
O uso de regressão logística para espacialização de probabilidades

EDUARDO M. VENTICINQUE^{1*}
JULIANA STROPP CARNEIRO²
MARCELO PAUSTEIN MOREIRA²
LEANDRO FERREIRA³

¹ Wildlife Conservation Society, Programa de Conservação Andes-Amazônia, Amazonas, Brasil.

² Instituto Nacional de Pesquisas da Amazônia – INPA, Amazonas, Brasil.

³ Museu Paraense Emílio Goeldi, Pará, Brasil.

* e-mail: eventicinque@wcs.org

RESUMO

Neste trabalho discute-se o uso de modelos de regressão logística em análises espaciais, fazendo uma breve introdução sobre regressões logísticas e usando estudos de casos da aplicação desta técnica em estudos ecológicos, utilizando aplicativos de Sistemas de Informação Geográfica.

ABSTRACT

In this chapter we discussed the use of logistic regression models in spatial analyses, doing a brief introduction on logistic regression and your application in some study cases related to ecology studies using with tools the Geographic Information System.

INTRODUÇÃO

A regressão logística vem sendo utilizada nas mais diversas áreas da ciência. Este método, assim como as regressões lineares e múltiplas, estuda a relação entre uma variável resposta e uma ou mais variáveis independentes. A diferença entre estas técnicas de regressão se deve ao fato de que na regressão logística as variáveis dependentes estão dispostas em categorias, enquanto na regressão linear estas variáveis são dados contínuos ou discretos. Outra diferença é que na regressão logística a resposta é expressa por meio de uma probabilidade de ocorrência, enquanto que na regressão simples obtém-se um valor numérico (Penha, 2002).

A estrutura do modelo logístico é apropriada para analisar o comportamento de uma variável dependente categórica. Geralmente, a regressão logística é realizada para dados binários (Cox, 1970), entretanto, também pode ser aplicada a dados multinominais. Tipicamente, a variável dependente é binária e codificada como 0 (ausência) ou 1 (presença); porém, pode ser multinomial, sendo codificada como um número inteiro, variando de 1 a $k - 1$, onde k é um número positivo qualquer. Embora a regressão logística possa ser aplicada a qualquer variável dependente categórica, ela é utilizada com maior frequência em análises de dados binários. Estes exemplos incluem a estimativa de probabilidade de ocorrência de uma espécie em

função de variações na altitude ou da quantidade de chuva, a estimativa da probabilidade de que uma área seja desflorestada em função de sua distância das estradas, rios ou sedes municipais, etc.

De forma sucinta, podemos dizer que existem três procedimentos distintos para manipular dados binários, ordinais e nominais em regressão logística. A escolha de qual método utilizar depende do número de categorias e das características da variável resposta, conforme mostra a Tabela 1.

TABELA 1 – Tipos de variável resposta.

TIPO	NÚMERO DE CATEGORIAS	CARACTERÍSTICAS
Binária	2	Dois níveis
Ordinal*	3 ou +	Ordenação natural de níveis
Nominal*	3 ou +	Sem ordenação natural de níveis

(*) São variações do estado multinominal ou politômico de uma variável (adaptado de Penha, 2002).

Uma variável binária é aquela que aceita apenas dois níveis de resposta, como sim ou não. Já uma variável ordinal segue uma ordenação natural dos fenômenos ou eventos, como pequeno, médio e grande, ou classificações como ruim, regular, bom, ou excelente (“ranks”). A nominal, por sua vez, pode ter mais de três níveis e não considera nenhuma ordenação. Um exemplo seria a classificação de algum objeto em azul, preto, amarelo e vermelho; ou a previsão do tempo como ensolarado, nublado e chuvoso (Penha, 2002).

Existem vários tipos de estudos que se pode analisar com modelos logísticos. Estes incluem bioensaios, epidemiologia, experimentos clínicos, pesquisa de mercado, distribuição de espécies, etc. Neste trabalho nós vamos nos ater às aplicações da regressão logística dentro de um Sistema de Informação Geográfica (SIG) com questões ligadas à ecologia e à conservação.

A Figura 1 compara o modelo linear com o logístico binário básico, utilizando os mesmos dados. Note que o modelo linear prediz valores de y contínuos infinitamente. Assim, se a predição é para compreensão das probabilidades, este modelo é claramente impróprio. Além disso, o modelo linear não se ajusta à média de x para qualquer um dos valores da resposta. Geralmente não se consegue ajustar estes dados satisfatoriamente. Assim, podemos dizer que o modelo linear não é apropriado para esta estrutura de dados. Já o modelo logístico é projetado para ajustar dados binários, quando é

assumido que y representa uma distribuição de probabilidades, ou quando é simplesmente expresso como uma medida binária que estamos tentando prever. Apesar da diferença entre os dois gráficos, o modelo linear e o logístico são variantes um ao outro. Assumindo a variável preditora (x), o modelo linear é:

$$y = xb + e,$$

onde y é um vetor de observações, x é uma matriz dos preditores, e e é um vetor de erros.

Enquanto que o modelo logístico é:

$$y = \exp(xb + e) / [1 + \exp(xb + e)],$$

onde y é a probabilidade de ocorrência de um evento, x é uma matriz dos preditores, e e é um vetor de erros.

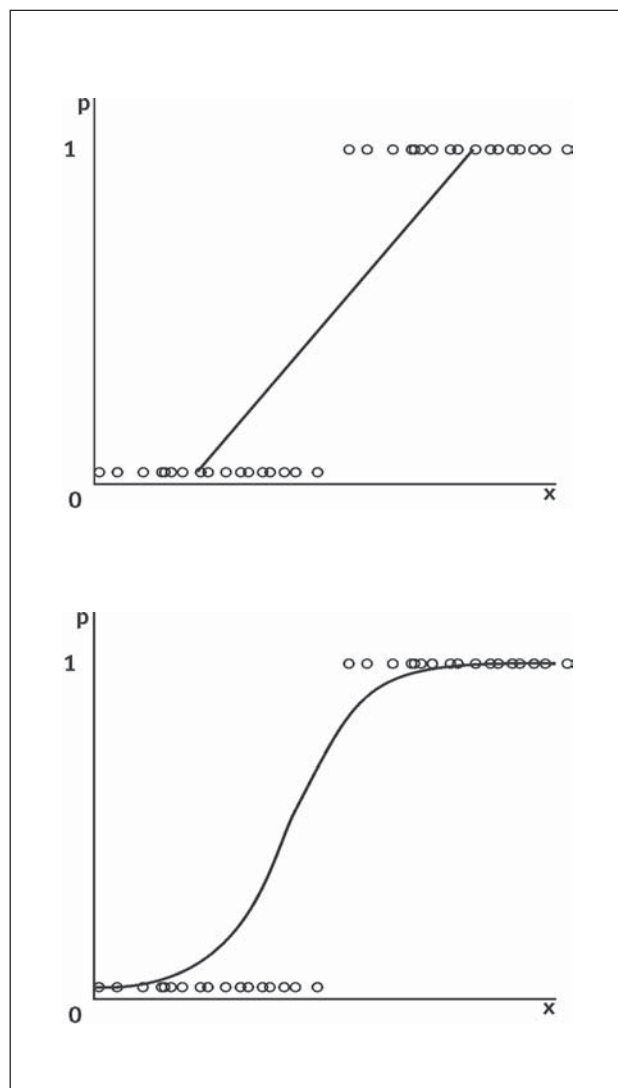


FIGURA 1 – Modelos de regressão linear e logística com dados binários.

Coeficientes e constantes

Podemos avaliar os coeficientes obtidos pela regressão logística de forma parecida com a que fazemos em uma regressão linear. No entanto, sua interpretação é diferente. O coeficiente da regressão logística indica o quanto aumenta a probabilidade de ocorrência de um evento para o aumento de uma unidade na variável independente. O coeficiente pode ser positivo ou negativo. No caso de um coeficiente positivo, quanto maior for seu valor, maior será o poder preditivo da variável independente sobre a probabilidade de ocorrência de um evento. No entanto, a probabilidade de 0 a 1 é resultado de uma função não linear da probabilidade de ocorrência de um evento.

É muito importante lembrar o que quer dizer, em termos de interpretação, uma função não linear. Na regressão linear o acréscimo (ou decréscimo) do valor de y em função do acréscimo de x é constante ao longo de toda escala de valores de x . Já na regressão logística isto não acontece, havendo áreas onde essa mudança é mais pronunciada e outras onde ela nem ocorre. As áreas onde pequenas variações nos valores de x causam grandes mudanças nos valores de y representam áreas de maior probabilidade de mudança de estado da variável y em função de x .

Na Figura 2 podemos visualizar o efeito da variação dos valores da constante e dos coeficientes sobre a curva de probabilidade estimada a partir de regressão logística. O gráfico da Figura 2a foi obtido somente com a troca dos valores da constante (intercepto) e podemos notar que as formas das curvas são exatamente as mesmas e a única mudança é sua localização no eixo x . Em outras palavras, todo modelo de regressão logística tem seus limites entre 0 e 1, só que muitas vezes estes limites estão fora do nosso intervalo de amostragem ou mesmo não são plausíveis de acontecer, por exemplo, como valores negativos de distância. Neste caso, não é possível visualizar em que intervalo de x as probabilidades alcançam valores próximos de 1. Já na Figura 2b temos uma situação distinta, onde a alteração dos coeficientes com uma constante fixa causa mudanças evidentes na distribuição da probabilidade de ocorrer um evento em função da mudança de valores no eixo x . Podemos notar que quanto maior o coeficiente, maior é a mudança na probabilidade estimada em função de mudanças no x . De forma simplificada, podemos dizer que o coeficiente modela a curva enquanto que a constante a localiza em função do x .

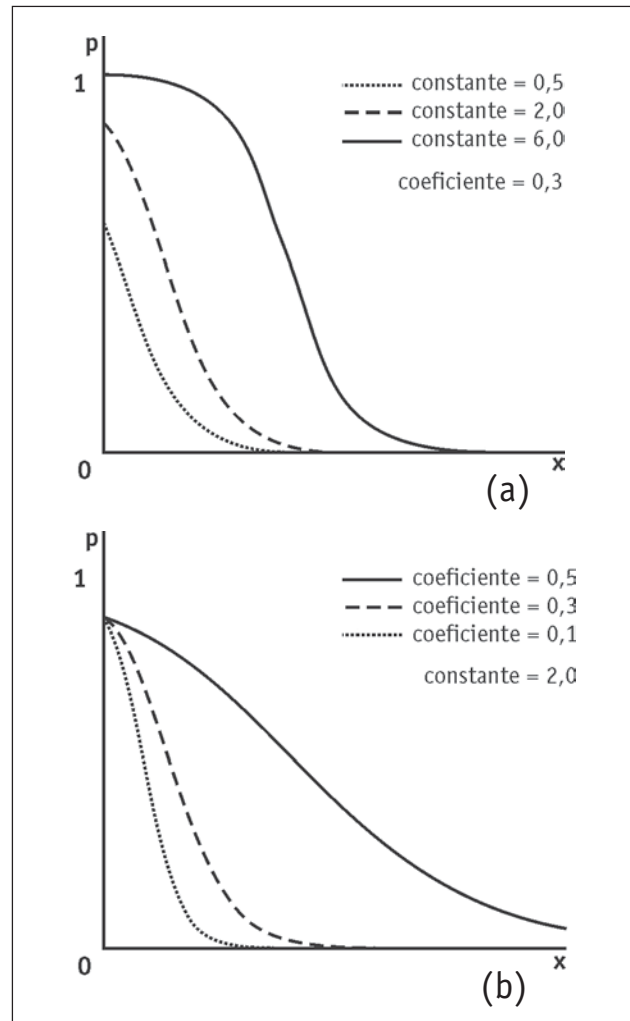


FIGURA 2 – Modelos de regressão logística obtidos com alterações somente na constante (a) e no coeficiente (b).

Razão de chances (*odds ratio*)

A razão de chances permite conhecer quais chances um evento tem de acontecer se, sob as mesmas condições, ele não acontecer. Ou seja, razão de chances é uma medida de associação e expressa a aproximação do quanto é mais provável (ou improvável) para o resultado estar presente entre aqueles com $x = 1$ do que entre aqueles com $x = 0$. Por exemplo, se y denota a presença ou ausência de uma determinada espécie e x denota se a área tem ou não tem floresta, o *Odds* = 2 indica que a presença daquela espécie é duas vezes mais esperada em áreas com floresta do que em áreas sem floresta. Ou seja, a presença de floresta é muito importante para aumentar a chance de ocorrência daquela espécie. Outro exemplo, que talvez possa ser mais intuitivo, seria a razão de chances de ser atropelado toda vez que se atravessa uma avenida. Mesmo que você

atravesse a avenida e não seja atropelado, existia uma chance deste evento ocorrer, essa chance é a “razão de chances” ou “*odds ratio*”. A razão de chances de resposta é dada por $p/(1-p)$ onde p é a probabilidade de resposta, e a razão de chances é o fator multiplicativo de mudança de estado de y quando a variável independente aumenta uma unidade. O livro de Hosmer & Lemeshow (1989) contém maiores explicações sobre a interpretação e forma de cálculo das razões de chance e de seus intervalos de confiança.

Estatística de *Likelihood-Ratio*

Uma vez definido o modelo, é necessário testar a sua validade. Em regressão logística há uma série de gráficos, testes de ajuste, e outras medidas para assegurar a validade do modelo. Estas estatísticas permitem identificar as variáveis que não se ajustam bem, ou que têm forte influência sobre a estimativa dos parâmetros.

Uma das formas mais comuns de se avaliar o modelo como um todo, é por meio da estatística de *Likelihood-Ratio*. Esta estatística testa a hipótese de que todos os coeficientes, menos a constante, são iguais a 0. A significância da estatística de *Likelihood-Ratio* (LR) é testada utilizando a distribuição do X^2 com os graus de liberdade iguais ao número de variáveis independentes no modelo, não incluindo a constante.

O teste de *Likelihood-Ratio*, ou teste G, é calculado utilizando o valor da estatística de log likelihood do modelo saturado e do insaturado. Tipicamente, o modelo saturado contém o conjunto de variáveis analisadas e o modelo insaturado omite um subconjunto selecionado, embora outras restrições sejam possíveis. A estatística do teste é duas vezes a diferença da *Likelihood-Ratio* do modelo saturado para o insaturado e é testada com a distribuição do X^2 , sendo o grau de liberdade igual ao número de restrições impostas. Se um modelo contém uma constante, podemos calcular um teste de *Likelihood-Ratio* da hipótese nula em que todos os coeficientes, exceto a constante, são iguais a 0. A fórmula da estatística G usada para testar o modelo é a seguinte:

$$G = 2*[LL(N)-LL(0)]$$

Onde:

LL(N) = log *likelihood* do modelo saturado
(todas as variáveis inclusas)

LL(0) = log *likelihood* do modelo insaturado
(somente a constante inclusa)

Quando fazemos essa subtração, estamos olhando, simplesmente, o quanto as variáveis estão causando mudanças nas probabilidades de ocorrência de um evento e se essas mudanças são maiores que esperadas ao acaso.

Para ilustrar o uso do teste de *Likelihood-Ratio*, considere o seguinte modelo:

Presença de uma espécie =
CONSTANTE + altitude + chuva + temperatura (saturado)
Presença de uma espécie =
CONSTANTE + altitude + chuva (insaturado)

Podemos formular a hipótese nula de que a temperatura não contribui para explicar a variação do modelo e proceder ao teste desta forma. Suponha que para este exemplo os valores de G são 12,05 e 5,01, com 3 e 2 graus de liberdade para os modelos saturado e insaturado, respectivamente. Agora podemos entender a variação que é explicada pela temperatura, entendendo quanto perdemos de poder de explicação ao removermos essa variável do modelo. Isso pode ser realizado da seguinte forma:

Efeito da temperatura = G (insaturado) – G (saturado),
com 3 – 2 graus de liberdade.

Essa expressão fica assim:

G = 12,05 – 5,01, com 1 grau de liberdade
G = 7,04, gl = 1 e $p < 0,05$, rejeitando-se a hipótese nula de que a temperatura não tem influência sobre a probabilidade de ocorrência de uma determinada espécie.

TESTES ESTATÍSTICOS DE AJUSTE DO MODELO ÀS OBSERVAÇÕES

Rho² de McFadden

Rho² de McFadden é uma transformação da estatística de LR para imitar um R² da regressão linear. Seus valores estão sempre entre 0 e 1 e, quanto mais alto, melhor é o ajuste do modelo aos resultados. Entretanto, o Rho² de McFadden tende a ser muito mais baixo que R². Porém, baixos valores não implicam, necessariamente, num ajuste pobre. Valores entre 0,2 e 0,4 são considerados satisfatórios (Hensher & Johnson, 1981).

Pearson

Mede quão bem a observação é prevista pelo modelo. Observações que não se ajustam bem ao modelo têm um alto valor de Pearson.

Hosmer-Lemeshow

Este teste avalia o modelo ajustado, comparando as frequências observadas e as esperadas. O teste associa os dados às suas probabilidades estimadas, da mais baixa à mais alta, e então faz um teste qui-quadrado para determinar se as frequências estimadas estão próximas das frequências observadas (Hosmer & Lemeshow, 1989).

Diagnósticos de regressão

Na regressão logística, a representação gráfica permite visualizar vários testes de ajuste, sendo que há gráficos relacionados à probabilidade do evento e outros relacionados à alavancagem (que diz se uma observação é um ponto extremo e possui uma forte influência na determinação da reta de regressão, o que diminui a capacidade de análise do modelo). A inspeção gráfica é realizada com base nos pontos extremos de influência (ou *outliers*). Em alguns casos, o ponto que foi identificado como extremo deve ser excluído da amostra e, em seguida, deve ser novamente calculada a equação e o gráfico. Quando os coeficientes desta nova equação forem muito diferentes dos coeficientes da antiga, significa que aquele era um ponto de influência. Se o contrário ocorrer, significa que aquele ponto era apenas um ponto extremo. A decisão de se remover dados da amostra deve ser procedida com muito critério e cuidado. Geralmente, existem informações importantes nestes pontos discrepantes. Por exemplo, os gráficos do delta qui-quadrado (DELPSTAT) *versus* probabilidade do evento identificam os pontos que não se ajustam bem aos modelos.

Em diversos pacotes estatísticos pode-se criar um arquivo para elaborar diagnósticos da regressão logística (Pregibon, 1981; Cook & Weisberg, 1984; Steinberg & Colla, 1998). No caso do programa SYSTAT, o arquivo contém as variáveis apresentadas na Tabela 2.

Podemos entender a variável: LEVERAGE (1) como uma medida da influência de uma observação no ajuste do modelo, e a variável DELBETA (1) como uma medida da mudança no vetor do coeficiente devido àquela observação. Por exemplo, os gráficos de PEARSON, DEVIANCE, LEVERAGE (1), DELPSTAT, com o CASO, destacam pontos de dados diferenciados. Para discussão adicional e interpretação de gráficos de diagnóstico, veja o Capítulo 5 de Hosmer & Lemeshow (1989).

TABELA 2 – Variáveis de diagnóstico da regressão logística geradas pelo programa estatístico SYSTAT. Uma descrição detalhada destas variáveis pode ser encontrada no manual do SYSTAT ou em Hosmer & Lemeshow (1989).

NOME NO ARQUIVO	LEGENDA
ACTUAL	Valor da variável dependente
PREDIGA	Valor predito (1 ou 0)
PROB	Probabilidade predita
LEVERAGE (1)	Elemento diagonal da matriz "chapéu" de Pregibon
LEVERAGE (2)	Componente de LEVERAGE (1)
PEARSON	Resíduo de observação de Pearson
VARIANCE	Variância de resíduo de Pearson
PADRÃO	Resíduo de Pearson padronizado
DEVIANCE	Desvios Residuais
DELPSTAT	Mudança no χ^2 de Pearson
DELBETA (1)	Mudança padronizada em Beta
DELBETA (2)	Mudança padronizada em Beta
DELBETA (3)	Mudança padronizada em Beta

Principais problemas e vantagens

PROBLEMAS

- Se o fenômeno de interesse não for monotônico e seu pico de frequência tiver valores intermediários, será difícil obter um bom ajuste do modelo. Esse tipo de problema pode ser identificado através da análise dos resíduos da regressão;
- Pontos discrepantes, principalmente nos limites da distribuição das variáveis explanatórias, podem causar resultados espúrios;
- Obter dados confiáveis e não viciados para alimentação do modelo;
- Autocorrelação espacial.

VANTAGENS

- O modelo logístico requer informações simples e, portanto pode ser alimentado com facilidade;
- Trabalha com N variáveis simultaneamente;
- Trabalha simultaneamente com N vetores direcionais de variação. Essa flexibilidade pode ser obtida quando construímos em um Sistema de Informação Geográfica uma camada de dados independente. No caso de se trabalhar com informações com formas irregulares, por exemplo, distância da estrada ou declividade, o sentido de variação espacial do fenômeno pode ocorrer em diferentes direções;
- As probabilidades obtidas podem ser espacializadas e então se aplicar um filtro para que os padrões possam ser mais facilmente visualizados;
- Facilidade de interpretação e uso dos resultados em probabilidades.

Validação do modelo

Assim como a maioria dos métodos, a regressão logística necessita de novos dados (nova amostra) ou de uma amostra reservada dos dados para verificar se o mesmo modelo pode ser satisfatoriamente ajustado a estes novos dados. Ou seja, é preciso determinar se os coeficientes e os erros-padrão obtidos a partir dos dados utilizados para elaboração do modelo são similares aos obtidos para os dados de validação.

O uso de regressão logística na análise espacial

Fizemos uma consulta bibliográfica (www.webofscience.com) a partir das palavras-chave “regressão logística” e “Sistemas de Informação Geográfica” e obtivemos 93 trabalhos publicados ao longo de 58 anos. O uso das regressões logísticas associado a Sistemas de Informação Geográfica começou a ser mais praticado a partir de 1990 (Figura 3). De 1999 até 2003 foram publicados, em média, 14 trabalhos por ano.

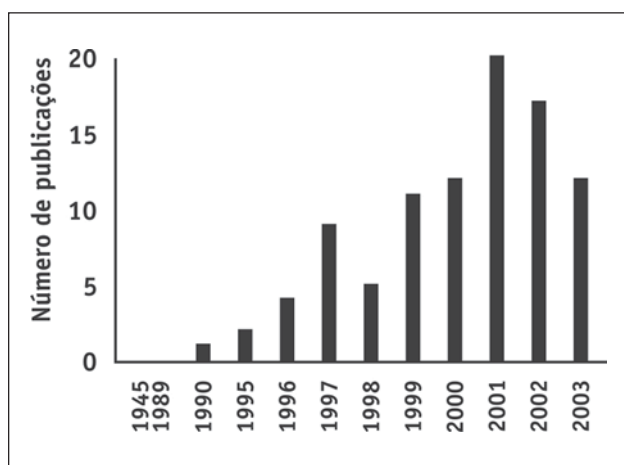


FIGURA 3 – Número de publicações encontradas no site www.webofscience.com, referente ao emprego de regressão logística em Sistemas de Informação Geográfica (SIG).

EXEMPLOS DE APLICAÇÃO DA REGRESSÃO LOGÍSTICA

Aqui vamos citar três exemplos onde técnicas de regressão logística simples são utilizadas. O leitor poderá reparar que o uso pode ter variação na escala espacial empregada, podendo ser usada desde a detecção de respostas de Odonata à proporção de florestas circundando igarapés na Amazônia central até modelos de desflorestamento em função da malha viária na Amazônia Legal. Outra característica relevante diz respeito às diferentes formas com que estes

modelos podem ser usados. Há casos, como no exemplo da probabilidade de ocorrência de espécies de Odonata em função da quantidade de floresta, onde os resultados não são reprojatados, ou seja, não há espacialização das probabilidades, pois estávamos interessados simplesmente em saber se há influência da quantidade de floresta preservada ao longo dos pequenos cursos d’água sobre a ocorrência de determinadas espécies.

No estudo realizado com distribuição de árvores na Mil Madeireira Itacoatiara Ltda, as probabilidades de ocorrência de cada espécie foram projetadas no espaço, utilizando como base os mapas de altitude e declividade (variáveis independentes). Neste caso, foi utilizada uma regressão logística múltipla e o modelo pode ser considerado espacialmente explícito, pois podemos localizar, no espaço, todas as probabilidades.

Outro exemplo em que o espaço continua implícito, mas as probabilidades não são projetadas no mapa, é o trabalho sobre a importância das unidades de conservação e terras indígenas, ajudando a conter o desmatamento na Amazônia brasileira. Nesse caso, a regressão logística foi utilizada para entender a probabilidade de uma área florestada ser convertida em área desflorestada, considerando se está localizada dentro ou fora de uma terra indígena ou unidade de conservação, e a distância que está da malha viária.

O que tentamos aqui, por meio destes exemplos, é oferecer ao leitor um panorama geral de alguns usos que podemos ter com regressão logística, e também chamar atenção para o uso da técnica com problemas espaciais.

Mudanças na fauna de odonatas em igarapés amazônicos, em função de alterações na cobertura vegetal

Este estudo, realizado na Amazônia central, estima a probabilidade de ocorrência de espécies de odonatas em função da quantidade de floresta. Neste estudo, o pesquisador Dr. Paulo de Marco estava interessado em saber se existe influência da quantidade de floresta preservada ao longo dos pequenos cursos de água sobre a ocorrência de determinadas espécies de odonatas. O resultado ajuda a avaliar se a largura da mata ripária prevista no Código Florestal Brasileiro é suficiente para manter o conjunto de espécies de odonatas.

Probabilidade de ocorrência de uma espécie = $\exp (\% \text{ de cobertura florestal} * \text{coeficiente} + \text{Intercepto} + \text{erro}) / [1 + \exp (\% \text{ de cobertura florestal} * \text{coeficiente} + \text{Intercepto} + \text{erro})]$

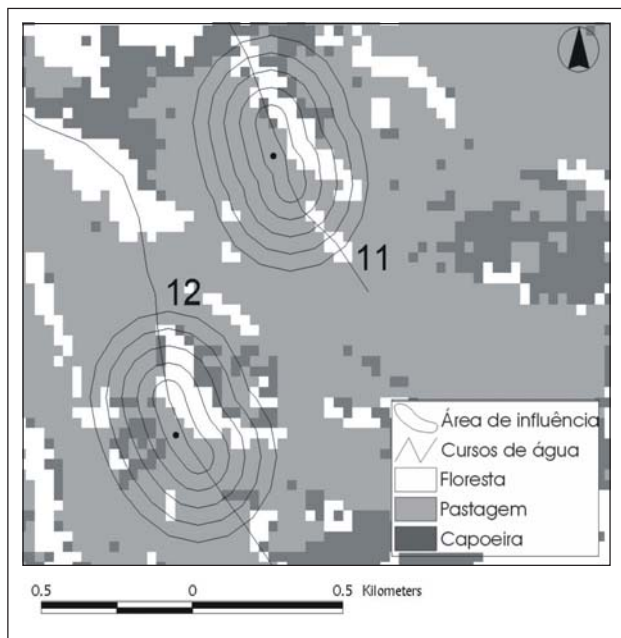


FIGURA 4 – Exemplo de como a paisagem é tratada neste estudo. A análise é realizada utilizando o valor de cobertura florestal contida dentro das áreas de influência. No caso deste estudo, os anéis são distanciados 50 metros.

Podemos notar nos resultados (Tabela 3) que somente uma espécie respondeu às alterações na cobertura de forma negativa, ou seja, quanto maior era a porcentagem de cobertura florestal menor era a probabilidade de encontrar a espécie. As demais espécies não responderam à proporção de mata ao redor dos pontos amostrais, na área do estudo.

Uso de regressão logística para modelar a distribuição espacial de espécies arbóreas na Amazônia central

O presente trabalho é parte dos resultados apresentados na dissertação de mestrado de Juliana Stropp Carneiro, sob a orientação do Dr. Eduardo Venticinque

(Carneiro, 2004). Este trabalho teve como objetivo elaborar modelos preditivos de ocorrência de *Aniba roseodora* (pau-rosa), *Cariniana micrantha*, *Caryocar villosum*, *Dinizia excelsa*, *Dipteryx odorata*, *Goupia glabra*, *Manilkara bidentata*, *Manilkara huberi*, *Parkia multijuga*, *Parkia pendula*, *Peltogyne paniculata* e *Pseudopiptadenia psilostachya* em função da topografia. Estimamos a probabilidade de ocorrência dos indivíduos com regressão logística múltipla, sendo a variável dicotômica a presença e a ausência dos indivíduos, e as variáveis contínuas a altitude e a declividade do terreno. As informações sobre a ocorrência das árvores foram cedidas pela Mil Madeireira Itacoatiara Ltda. Os dados sobre a ocorrência dos indivíduos arbóreos foram coletados pela empresa durante a prospecção e o mapeamento das árvores com DAP ≥ 40 cm. As informações sobre a ocorrência das árvores consistem em um arquivo do tipo pontos, em formato *shapefile*, com a lista de espécies e as coordenadas da localização dos indivíduos em UTM. Convertimos esse arquivo para o formato matricial e obtivemos um arquivo do tipo GRID, com células de 93 m. Elaboramos o Modelo Digital do Terreno a partir dos dados do Shuttle Radar Topography Mission (SRTM) e adquirimos os dados sobre altitude no site <http://seamless.usg.gov>. Para a correção geométrica, utilizamos como base uma imagem Landsat TM 7 (órbita/ponto 230/62) georreferenciada (projeção UTM – zona 21; datum WGS 84). Posteriormente, coregistramos a imagem SRTM com a base dos igarapés da área de interesse digitalizada. Para obter os parâmetros da regressão logística, exportamos os dados do ArcView 3.2 e os analisamos em um pacote estatístico. Posteriormente, aplicamos as equações obtidas nos modelos logísticos aos temas de altitude e declividade e obtivemos os mapas de probabilidade de ocorrência de indivíduos (ver anexo). Deste modo, estes mapas expressam a probabilidade de ocorrência dos indivíduos em células de 93 m, dada a altitude e declividade daquela célula.

TABELA 3 – Análise de regressão logística para a dependência da presença de algumas espécies de Odonata em relação à proporção de mata ao redor dos pontos amostrais, na área do Projeto Dinâmica Biológica de Fragmentos Florestais (PDBFF), Manaus, AM. Valores entre parênteses são os erros padrões dos parâmetros estimados.

ESPÉCIE	COEFICIENTE BO	% DE MATA	X ² (VALOR DE p)
<i>Argia</i> sp.1	-1,086 (1,138)	2,305 (1,631)	2,145 (0,143)
<i>Argia</i> sp. 2	2,665 (1,461)	-3,233 (1,861)	3,656 (0,050)
<i>Chalcopteryx scintilans</i>	-0,782 (1,114)	1,490 (1,549)	0,954 (0,329)
<i>Dicterias atosanguinea</i>	-0,080 (1,094)	-0,181 (1,504)	0,014 (0,904)

Para determinar a capacidade preditiva do modelo, obtivemos a tabela de sucesso de predição para cada um dos modelos gerados. Esta tabela é composta pelas variáveis expressas abaixo:

$$resposta = \sum P_i,$$

onde P_i é a probabilidade estimada para as células de presença;

$$referência = \sum P_j,$$

onde P_j é a probabilidade estimada para as células de ausência;

$$\text{Índice de acerto de presença} = \frac{resposta}{N_i},$$

onde N_i é o número de células de presença;

$$\text{Índice de acerto de ausência} = \frac{referência}{N_j},$$

onde N_j é o número de células de ausência.

As variáveis Índice de acerto de presença, Índice de acerto de ausência e Índice de acerto total refletem a relação entre a distribuição observada e a esperada, indicando o nível de acerto do modelo.

Ainda com o objetivo de determinar se as probabilidades geradas pelos modelos refletem aumento no acerto de ocorrência de um indivíduo, comparamos a probabilidade de acerto usando o modelo com a probabilidade de acerto ao acaso.

A análise de regressão logística indicou associação entre a ocorrência dos indivíduos e a topografia para 10 espécies. As espécies *D. excelsa*, *A. rosaeodora* e *C. villosum* tiveram o padrão de distribuição distinto das demais, em relação à topografia. *D. excelsa* teve relação positiva tanto com a declividade quanto com a altitude, sugerindo que a probabilidade de encontrar indivíduos dessa espécie é maior em lugares altos e íngremes, ou seja, no início dos platôs. Já a ocorrência de *C. villosum* teve relação positiva com a declividade e negativa com a altitude, indicando que esta espécie ocorre nas vertentes e em baixas altitudes. *A. rosaeodora* mostrou-se negativamente relacionada com as variáveis topográficas analisadas, ocorrendo em locais de altitude e declividade baixas, estando assim associada às regiões de baixo. Entretanto as espécies *C. micrantha*, *G. glabra*, *M. huberi*, *M. bidentata*, *P. multijuga*, *P. pendula* e

P. psilostachya evidenciaram semelhanças na maneira em que se distribuem ao longo da toposequência. Essas espécies mostraram-se associadas a locais de altitude elevada e baixa declividade, características que definem os ambientes de platô. As espécies *D. odorata* e *P. paniculata* não tiveram a distribuição estruturada pela topografia.

Os mapas de probabilidade de ocorrência das espécies estudadas representam a configuração espacial da distribuição prevista para cada espécie. Nos mapas das espécies que têm sua ocorrência influenciada pela topografia, é possível visualizar concordância entre as probabilidades mapeadas e a variável topográfica de maior influência sobre a distribuição de seus indivíduos. A Figura 5 ilustra o exemplo do mapa de probabilidade de ocorrência de *P. multijuga*. Neste caso, o índice de acerto de presença foi maior que a probabilidade de encontrar indivíduos ao acaso, indicando que o modelo é capaz de prever a distribuição desta espécie na área onde foi elaborado.

O modelo preditivo foi capaz de prever corretamente a ocorrência de *A. rosaeodora*, *C. micrantha*, *C. villosum*, *D. excelsa*, *M. huberi*, *M. bidentata*, *P. multijuga*, *P. pendula* e *P. psilostachya* na área onde foi elaborado. Esse resultado indica que, em determinados compartimentos da paisagem, a topografia pode condicionar a distribuição de algumas espécies. Em geral, as características edáficas das florestas de terra firme da Amazônia central se alteram ao longo do gradiente de altitude. Dessa forma, a topografia é uma medida indireta das mudanças do ambiente na paisagem (Tuomisto & Ruokolainen, 1994) e, portanto, as diferentes respostas das espécies frente às posições topográficas refletem a influência que a variação ambiental pode ter sobre a estrutura espacial dessas espécies. No caso do trabalho aqui apresentado, a liberação dos dados SRTM na Internet foi fundamental para alcançar os objetivos propostos. Vale lembrar que já existem imagens do SRTM com resolução de 30 metros disponíveis para os Estados Unidos da América e, em breve, é provável que tenhamos acesso a esta informação para a região Amazônica. Caso isso ocorra, poderemos fazer modelos ainda mais precisos, baseados na topografia. Outro fato importante é a ausência de custo para se trabalhar com estas imagens. Se conseguirmos gerar bons modelos preditivos sobre a distribuição de espécies arbóreas com dados provenientes do SRTM, passaremos a ter uma ferramenta eficiente e de baixo custo para modelar a distribuição destas espécies.

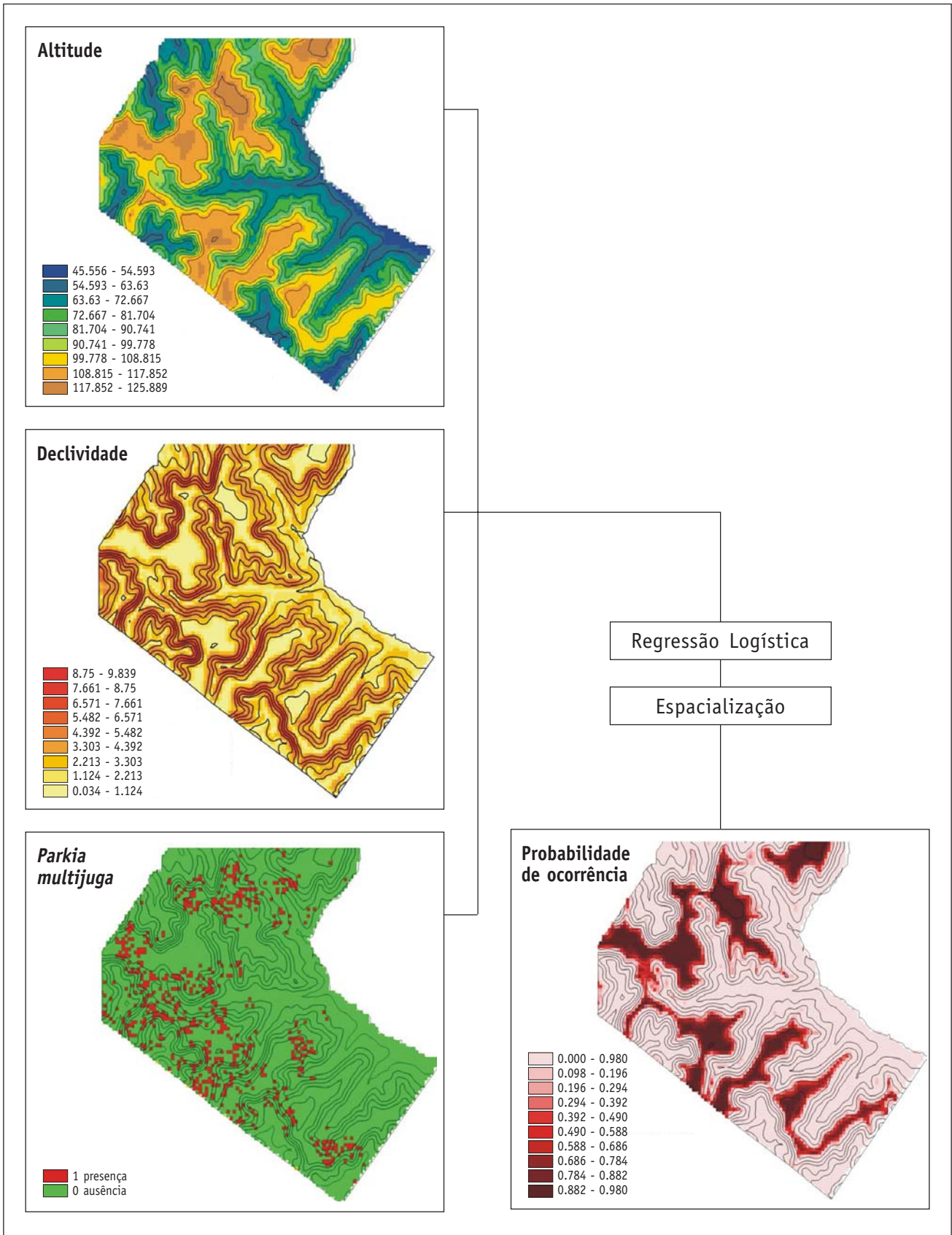


FIGURA 5 – Mapa de probabilidade de ocorrência de *Parkia multijuga*, obtido a partir dos dados de ocorrência dos indivíduos, altitude e declividade, na área da Mil Madeireira Itacoatiara Ltda.

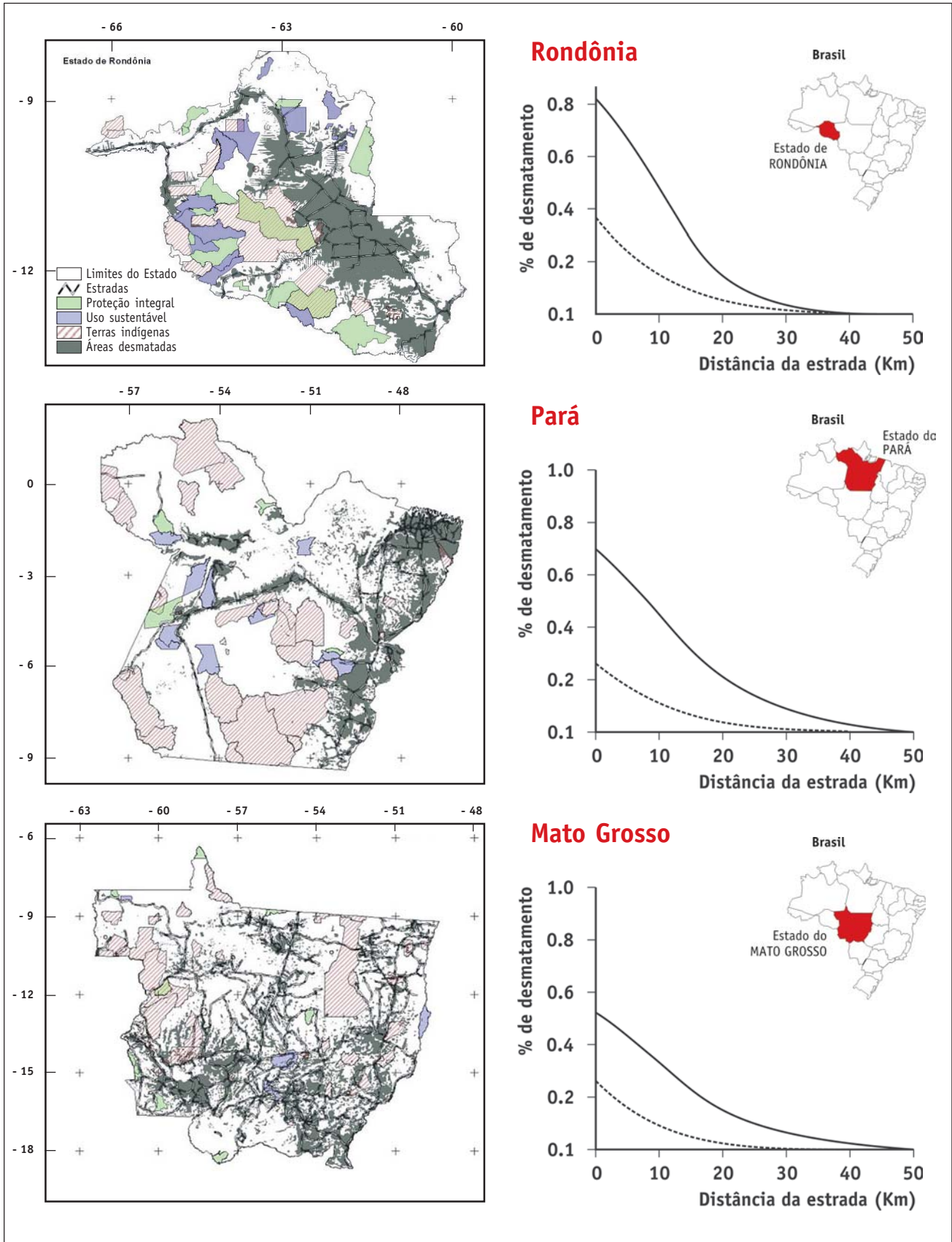


FIGURA 6 – Proporção de área desmatada em função da distância das estradas, dentro (tracejado) e fora (contínua) de áreas protegidas, em Rondônia, no Pará e no Mato Grosso.

O aumento do poder de predição providenciado pelo modelo é influenciado pelos fatores que estruturam espacialmente as espécies. Assim, se o modelo contempla os fatores preponderantes na ocorrência das espécies, o poder de predição é maior. Deste modo, a compreensão dos fatores que interferem na distribuição espacial das espécies e sua incorporação aos modelos preditivos podem providenciar modelos mais próximos da realidade. Portanto, a incorporação de informações da variação ambiental, bem como estudos aprofundados da relação espécie-ambiente (Pitman *et al.*, 2001) e a análise da distribuição das árvores em escala regional, podem contribuir para a modelagem da distribuição espacial das árvores. Dado o contexto em que se insere a análise da configuração espacial da vegetação, a análise dos dados sobre variáveis ambientais relacionadas com a ocorrência de espécies pode ser proveitosa para a compreensão da distribuição da diversidade na Amazônia.

Unidades de conservação e terras indígenas ajudam a conter desmatamento na Amazônia brasileira

O objetivo deste tópico foi testar diferenças no nível de desmatamento dentro e fora de unidades de conservação (proteção integral e uso sustentável) e terras indígenas (denominadas aqui como áreas protegidas) em relação à distância das estradas, nos Estados de Rondônia, Pará e Mato Grosso, para ilustrar a importância de unidades de conservação como redutores do efeito do desmatamento na Amazônia. Estes estados foram escolhidos como estudos de caso devido a sua importância na participação do desmatamento da Amazônia, já que somam cerca de 70% do total da área desmatada nesta região entre 2000-2001 (INPE, 2003). Os Estados de Rondônia, Pará e Mato Grosso têm cerca de 29,2%, 20,4%, e 28,4% de sua área já desmatada, respectivamente.

A análise demonstra que a proporção total da área desmatada fora das áreas protegidas sempre foi significativamente mais elevada do que no interior destas. Uma diferença que pode variar de 9,8 a 19,6 vezes, dependendo do estado analisado. A regressão logística também permite demonstrar que a proporção do desmatamento decai exponencialmente em função da distância das estradas. Contudo, o desmatamento dentro das áreas protegidas é sempre menor do que fora delas nos três estados analisados, mesmo quando estas áreas situam-se próximas a estradas. Isso derruba a crença de que as áreas protegidas sofreriam menos desmatamento somente pelo fato de estarem situadas mais distantes das estradas (Figura 6).

AGRADECIMENTOS

Agradecemos à Mil Madeireira Itacoatiara Ltda. pela disponibilização da base de dados, ao Projeto Dinâmica Biológica de Fragmentos Florestais (PDBFF), ao WWF – Brasil, Projeto Experimento de Grande Escala da Biosfera-Atmosfera da Amazônia (LBA) e ao Projeto Geoma, pelo suporte financeiro (Bolsa de Marcelo Moreira). À Marina Antongiovanni, Amanda Mortati e Ana Albernaz pela revisão do texto. Ao Dr. Paulo De Marco por nos autorizar a usar seus dados em um dos exemplos.

REFERÊNCIAS BIBLIOGRÁFICAS

- Carneiro, J.S. 2004. Mapeamento preditivo da vegetação: uso de SIG para modelar a distribuição espacial de espécies arbóreas na Amazônia central. 2004. Dissertação de Mestrado. Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus.
- Cook, D.R. & S. Weisberg. 1984. Residuals and influence in regression. Chapman and Hall, New York.
- Cox, D.R. 1970. The analysis of binary data. Methuen, Londres.
- Hensher, D. & L.W. Johnson. 1981. Applied discrete choice modelling. Croom Helm, London.
- Hosmer, D.W. & S. Lemeshow. 1989. Applied logistic regression. John Wiley & Sons, Inc., New York.
- INPE. 2003. Disponível em: <http://www.obt.inpe.br/prodes> (acessado em janeiro 2007).
- Penha, R.N. 2002. Um estudo sobre regressão logística binária. Disponível em: <http://www.iem.efei.br/dpr/td/producao2002/PDF/Renata.PDF> (acessado em novembro de 2003).
- Pitman, N.C.A., J. Terborgh, M.R. Silman, P.V. Núñez, D.A. Neill, C.E. Cerón, W.A. Palacios & M. Aulestia. 2001. Dominance and distribution of tree species in upper Amazonian terra firme forests. *Ecology* 82: 2101-2117.
- Pregibon, D. 1981. Logistic regression diagnostics. *Annals of Statistics* 9: 705-724.
- Steinberg, D. & P. Colla. 1998. Logistic regression. In: Wilkinson, L. (ed). SYSTAT 8.0 Statistics. pp. 517-584. Chicago.
- Tuomisto, H. & K. Ruokolainen. 1994. Distribution of Pteridophyta and Melastomataceae along an edaphic gradient in an Amazonian rain forest. *Journal of Vegetation Science* 5: 25-34.

ANEXO

Como espacializar regressão logística no ArcView versão 3.2

A espacialização da regressão logística no ArcView 3.2 se inicia pela determinação dos temas a serem relacionados. Para isso, define-se o tema que contém a variável dependente binária (1/0 – presença e ausência

do fenômeno de interesse) e o tema que representa a variável preditora contínua. No caso da regressão logística múltipla, é possível estabelecer dois ou mais temas referentes às variáveis preditoras. Após essa determinação, é feita a análise estatística dos dados, a fim de se obter os parâmetros da regressão logística. Finalmente, os parâmetros da regressão são incorporados ao ArcView 3.2 e então realizadas as operações para sua espacialização. Os tópicos abaixo descrevem detalhadamente esses três procedimentos:

Obtenção dos temas referentes à variável dicotômica e contínua

Os temas (camadas digitais) deste tópico devem estar em formato GRID e apresentar a mesma resolução espacial, número de linhas e colunas. O arquivo GRID referente ao tema da variável binária deve ter os valores das células 0 e 1. Assim, se a representação do evento for do tipo ponto, linha ou polígono, é necessário converter o arquivo para GRID e atribuir o valor 1 e 0 às células correspondentes à presença e ausência do evento, respectivamente. A conversão para GRID e a associação do valor 1 às células de presença pode ser feita a partir do menu do ArcView 3.2 e a associação do valor 0 pode ser feita a partir da extensão Grid PIG Tolls (<http://arcscripsts.esri.com> ou <http://arcscripsts.esri.com/details.asp?dbid=11872>). Este tema será utilizado nas operações de obtenção dos valores a serem utilizados na análise estatística. Assim, é necessário que a tabela associada a ele tenha quatro campos, como na Figura 7.

Os campos *value* e *count* são criados automaticamente pelo ArcView 3.2 e indicam o valor numérico do pixel (*value*) e o respectivo número de pixels (*count*) com valor 0 e 1. Os campos *presença* e *ausência* são criados pelo usuário, sendo que o campo *presença* apresenta valor 1 para presença e 0 para ausência e o campo *ausência* valor 1 para ausência e 0 para presença. Até aqui definimos o tema referente à variável categórica. O próximo passo é determinar os temas com as variáveis contínuas. Isso é definido pela experiência e pela disponibilidade de dados do usuário.

Obtenção dos dados para a análise estatística

Neste tópico será descrito como obter os valores das variáveis contínuas na área de estudo do evento de interesse. Para isso, realizaremos algumas operações matemáticas na função MAP CALCULATOR do ArcView 3.2, com os temas definidos anteriormente. O esquema das operações entre as camadas é mostrado na Figura 8.

- Obtenção dos valores da variável contínua nas células de presença do evento de interesse:
[GRID variável contínua] ÷ [GRID variável categórica “campo presença = 1”]
- Obtenção dos valores da variável contínua nas células de ausência do evento de interesse:
[GRID variável contínua] ÷ [GRID variável categórica “campo ausência = 1”]

Os GRIDs gerados por essas operações devem ser exportados no formato ASCII Raster (opção disponível no menu do programa). A planilha da primeira operação contém os valores das variáveis contínuas nos pixels referentes à presença e a da segunda, os valores referentes à ausência. O valor “-9999” é atribuído à ausência de dados.

As planilhas podem ser editadas no Excel. Sugerimos a elaboração de uma única planilha com duas colunas: uma contendo a variável contínua e outra a informação de presença e ausência. No caso da regressão logística múltipla, a planilha pode conter três ou mais colunas. A partir dessas planilhas é possível se obter os parâmetros necessários para espacialização da regressão logística em um pacote estatístico.

Espacialização da regressão logística no ArcView 3.2

A equação da regressão logística simples pode ser espacializada no ArcView 3.2 a partir das seguintes operações:

- ([GRID variável contínua]) * coeficiente - > [GRID A]
- ([GRID A] + Constante) - > [GRID B]
- ([GRID B] .Exp) - > [GRID C]
- ([GRID C] + 1) - > [GRID D]
- ([GRID C] / [GRID D]) - > [GRID E]

Já para a equação da regressão logística múltipla, sua espacialização é feita com as operações indicadas abaixo.

- ([GRID variável contínua₁] * (coeficiente₁)) + ([GRID variável contínua₂] * (coeficiente₂)) - > A
- ([GRID A] + Constante) - > [GRID B]
- ([GRID B] .Exp) - > [GRID C]
- ([GRID C] + 1) - > [GRID D]
- ([GRID C] / [GRID D]) - > [GRID E]

Todas essas operações podem ser realizadas a partir da função MAP CALCULATOR do módulo Spatial Analyst do ArcView 3.2.

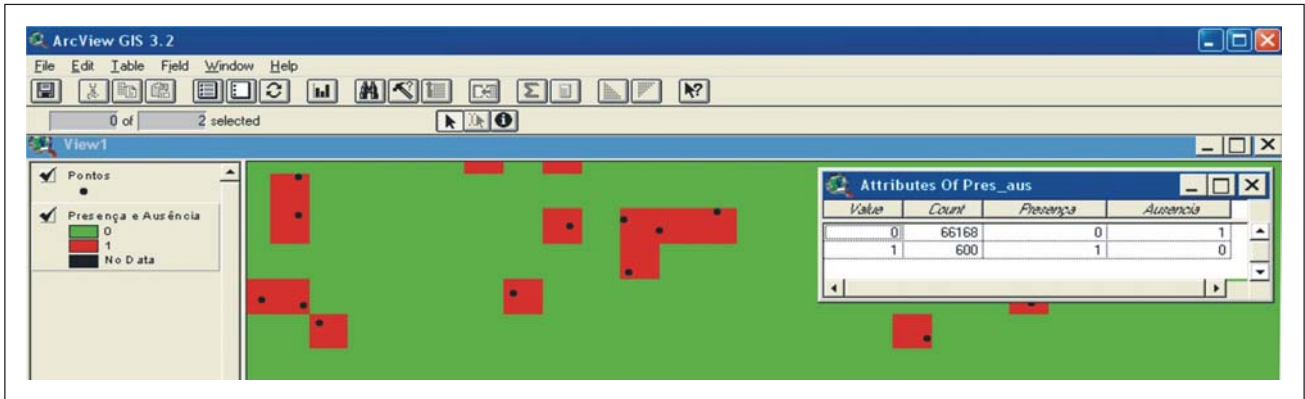


FIGURA 7 – Tabela de atributos do arquivo formato GRID da variável dependente. Os pontos eram um arquivo que estava em formato vetorial que foi transformado para GRID. Os pixels em vermelho correspondem a pelo menos uma presença e os pixels em verde às ausências.

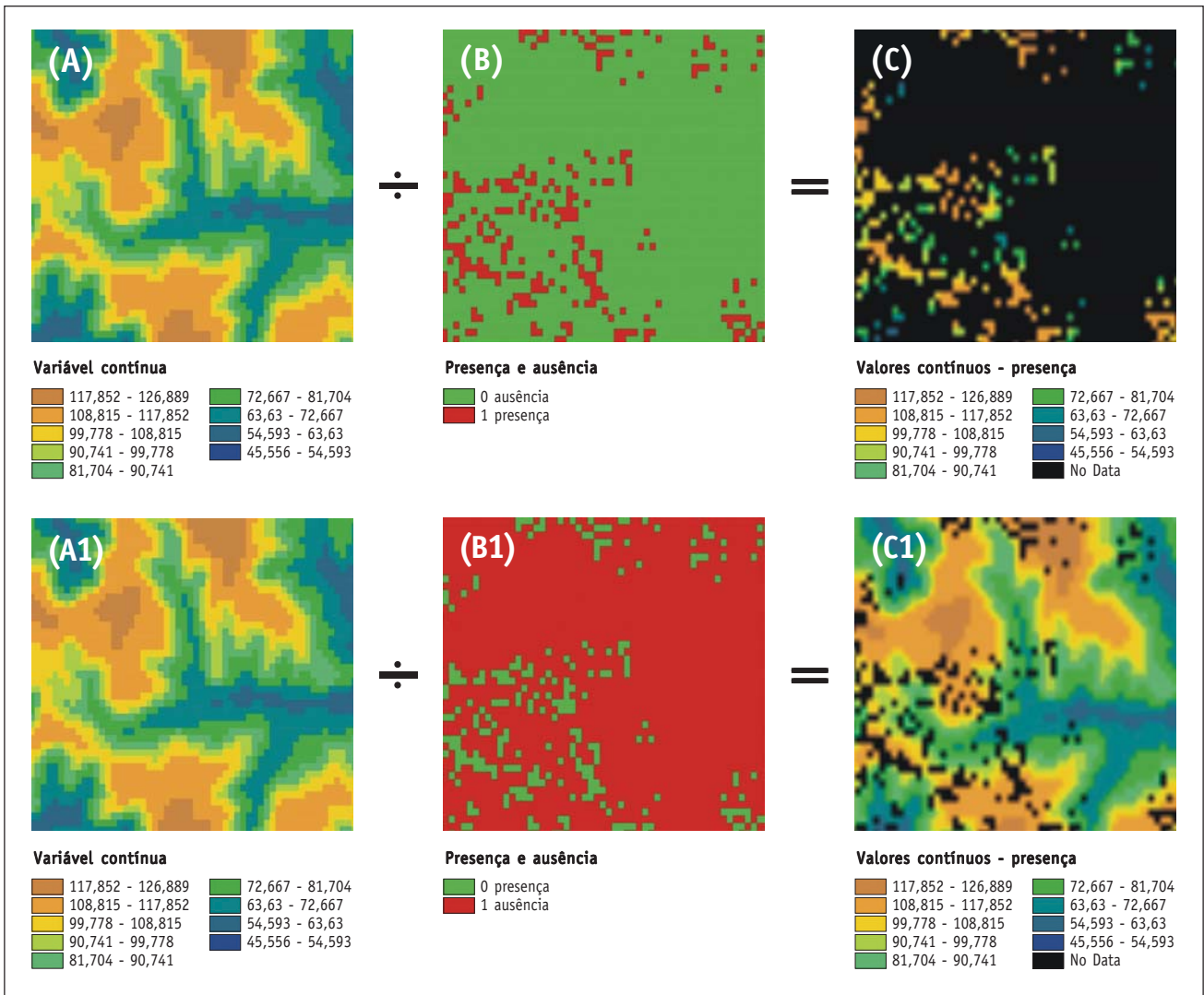


FIGURA 8 – Representação das operações para obtenção dos valores das variáveis contínuas nas células de ausência e presença do evento de interesse. A e A1 são variáveis contínuas; B é a variável dicotômica com valor 1 para presença; B1 é a variável dicotômica com 1 para ausência; C são os valores das células da camada digital da variável contínua com presença e C1 a mesma operação para as células com ausência.