correlação e com dependência espacial, observa-se que, de uma forma geral, as médias das estimativas dos parâmetros β_0, β_1, β_2 e γ aumentam à medida que aumenta o valor da correlação espacial. Nota-se ainda que, no modelo completo, as médias das estimativas de β_0 são mais próximas dos valores verdadeiros, para o caso de baixa infestação, quando a correlação entre vizinhos é maior e são mais próximas dos valores verdadeiros quando a correlação entre vizinhos é menor para média e alta infestação. Verifica-se também, que os erros padrões das estimativas têm valores muito próximos da média das estimativas dos erros padrões dados pelo modelo.

Observa-se que quando as covariáveis foram geradas com correlação e com dependência espacial, de uma forma geral, as médias das estimativas dos parâmetros β_1 e γ aumentam à medida em que aumenta a correlação entre vizinhos, já as médias das estimativas de β_2 são muito próximas dos valores verdadeiros. Observa-se ainda, que a média das estimativas de β_0 não alteram muito no caso de baixa infestação e aumentam à medida em que aumenta a correlação entre os vizinhos para o caso de média e alta infestação. Verifica-se ainda que os erros padrões das estimativas têm valores próximos da média das estimativas dos erros padrões dados pelo modelo.

Tabela 1 – Valores obtidos para a média das estimativas dos parâmetros a partir de 1000 simulações, supondo duas covariáveis independentes e sem dependência espacial, para baixa ($\beta_0 = -3,00$, $\beta_1 = -1,00$, $\beta_2 = -1,00$), média ($\beta_0 = -1,00$, $\beta_1 = 0,25$, $\beta_2 = 0,25$) e alta ($\beta_0 = 0,00$, $\beta_1 = 1,00$, $\beta_2 = 1,00$) infestação e γ v = 0,00; 0,25; 0,50; 0,75 e 1,00 considerando quatro vizinhos

		γ v														
			0,00			0,25			0,50			0,75			1,00	
Mod1	Par	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP
M1	β_0	-3,10	0,41	0,36	-3,08	0,41	0,35	-3,05	0,41	0,35	-3,03	0,40	0,35	-3,01	0,39	0,34
	β_1	-1,03	0,24	0,23	-1,03	0,24	0,23	-1,03	0,24	0,23	-1,03	0,24	0,23	-1,02	0,24	0,22
	β_2	-1,05	0,25	0,24	-1,05	0,25	0,24	-1,05	0,24	0,24	-1,05	0,24	0,24	-1,04	0,24	0,24
	γ	-0,87	6,34	47,45	-0,62	4,90	25,83	-0,69	5,84	36,28	-0,50	4,89	25,06	-0,41	4,96	27,34
M2	β_0	-2,23	0,24	0,23	-2,22	0,24	0,22	-2,20	0,24	0,22	-2,19	0,24	0,22	-2,17	0,24	0,22
	γ	-1,14	5,93	32,36	-0,88	4,65	19,51	-0,91	5,51	26,95	-0,70	4,63	19,07	-0,59	4,64	18,84
M3	β_0	-3,10	0,36	0,32	-3,07	0,35	0,32	-3,04	0,35	0,31	-3,01	0,34	0,31	-2,98	0,33	0,31
	β_1	-1,03	0,24	0,23	-1,02	0,24	0,23	-1,02	0,24	0,23	-1,02	0,23	0,22	-1,01	0,23	0,22
	β_2	-1,04	0,25	0,24	-1,04	0,24	0,24	-1,04	0,24	0,24	-1,04	0,24	0,23	-1,04	0,24	0,23
M1	β_0	-1,02	0,26	0,20	-0,94	0,26	0,21	-0,87	0,26	0,21	-0,78	0,26	0,21	-0,69	0,26	0,21
	β_1	0,26	0,12	0,12	0,25	0,12	0,12	0,25	0,11	0,12	0,25	0,11	0,12	0,25	0,11	0,11
	β_2	0,26	0,14	0,14	0,25	0,14	0,13	0,25	0,14	0,13	0,25	0,13	0,13	0,25	0,13	0,13
	γ	-0,04	0,83	0,59	-0,05	0,78	0,57	-0,03	0,75	0,55	-0,02	0,71	0,53	-0,02	0,70	0,52
M2	β_0	-0,98	0,26	0,20	-0,91	0,25	0,20	-0,84	0,25	0,20	-0,76	0,25	0,20	-0,67	0,26	0,21
	γ	-0,06	0,82	0,57	-0,06	0,77	0,56	-0,03	0,74	0,54	-0,01	0,71	0,52	0,00	0,70	0,51
M3	β_0	-1,02	0,13	0,13	-0,95	0,13	0,13	-0,87	0,13	0,12	-0,78	0,13	0,12	-0,69	0,12	0,12
	β_1	0,26	0,12	0,12	0,25	0,12	0,12	0,25	0,11	0,12	0,25	0,11	0,11	0,25	0,11	0,11
	β_2	0,26	0,14	0,14	0,25	0,14	0,13	0,25	0,14	0,13	0,25	0,13	0,13	0,25	0,13	0,13
M1	β_0	0,05	0,36	0,30	0,13	0,37	0,31	0,22	0,39	0,32	0,33	0,43	0,34	0,45	0,47	0,36
	β_1	1,03	0,16	0,15	1,03	0,16	0,15	1,03	0,16	0,15	1,03	0,16	0,15	1,04	0,16	0,16
	β_2	1,03	0,17	0,17	1,03	0,17	0,17	1,02	0,17	0,17	1,02	0,17	0,17	1,02	0,18	0,17
	γ	-0,09	0,67	0,54	0,00	0,67	0,54	0,10	0,68	0,54	0,18	0,71	0,54	0,24	0,75	0,55
M1	β_0	0,08	0,31	0,25	0,12	0,33	0,26	0,15	0,34	0,27	0,21	0,37	0,29	0,27	0,41	0,30
	γ	-0,17	0,59	0,45	-0,06	0,59	0,45	0,07	0,60	0,45	0,17	0,62	0,46	0,25	0,66	0,47
M3	β_0	0,00	0,13	0,13	0,13	0,13	0,13	0,27	0,13	0,13	0,43	0,14	0,14	0,59	0,14	0,14
	β_1	1,03	0,16	0,15	1,03	0,16	0,15	1,03	0,15	0,15	1,03	0,16	0,15	1,04	0,16	0,16
	β_2	1,02	0,17	0,16	1,02	0,17	0,16	1,02	0,17	0,17	1,02	0,17	0,17	1,02	0,18	0,17

Tabela 2 – Valores obtidos para a média das estimativas dos parãmetros a partir de 1000 simulações, supondo duas covariáveis independentes e sem dependência espacial, para baixa ($\beta_0 = -3,00$, $\beta_1 = -1,00$, $\beta_2 = -1,00$), média ($\beta_0 = -1,00$, $\beta_1 = 0,25$, $\beta_2 = 0,25$) e alta ($\beta_0 = 0,00$, $\beta_1 = 1,00$, $\beta_2 = 1,00$) infestação e γ v = 0,00; 0,25; 0,50; 0,75 e 1,00 considerando quatro vizinhos

		γ v														
		0,00			0,25			0,50			0,75			1,00		
		Est	EPC	EP	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP
M1	β_0	-3,09	0,38	0,35	-3,07	0,39	0,35	-3,05	0,39	3,45	-3,03	0,38	0,34	-3,02	0,39	0,34
	β_1	-1,08	0,33	0,29	-1,09	0,33	0,29	-1,11	0,33	0,29	-1,12	0,32	0,29	-1,14	0,32	0,29
	β_2	-1,08	0,31	0,29	-1,10	0,31	0,29	-1,12	0,31	0,29	-1,15	0,32	0,29	-1,17	0,32	0,29
	γ	-0,66	2,85	6,10	-0,53	1,78	1,22	-0,53	2,91	8,29	-0,38	1,56	1,14	-0,29	1,54	1,10
M2	β_0	-2,61	0,27	0,25	-2,60	0,27	0,25	-2,58	0,26	0,25	-2,57	0,26	0,24	-2,56	0,26	0,24
	γ	2,33	2,70	6,13	2,50	1,53	1,03	2,53	2,62	5,58	2,71	1,32	0,94	2,83	1,29	0,90
M3	β_0	-3,10	0,36	0,34	-3,08	0,36	0,34	-3,06	0,36	0,33	-3,04	0,37	0,33	-3,02	0,36	0,33
	β_1	-1,04	0,30	0,28	-1,06	0,30	0,27	-1,08	0,29	0,27	-1,10	0,29	0,27	-1,12	0,28	0,27
	β_2	-1,04	0,27	0,27	-1,07	0,27	0,27	-1,09	0,27	0,27	-1,12	0,27	0,27	-1,14	0,27	0,27
M1	β_0	-0,98	0,25	0,20	-0,91	0,25	0,21	-0,83	0,25	0,21	-0,74	0,26	0,21	-0,65	0,28	0,21
	β_1	0,26	0,14	0,14	0,27	0,14	0,13	0,28	0,14	0,13	0,30	0,14	0,13	0,31	0,14	0,13
	β_2	0,26	0,14	0,13	0,27	0,14	0,13	0,29	0,14	0,13	0,30	0,14	0,13	0,31	0,14	0,13
	γ	-0,18	0,84	0,59	-0,17	0,80	0,58	-0,15	0,77	0,56	-0,14	0,75	0,54	-0,12	0,73	0,53
M2	β_0	-1,09	0,25	0,20	-1,04	0,24	0,20	-0,97	0,24	0,20	-0,91	0,25	0,20	-0,84	0,26	0,20
	γ	0,27	0,82	0,57	0,30	0,77	0,55	0,35	0,74	0,53	0,41	0,71	0,51	0,45	0,70	0,49
M3	β_0	-1,02	0,13	0,13	-0,95	0,13	0,13	-0,87	0,13	0,12	-0,78	0,12	0,12	-0,69	0,12	0,12
	β_1	0,25	0,14	0,13	0,26	0,13	0,13	0,27	0,13	0,13	0,29	0,13	0,13	0,30	0,13	0,13
	β_2	0,25	0,13	0,13	0,26	0,13	0,13	0,28	0,13	0,13	0,30	0,13	0,13	0,31	0,13	0,13
M1	β_0	0,08	0,40	0,31	0,18	0,42	0,32	0,29	0,43	0,34	0,40	0,45	0,36	0,53	0,47	0,38
	β_1	1,04	0,21	0,18	1,08	0,21	0,18	1,12	0,22	0,19	1,15	0,22	0,19	1,19	0,23	0,20
	β_2	1,06	0,23	0,19	1,09	0,23	0,19	1,12	0,23	0,20	1,16	0,23	0,21	1,20	0,24	0,21
	γ	-0,15	0,74	0,56	-0,08	0,76	0,57	-0,02	0,75	0,57	0,05	0,76	0,58	0,10	0,75	0,59
M2	β_0	-1,30	0,26	0,23	-1,32	0,26	0,24	-1,34	0,26	0,24	-1,36	0,27	0,25	-1,36	0,27	0,26
	γ	2,51	0,52	0,41	2,64	0,51	0,41	2,77	0,51	0,41	2,90	0,51	0,42	3,01	0,51	0,42
M3	β_0	0,00	0,13	0,13	0,13	0,13	0,13	0,28	0,14	0,14	0,42	0,14	0,14	0,59	0,15	0,15
	β_1	1,01	0,15	0,15	1,06	0,15	0,15	1,11	0,16	0,16	1,16	0,16	0,16	1,21	0,18	0,17
	β_2	1,03	0,17	0,16	1,07	0,17	0,16	1,12	0,17	0,17	1,17	0,18	0,17	1,21	0,19	0,18

Tabela 3 – Valores obtidos para a média das estimativas dos parãmetros a partir de 1000 simulações, supondo duas covariáveis dependentes e com dependência espacial, para baixa ($\beta_0 = -3,00$, $\beta_1 = -1,00$, $\beta_2 = -1,00$), média ($\beta_0 = -1,00$, $\beta_1 = 0,25$, $\beta_2 = 0,25$) e alta ($\beta_0 = 0,00$, $\beta_1 = 1,00$, $\beta_2 = 1,00$) infestação e γ v =0,00; 0,25; 0,50; 0,75 e 1,00 considerando quatro vizinhos

		γ v														
		0,00			0,25			0,50			0,75			1,00		
		Est	EPC	EP	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP	Est	EPC	EP
M1	β_0	-3,09	0,38	0,35	-3,07	0,37	0,35	-3,06	0,37	0,34	-3,05	0,38	0,34	-3,03	0,37	0,34
	β_1	-1,12	0,75	0,70	-1,16	0,72	0,69	-1,19	0,73	0,69	-1,22	0,74	0,68	-1,25	0,73	0,68
	β_2	-0,99	0,71	0,69	-0,99	0,69	0,68	-1,00	0,69	0,68	-1,00	0,69	0,67	-1,01	0,68	0,67
	γ	-0,24	1,35	1,00	-0,20	1,33	0,97	-0,09	1,31	0,95	0,03	1,25	0,92	0,11	1,22	0,90
M2	β_0	-2,52	0,24	0,24	-2,51	0,24	0,24	-2,51	0,24	0,24	-2,51	0,23	0,24	-2,51	0,24	0,24
	γ	3,24	1,00	0,76	3,31	0,98	0,74	3,43	0,95	0,72	3,54	0,90	0,70	3,62	0,86	0,68
M3	β_0	-3,08	0,36	0,34	-3,06	0,36	0,34	-3,04	0,36	0,33	-3,03	0,36	0,33	-3,01	0,36	0,33
	β_1	-1,08	0,68	0,66	-1,12	0,66	0,65	-1,17	0,66	0,65	-1,23	0,67	0,65	-1,28	0,65	0,64
	β_2	-0,99	0,70	0,68	-0,99	0,68	0,68	-0,99	0,68	0,67	-0,99	0,68	0,66	-1,00	0,67	0,66
M1	β_0	-0,99	0,27	0,21	-0,92	0,26	0,21	-0,84	0,27	0,21	-0,75	0,27	0,21	-0,65	0,28	0,22
	β_1	0,24	0,45	0,43	0,26	0,44	0,43	0,29	0,43	0,42	0,33	0,41	0,42	0,36	0,41	0,41
	β_2	0,27	0,46	0,44	0,28	0,46	0,44	0,27	0,44	0,43	0,27	0,43	0,43	0,27	0,42	0,42
	γ	-0,12	0,84	0,59	-0,10	0,78	0,57	-0,09	0,78	0,56	-0,09	0,75	0,55	-0,09	0,72	0,53
M2	β_0	-1,15	0,25	0,20	-1,11	0,25	0,20	-1,06	0,26	0,20	-1,01	0,26	0,20	-0,96	0,26	0,20
	γ	0,57	0,78	0,54	0,64	0,72	0,52	0,71	0,72	0,50	0,77	0,69	0,49	0,84	0,66	0,47
M3	β_0	-1,02	0,13	0,13	-0,94	0,13	0,13	-0,86	0,13	0,13	-0,77	0,13	0,12	-0,68	0,12	0,12
	β_1	0,23	0,44	0,43	0,25	0,43	0,42	0,29	0,42	0,42	0,32	0,40	0,41	0,35	0,40	0,41
	β_2	0,27	0,45	0,44	0,28	0,45	0,43	0,27	0,44	0,43	0,27	0,43	0,42	0,27	0,42	0,42
M1	β_0	0,04	0,42	0,33	0,15	0,43	0,34	0,26	0,45	0,35	0,36	0,46	0,37	0,48	0,48	0,38
	β_1	1,02	0,51	0,51	1,07	0,52	0,51	1,13	0,54	0,52	1,19	0,54	0,52	1,27	0,54	0,53
	β_2	1,06	0,50	0,50	1,07	0,49	0,50	1,09	0,50	0,51	1,08	0,50	0,51	1,08	0,52	0,52
	γ	-0,08	0,81	0,61	-0,04	0,81	0,61	0,02	0,81	0,61	0,11	0,80	0,61	0,17	0,80	0,61
M2	β_0	-1,63	0,26	0,23	-1,64	0,25	0,24	-1,64	0,25	0,24	-1,66	0,26	0,25	-1,67	0,26	0,25
	γ	3,17	0,51	0,41	3,25	0,50	0,41	3,33	0,50	0,42	3,43	0,49	0,42	3,52	0,50	0,42
M3	β_0	0,00	0,14	0,14	0,02	0,20	0,15	0,27	0,15	0,15	0,41	0,16	0,16	0,57	0,16	0,16
	β_1	0,99	0,46	0,46	1,06	0,46	0,47	1,13	0,47	0,48	1,22	0,48	0,48	1,32	0,49	0,49
	β_2	1,06	0,49	0,50	1,07	0,49	0,50	1,08	0,50	0,51	1,08	0,50	0,51	1,07	0,51	0,52

0.6.1 Aplicação

Uma aplicação da metodologia foi feita usando-se dados de presença/ausência do patógeno *Phytophthora capsici* em *quadrats* de pimenta de sino, tendo como covariáveis o conteúdo de água no solo e o número de discos de folhas colonizados pelo patógeno, apresentados em Gumpertz, Graham e Ristiano (1997). É importante observar que a porcentagem de infecção nesse caso é de $13,5\%$, o que ser considerado como baixa. Além disso, a correlação estimada entre X_1 e X_2 é apenas 0,27. A esse conjunto de dados, foram ajustados os cinco modelos usados no estudo de simulação, considerando X_1 como a covariável conteúdo de água no solo e X_2 como o número de discos de folhas colonizados pelo patógeno e o método de pseudo-verossimilhança para estimação dos parâmetros. Embora, no artigo original tenha sido considerada borda dupla, usou-se borda simples aqui. Os valores faltantes da covariável X_1 foram estimados usando-se as expressões (1) e (**??**) para estruturas de vizinhança de primeira e segunda ordens, respectivamente. A seleção de modelos foi feita usando-se o critério de informação de Akaike (AIC), dado por: $-2 * LVM + 2 * p$, em que p representa o número de parâmetros e LVM é o logaritmo da função de verossimilhança, considerando-se como melhor modelo aquele com menor AIC.

Os resultados obtidos para o ajuste dos modelos ao conjunto original de dados estão nas Tabelas 4 e 5. Pelo critério de informação de Akaike, considerando-se estrutura de vizinhança de primeira ordem foi escolhido o modelo que inclui apenas a constante e o efeito de linha, enquanto que estrutura de vizinhança de segunda ordem, o melhor modelo inclui a constante, e os efeitos de linha e da diagonal B. Verifica-se, portanto, que nenhuma covariável foi significativa, como obtido por Gumpertz; Graham; Ristaino (1997). A probabilidade de um “quadrat” ter presença da doença é dada por

$$P(Y_i = 1/y_j, j \neq i) = \frac{\exp\{-2,83 + 1,07dB + 1,29L\}}{1 + \exp\{-2,83 + 1,07dB + 1,29L\}}$$

Tabela 4 – Estimativas dos parâmetros e estatísticas dos diversos modelos ajustados aos dados originais de pimenta do sino, com estrutura de vizinhança de primeira ordem

Modelo	Parâmetro	Estimativa	Erro Padrão	Z	$valor - p$	AIC
M1	β_0	-3,29	1,10	-3,00	**0,003**	245,16
	β_1	-0,05	0,13	-0,35	0,73	
	β_2	0,08	0,10	0,75	0,454	
	γ	3,57	0,67	5,31	**0,000**	
M2	β_0	-2,50	0,24	-10,51	**0,000**	241,73
	γ	3,56	0,66	5,38	**0,000**	
M3	β_0	-2,66	1,07	-2,48	**0,013**	272,85
	β_1	0,08	0,12	0,65	0,519	
	β_2	0,08	0,10	0,77	0,440	
M4	β_0	-3,07	1,11	-2,75	**0,006**	243,00
	β_1	-0,04	0,14	-0,262	0,793	
	β_2	0,06	0,10	0,54	0,591	
	γ_1	1,30	0,26	4,93	**0,000**	
	γ_2	0,36	0,32	1,12	0,26	
M5	β_0	-2,49	0,24	-10,47	**0,000**	239,30
	γ_1	1,31	0,26	5,01	**0,000**	
	γ_2	0,34	0,32	1,08	0,279	
M5	β_0	-2,41	0,22	-10,85	**0,000**	238,43
	β_1	1,41	0,25	5,73	**0,000**	

Tabela 5 – Estimativas dos parâmetros e estatísticas dos diversos modelos ajustados aos dados originais de pimenta do sino, com estrutura de vizinhança de segunda ordem

Modelo	Parâmetro	Estimativa	Erro Padrão	Z	$valor - p$	AIC
M1	β_0	-3,74	1,13	-3,30	**0,000**	233,86
	β_1	-0,11	0,14	-0,79	0,433	
	β_2	0,09	0,10	0,86	0,390	
	γ	5,22	0,88	5,96	**0,000**	
M2	β_0	-2,82	0,27	-10,28	**0,000**	230,88
	γ	5,09	0,85	6,02	**0,000**	
M3	β_0	-2,68	1,08	-2,49	**0,013**	272,82
	β_1	0,08	0,12	0,64	0,52	
	β_0	0,08	0,10	0,79	0,429	
M4	β_0	-3,60	1,19	-3,04	**0,002**	229,31
	β_1	-0,10	0,15	-0,70	0,485	
	β_2	0,06	0,11	0,60	0,550	
	γ_3	0,57	0,33	1,71	0,088	
	γ_4	1,04	0,28	3,75	**0,000**	
	γ_1	1,25	0,28	4,53	**0,000**	
	γ_2	-0,15	0,37	-0,41	0,682	
M5	β_0	-2,94	0,29	-10,16	**0,000**	225,94
	γ_3	0,56	0,33	1,68	0,092	
	γ_4	1,02	0,27	3,72	**0,000**	
	γ_1	1,25	0,27	4,58	**0,000**	
	γ_2	-0,19	0,37	-0,51	0,61	
M5	β_0	-2,83	0,27	-10,48	**0,000**	224,74
	γ_4	1,07	0,27	4,00	**0,000**	
	γ_1	1,29	0,25	5,12	**0,000**	

0.7 Considerações finais

O estudo de simulação com o objetivo de verificar o efeito causado por diferentes estruturas de covariáveis e dependência espacial sobre os estimadores de pseudo-verossimilhança dos parâmetros do modelo autologístico permitiu verificar que as médias das estimativas dos parâmetros associados às covariáveis têm valores não muito distantes dos valores verdadeiros, mas com variações dependendo da correlação espacial, e da forma como as covariáveis foram geradas, mostrando uma robustez no processo de estimação. Os erros padrões de suas estimativas são muito próximos da média dos erros padrões fornecidos pelo modelo. Entretanto, as médias das estimativas do parâmetro de correlação espacial têm uma disparidade muito grande em relação ao valor verdadeiro com o qual foram gerados os dados. De forma semelhante existem disparidades entre o erro padrão obtido a partir das estimativas dos parâmetros e a média dos erros padrões fornecidos pelo modelo.

As médias das estimativas dos parâmetros, geralmente, aumentam com o aumento da correlação espacial, evidenciando a presença de um pequeno vício, que praticamente desaparece no caso em que as covariáveis não são correlacionadas e não têm dependência espacial.

O coeficiente de correlação espacial é estimado com vício muito grande, fazendo com que a correlação espacial se torne muito maior do que o valor verdadeiro.

Portanto, o método de estimação por pseudo-verossimilhança pode ser usado, com certa cautela, quando o interesse está na contribuição das covariáveis, mas não deve ser usado quando o interesse está na estimação da correlação espacial.

Estudos adicionais por simulação são necessários para verificar o efeito de observações faltantes nas estimativas dos parâmetros do modelo autologístico.

0.7.1 Referências

ABEL, L; GOLMARD, J. L.; MALLET, A. An autologistic model for the Genetic Analysis of familial Binary data. **American Society of Human Genetics**, San Francisco, v. 53, p. 894-907, 1993. Bethesda, MD.

AUGUSTIN, N. H.; MUGGLESTONE, M. A.; BUCKLAND, S. T. An autologistic model for the spatial distribution of wildlife. **Journal of Applied Ecology**, v. 33, p. 339-347, 1996. http://www.britishecologicalsociety.org/art

BESSAG, J. Nearest-neighbour systems and the auto-logistic model for binary data (with discussion). **Journal of the Royal Statistical Society**, v. 34, p. 75-83, 1972.

BESAG, J. Efficience of pseudolikelihood estimators for simple Gaussian fields. **Biometrika**, v. 64, p. 616-618, 1977.

BESAG, J. On the statistical analusis of dirty pictures (with discussion). **Journal Royal Statistic Society**, v. 48, p. 259-302,1986.

DEMÉTRIO, C. G. B. Modelos lineares generalizados In: 46^{O} Reunião Anual da RBRAS, 2001, Piracicaba. **Minicurso** Piracicaba, 2001, 1-113.

GEYER, C. J. Markov Chain Monte Carlo maximum likelihood. **Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface**, p. 156-163, 1991.

GEYER C. J. Pratical Markov chain Monte Carlo (with discussion). **Statistical Science**, v. 7(4), p. 473-511, 1992.

GEYER, C. J. On the convergence Of Monte Carlo Maximum likelihood calculations. **Journal of the Royal Statistical Society**, v. 56, p. 261-274, 1994.

GRIFFITH, D. A. A spatial filtering specification for the autologistic model. Enviromental and Planning A, v. 36, p. 1791-1811, 2002.

GU, M. G.; KONG, F. H. A stochastic aproximation algorithm with Markov Chain Monte Carlo method for incomplete data estimation problems **proceedings of National Academic Science of USA** v. 95, p. 7270-7274, 1998.

GU, M. G.; ZHU, H. T. Maximum likelihood estimation for spatial models by Markov Chain Monte Carlo Stocastic Approximation. **Journal of the Royal Statistics Society**, v. 63, p. 339-355, 2001.

GUMPERTZ M. L.; GRAHAM, J. M; RISTANO, J. B. Autologistic model of spatial pattern of Phytophthora epidemic in bell pepper: effects of soil variables on disease presence. Journal of agricultural, biological end Environmental Statistics, v. 2, p. 131-156, 1997.

HE, F.; ZHOU, J.; ZHU, H. Autologistic regression model for the distribution of vegetation. **Journal of Agricultural, Biological and Environmental Statistics**, v. 8(2), p. 205-222, 2003.

HUANG, F.; OGATA, Y. Comparison of two methods for calculating the partition functions of various spatial statistical models **the Australian and New Zeland Journal of Statistics**, v. 43, p. 47-65, 2002.

HUFFER ; WU, 1998. MOON, S; RUSSEL, G. R. Predicting product purchase from inferred customer similarity: an autologistic model approach. 2005.

PETRUCCI, A. ; SALVATI, N.; SEGHIERI, C. Autologistic regression model for poverty mapping and analysis. Metodolo$\widehat{s}$ki Zveszki, v. 1(1), p. 225-234, 2004.

PETTITT, A. N.; FRIEL, N.; REEVES, R. Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. **Journal Royal Statistical Society**, v. 65(1), p. 235-246, 2003.

SANDERSON, R. A.; EYRE, M. D.; RUSHTON, S. P.; Distribution of selected macroinvertebrates in a mosaic of temporary and permanent freshwater ponds as explained by autologistic models **Ecographi**, v. 28, p. 355-362, 2005.

SHERMAN, M.; APANOSOVICH, T. V.; CARROLL, R. J. On estimation in binary autologistic spatial models **journal of Statistical Computation and Simulation**, v. 76(2), p. 167-179, 2006.

TETERUKOVSKIY, A.; EDEMIRS, L. Effective field sampling for predicting the spatial distribution of reindeeer (Rangifer tarandus) with help of the gibbs sampler **Ambio**, v. 32(8), p.568-572, 2003.

WARD, M.; GLEDITSCH, K. S. Location, location, location: An MCMC approach to modeling the spatial context of war and peace **Political Analysis**, v. 10(3), p. 244-260, 2002.

WU, H; HUFFER, F. W. Modelling distribution of plant species using the autolologistic regression model. **Environmental and ecological Statistics**, v. 4, p. 49-64, 1997.