

Universidade do Minho  
2ºSemestre 2019/20  
(MIEI, 3ºAno)

Modelos Estocásticos de Investigação Operacional  
Trabalho Prático

Identificação do Grupo

<u>Número:</u>	<u>Nome completo:</u>	<u>Rubrica:</u>
A81283	Hugo Filipe Oliveira de Sousa Faria	Hugo Faria
A80791	João Diogo Mendes Teixeira da Mota	João Diogo Mota
A82313	Pedro Teixeira Gonçalves	Pedro Gonçalves
A81716	Rodolfo António Vieira da Silva	Rodolfo Silva

Data de entrega: 2020- 05 - 11

---

## Conteúdo

1	Introdução	3
2	Caso de Estudo	3
3	Modelo Oculto de Markov (MOC) e Markov Chain Monte Carlo (MCMC)	5
4	Aplicação e Resultados	8

## Lista de Figuras

1	Mapa distribuição espacial Francolim-de-asa-cinzenta . . . . .	4
2	Lista variáveis com potencial impacto na distribuição espacial	
	das aves . . . . .	5
3	Algoritmo geração parâmetros desconhecidos . . . . .	6
4	Algoritmo geração parâmetros desconhecidos . . . . .	7
5	Algoritmo geração parâmetros desconhecidos . . . . .	8
6	Mapa restaurado da distribuição espacial Francolim-de-asa-	
	cinzenta . . . . .	9

## 1 Introdução

O artigo que o grupo escolheu para análise foi o **Spatial hidden Markov models and species distributions** [1], um artigo onde se discute a introdução de um modelo oculto de Markov para analisar a distribuição de uma espécie de aves num mapa, tendo em consideração a possível existência de erros de observação da espécie durante o processo de criação do mapa. O objetivo do estudo deste artigo é, utilizando algoritmos de Markov chain Monte Carlo e do paradigma de Bayesian, reconstruir o mapa original eliminando os erros existentes no original devido a más observações no processo de construção.

Neste artigo, os autores estão interessados na distribuição espacial da espécie Francolim-de-asa-cinzenta no Sul de África (Lesoto, África do Sul e Suazilândia), na seleção de variáveis que influenciem a presença da espécie numa localização e na escolha de variáveis que possam afetar a veracidade das probabilidades de deteção da espécie.

## 2 Caso de Estudo

O caso de estudo do artigo, é a distribuição espacial da ave Francolim-de-asa-cinzenta, tendo por base um mapa, apresentado em baixo, onde as áreas negras significam que foi avistado pelo menos um elemento da espécie e as brancas correspondem a áreas onde não foi observado qualquer elemento. Este mapa encontra-se dividido em 1949 áreas com  $625\text{km}^2$  cada, e o processo de recolha de dados para a sua criação foi efetuado pela **Southern African Bird Atlas Project** (SABAP) entre 1987 e 1991. Convém notar que este mapa pode conter incerteza sobre a veracidade da presença/absência das espécies nas diferentes áreas, devido ao facto de a recolha de dados ter sido feita maioritariamente por voluntários e as áreas de observação serem bastante extensas, sendo difícil cobrir com precisão a totalidade das diferentes áreas.



Figura 1 - Mapa distribuição espacial Francolim-de-asa-cinzenta.

O objetivo do estudo é relacionar o mapa da distribuição espacial com uma lista de variáveis espaciais que fornecem informações úteis sobre a potencial distribuição espacial da espécie. Essas variáveis representam desde fatores meteorológicos ao tipo de ecossistema de cada seção do mapa, de modo a ser possível perceber quais as áreas com o melhor ambiente para a presença da espécie. Para este estudo foram selecionadas 18 variáveis dessa lista, apresentada em baixo, sendo elas as [1...10], [31...34] e da conjunção das variáveis semelhantes [11,12,23], [13,24], [14,15,26,27] e [16,17,25].

	Explanatory variables	Min	Max	% of zeros
[1]	Average annual evapotranspiration	1062.25	1934.33	0
[2]	Cumulative growing degree days above 10 degrees (Jan–June)	0	263.25	31
[3]	Cumulative growing degree days above 10 degrees (Jan–Dec)	14.90	533.40	0
[4]	Cumulative growing degree days above 18 degrees (Jan–Dec)	0	285.40	2
[5]	Average annual precipitation	39.23	1293.92	0
[6]	Maximum annual precipitation (of monthly means)	6.51	232.59	0
[7]	Minimum annual precipitation (of monthly means)	0	76.8	0
[8]	Average annual temperature	6.40	23.63	0
[9]	Minimum of all minimum monthly temperatures of all the period recorded	−5.90	12.85	0
[10]	Maximum of all maximum monthly temperatures of all the period recorded	17.08	36.75	0
[11]	Forest and woodland	0	625	77
[12]	Forest	0	142	88
[13]	Thicket and bushland	0	625	27
[14]	Shrubland and low fynbos	0	625	51
[15]	Herbland	0	526	99
[16]	Unimproved grassland	0	625	41
[17]	Improved grassland	0	80	93
[18]	Forest plantations	0	493	81
[19]	Waterbodies	0	555	81
[20]	Wetlands	0	184	80
[21]	Barren rock	0	94	89
[22]	Dongas and sheet erosion scars	0	118	94
[23]	Degraded: forest and woodland	0	409	91
[24]	Degraded: thicket and bushland	0	615	79
[25]	Degraded: unimproved grassland	0	520	83
[26]	Degraded: shrubland and low fynbos	0	272	89
[27]	Degraded: herbland	0	1	99
[28]	Cultivated permanent – commercial irrigated	0	339	93
[29]	Cultivated permanent – commercial dryland	0	109	96
[30]	Cultivated permanent – commercial sugarcane	0	421	96
[31]	Cultivated temporary – commercial irrigated	0	213	67
[32]	Cultivated temporary – commercial dryland	0	600	55
[33]	Cultivated temporary – commercial sugarcane	0	553	74
[34]	Urban/built-up land: residential	0	324	65
[35]	Urban/built-up land: residential (small holdings: woodland)	0	131	99
[36]	Urban/built-up land: residential (small holdings: bushland)	0	40	98
[37]	Urban/built-up land: residential (small holdings: shubland)	0	29	99
[38]	Urban/built-up land: residential (small holdings: grassland)	0	167	97
[39]	Urban/built-up land: commercial	0	53	97
[40]	Urban/built-up land: industrial/transport	0	35	94
[41]	Mines and quarries	0	68	92

Figura 2 - Lista variáveis com potencial impacto na distribuição espacial das aves.

### 3 Modelo Oculto de Markov (MOC) e Markov Chain Monte Carlo (MCMC)

MOCs consistem em pares de processos estocásticos ( $\{X_i\}, \{Y_i\}$ ), onde  $\{X_i\}$  correspondem a processos Markovianos ocultos ou não observáveis e  $\{Y_i\}$  processos observáveis que satisfazem duas condições. Neste artigo é assumido que  $\{X_i\}$  é um modelo autologista onde as condições de distribuição

das observações são estados dependentes de Bernoulli.

Em seguida, para a geração dos parâmetros desconhecidos do MOC é utilizado o algoritmo MCMC explicado em baico. Devido a ser um algoritmo extenso, o grupo decidiu mostrar o mesmo como se encontrava descrito no artigo.

The MCMC algorithm can be summarized as follows.

- (1) Generate the starting configuration  $\mathbf{x}$  by drawing each cell  $x_i$  ( $i = 1, \dots, n$ ) from an independent Bernoulli of probability 0.5.
- (2) Generate the starting values  $\theta$ ,  $\lambda$ ,  $\alpha$ , and  $\beta$  from their respective prior densities, and set the entries of all indicators  $\omega$ ,  $\gamma$ , and  $\rho$  equal to 1.
- (3) Generate any  $x_i$  ( $i = 1, \dots, n$ ) by single-site Gibbs sampling:

$$p(x_i | x_{\eta_i}, y_i, \theta, \lambda, \alpha, \beta, \omega, \gamma, \rho, \mathbf{w}, \mathbf{z}) \propto p(y_i | x_i, \alpha, \beta, \gamma, \rho, \mathbf{z}) p(x_i | x_{\eta_i}, \theta, \lambda, \omega, \mathbf{w}).$$

- (4) Draw the proposals  $\alpha_0^{\text{new}}, \alpha_1^{\text{new}}, \dots, \alpha_q^{\text{new}}, \beta_0^{\text{new}}, \beta_1^{\text{new}}, \dots, \beta_q^{\text{new}}$  from independent random walks:

$$\alpha_h^{\text{new}} = \alpha_h + U_A \quad \beta_h^{\text{new}} = \beta_h + U_B,$$

where  $U_A \sim \mathcal{N}(0; \sigma_A^2)$  and  $U_B \sim \mathcal{N}(0; \sigma_B^2)$ , with  $h = 0, 1, \dots, q$ .

- (5) Propose to move from  $(\alpha, \beta)$  to  $(\alpha^{\text{new}}, \beta^{\text{new}})$ , with probability  $\min\{1; R_{A,B}\}$ , where

$$R_{A,B} = \frac{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \alpha^{\text{new}}, \beta^{\text{new}}, \gamma, \rho)}{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \alpha, \beta, \gamma, \rho)} \frac{p(\alpha^{\text{new}}) p(\beta^{\text{new}})}{p(\alpha) p(\beta)} \frac{g(\alpha, \beta | \alpha^{\text{new}}, \beta^{\text{new}})}{g(\alpha^{\text{new}}, \beta^{\text{new}} | \alpha, \beta)},$$

If the move is accepted, set  $\alpha = \alpha^{\text{new}}$ , and  $\beta = \beta^{\text{new}}$ .

- (6) Draw the proposals  $\theta_0^{\text{new}}, \theta_1^{\text{new}}, \lambda_1^{\text{new}}, \dots, \lambda_p^{\text{new}}$  from independent random walks:

$$\theta_0^{\text{new}} = \theta_0 + U_\Theta \quad \theta_1^{\text{new}} = \theta_1 + U_\Theta \quad \lambda_k^{\text{new}} = \lambda_k + U_\Lambda,$$

where  $U_\Theta \sim \mathcal{N}(0; \sigma_\Theta^2)$  and  $U_\Lambda \sim \mathcal{N}(0; \sigma_\Lambda^2)$ , with  $k = 1, \dots, p$ .

- (7) Draw the auxiliary lattice  $\mathbf{s}$  from  $p(\cdot | \theta^{\text{new}}, \lambda^{\text{new}}, \omega, \mathbf{w})$ . Use the Gibbs sample obtained at the  $(N + 1)$ th iteration, where  $N$  is a sufficiently large number of iterations.
- (8) Propose to move from  $[(\theta, \lambda, \mathbf{x}), (\theta^{\text{new}}, \lambda^{\text{new}}, \mathbf{s})]$  to  $[(\theta^{\text{new}}, \lambda^{\text{new}}, \mathbf{x}), (\theta, \lambda, \mathbf{s})]$ , with probability  $\min\{1; R_{\Theta,\Lambda}\}$ , where

$$R_{\Theta,\Lambda} = \frac{q(\mathbf{x} | \theta^{\text{new}}, \lambda^{\text{new}}, \omega, \mathbf{w}) / Z(\theta^{\text{new}}, \lambda^{\text{new}}, \omega, \mathbf{w})}{q(\mathbf{x} | \theta, \lambda, \omega, \mathbf{w}) / Z(\theta, \lambda, \omega, \mathbf{w})} \frac{p(\theta^{\text{new}})}{p(\theta)} \frac{p(\lambda^{\text{new}})}{p(\lambda)} \\ \times \frac{q(\mathbf{ws} | \theta, \lambda, \omega, \mathbf{w}) / Z(\theta, \lambda, \omega, \mathbf{w})}{q(\mathbf{s} | \theta^{\text{new}}, \lambda^{\text{new}}, \omega, \mathbf{w}) / Z(\theta^{\text{new}}, \lambda^{\text{new}}, \omega, \mathbf{w})} \frac{g(\theta, \lambda | \theta^{\text{new}}, \lambda^{\text{new}})}{g(\theta^{\text{new}}, \lambda^{\text{new}} | \theta, \lambda)}. \quad (9)$$

Figura 3 - Algoritmo geração parâmetros desconhecidos.

The acceptance ratio (9) becomes

$$R_{\Theta, \Lambda} = \frac{q(\mathbf{x} | \boldsymbol{\theta}^{\text{new}}, \boldsymbol{\lambda}^{\text{new}}, \boldsymbol{\omega}, \mathbf{w})}{q(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w})} \frac{q(\mathbf{s} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w})}{q(\mathbf{s} | \boldsymbol{\theta}^{\text{new}}, \boldsymbol{\lambda}^{\text{new}}, \boldsymbol{\omega}, \mathbf{w})} \frac{p(\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{\theta})} \frac{p(\boldsymbol{\lambda}^{\text{new}})}{p(\boldsymbol{\lambda})},$$

after cancelling out the normalizing constants, and the proposal ratio of  $\theta_0, \theta_1, \lambda_1, \dots, \lambda_p$  because of the symmetry of the proposal density. If the exchange move is accepted, set  $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{new}}$  and  $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{new}}$ .

- (9) With probabilities  $\xi_1, \xi_2, \xi_u$ , choose randomly one of the following three mutation operators ( $\xi_1 + \xi_2 + \xi_u = 1$ ): 1-point mutation – Select at random one of the  $p$  indicators, for example,  $\omega_i$ , and reverse its value to obtain  $\omega_i^{\text{new}} = 1 - \omega_i$ , so that  $\boldsymbol{\omega}^{\text{new}} = (\omega_1, \dots, \omega_i^{\text{new}}, \dots, \omega_p)'$ . 2-point mutation – Select at random two of the  $p$  indicators, e.g.  $\omega_i$  and  $\omega_j$ , and reverse their values to obtain  $\boldsymbol{\omega}^{\text{new}} = (\omega_1, \dots, \omega_i^{\text{new}}, \dots, \omega_j^{\text{new}}, \dots, \omega_p)'$ . Uniform mutation – Draw independent random values  $u_1, \dots, u_p$  from the Uniform distribution defined between 0 and 1. Then, for any  $k = 1, \dots, p$ , if  $u_k \leq 0.25$ , reverse the indicator and set  $\omega_k^{\text{new}} = 1 - \omega_k$ , otherwise leave  $\omega_k^{\text{new}} = \omega_k$ .
- (10) Draw the auxiliary lattice  $\mathbf{s}$  from  $p(\cdot | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w})$ . Use the Gibbs sample obtained at the  $(N + 1)$ th iteration.
- (11) Propose to move from  $[(\boldsymbol{\omega}, \mathbf{x}), (\boldsymbol{\omega}^{\text{new}}, \mathbf{s})]$  to  $[(\boldsymbol{\omega}^{\text{new}}, \mathbf{x}), (\boldsymbol{\omega}, \mathbf{s})]$ , with probability  $\min\{1; R_{\Omega}\}$ , where

$$R_{\Omega} = \frac{q(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w}) / Z(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w})}{q(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w}) / Z(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w})} \frac{p(\boldsymbol{\omega}^{\text{new}})}{p(\boldsymbol{\omega})} \times \frac{q(\mathbf{s} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w}) / Z(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w})}{q(\mathbf{s} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w}) / Z(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w})} \frac{g(\boldsymbol{\omega} | \boldsymbol{\omega}^{\text{new}})}{g(\boldsymbol{\omega}^{\text{new}} | \boldsymbol{\omega})}. \quad (10)$$

The acceptance ratio (10) becomes

$$R_{\Omega} = \frac{q(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w})}{q(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w})} \frac{q(\mathbf{s} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{w})}{q(\mathbf{s} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\omega}^{\text{new}}, \mathbf{w})},$$

after cancelling out the normalizing constants, the proposal ratio of  $\boldsymbol{\omega}$  because of the symmetry of the proposal density, and the prior ratio of  $\boldsymbol{\omega}$  because the probability of inclusion of each covariate is assumed a priori equal to its probability of exclusion. If the exchange move is accepted, set  $\boldsymbol{\omega} = \boldsymbol{\omega}^{\text{new}}$ .

- (12) Apply one of the randomly selected mutation operators as in step [9] and propose a new indicator  $\boldsymbol{\gamma}^{\text{new}}$ .
- (13) Propose to move from  $\boldsymbol{\gamma}$  to  $\boldsymbol{\gamma}^{\text{new}}$ , with probability  $\min\{1; R_{\Gamma}\}$ , where

$$R_{\Gamma} = \frac{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}^{\text{new}}, \boldsymbol{\rho})}{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho})} \frac{p(\boldsymbol{\gamma}^{\text{new}})}{p(\boldsymbol{\gamma})} \frac{g(\boldsymbol{\gamma} | \boldsymbol{\gamma}^{\text{new}})}{g(\boldsymbol{\gamma}^{\text{new}} | \boldsymbol{\gamma})} = \frac{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}^{\text{new}}, \boldsymbol{\rho})}{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho})},$$

after cancelling out the proposal ratio of  $\boldsymbol{\gamma}$  because of the symmetry of the proposal density, and the prior ratio of  $\boldsymbol{\gamma}$  because the probability of inclusion of each covariate is assumed a priori equal to its probability of exclusion. If the move is accepted, set  $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{\text{new}}$ .

- (14) Apply one of the randomly selected mutation operators as in step [9] and propose a new indicator  $\boldsymbol{\rho}^{\text{new}}$ .

Figura 4 - Algoritmo geração parâmetros desconhecidos.



(15) Propose to move from  $\rho$  to  $\rho^{\text{new}}$ , with probability  $\min\{1; R_p\}$ , where

$$R_p = \frac{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho^{\text{new}})}{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)} \frac{p(\rho^{\text{new}})}{p(\rho)} \frac{g(\rho | \rho^{\text{new}})}{g(\rho^{\text{new}} | \rho)} = \frac{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho^{\text{new}})}{p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)},$$

after cancelling out the proposal ratio of  $\rho$  because of the symmetry of the proposal density, and the prior ratio of  $\rho$  because the probability of inclusion of each covariate is assumed a priori equal to its probability of exclusion. If the move is accepted, set  $\rho = \rho^{\text{new}}$ .

(16) Repeat from step [3], until the MCMC sample is large enough.

(17) When the MCMC sample is large enough, compute the marginal posterior median of the entries of  $\omega$ ,  $\gamma$ , and  $\rho$ , or the ergodic mean of  $\theta$ ,  $\lambda$ ,  $\alpha$ , and  $\beta$ , and the marginal posterior mode of  $\mathbf{x}$ , with  $\omega$ ,  $\gamma$ ,  $\rho$  skipping steps [9–15]).

Figura 5 - Algoritmo geração parâmetros desconhecidos.

O algoritmo apresentado em cima, é iterado até ser possível obter uma amostra grande o suficiente de modo a poder tirar conclusões sobre os parâmetros desconhecidos do modelo.

## 4 Aplicação e Resultados

A distribuição espacial do Francolim-de-asa-cinzenta no Sul de África foi analisada aplicando os algoritmos MOC e MCMC discutidos anteriormente, ao mapa da primeira figura. Nesse mapa, estão representadas 1949 áreas, sendo que 483 dessas estão assinaladas a negro, cerca de 24%, pois foi avistado pelo menos um elemento da espécie nessa secção. A análise de Bayesian ao problema foi desenvolvida em três passos:

- Seleção das variáveis que afetassem o modelo oculto e a detecção da probabilidade de falsas e verdadeiras detecções da espécie;
- Estimativa de parâmetros e a restauração do mapa verdadeiro;
- Comparação entre diferentes modelos.

Para o primeiro passo, o MOC efetuou um milhão de iterações, sendo que no segundo passo correu durante cem mil iterações. Os algoritmos foram testados com dados 'sintético', tendo os resultados obtidos com esses dados sido satisfatórios, sendo assim aceites como algoritmos com capacidade para responder ao problema em questão.

Foram seleccionadas para a primeira fase três variáveis, evapotranspiração anual, o acumulado de dias acima dos 10 graus entre janeiro e junho, o

máximo das temperaturas máximas registadas nas diferentes áreas. Para o segundo passo foram também utilizadas variáveis que representavam o tipo de pastagem e de matagal.

Após a execução dos algoritmos, foi notado o facto de que os Francolim-de-asa-cinzenta preferem prados alpinos juntamente com zonas que evitem alto 'stress térmico'.

A restauração do mapa verdadeiro da distribuição espacial da espécie, figura 3, mostrou que 115(5.8%) áreas que no mapa original possuíam cor branca passaram a ser preenchido por um bloco preto, e que nenhuma área preenchida a preto foi alterada para branco.

Assim sendo, este artigo permitiu uma maior compreensão da dispersão espacial do Francolim-de-asa-cinzenta, tendo por base a junção de um mapa com observações da espécie com uma série de variáveis dos locais observados.



Figura 6 - Mapa restaurado da distribuição espacial Francolim-de-asa-cinzenta.

## Referências

- [1] Spezia L., Friel N. and Gimona A. (2017). Spatial hidden Markov models and species distributions, 1-22. doi: <https://doi.org/10.1080/02664763.2017.1386771>