Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 1323-1330.

Análise multivariada aplicada a zoneamento: o método AMAZON

Darcton Policarpo Damião [1]
Newton Moreira de Sousa [2]

[1] Instituto de Estudos Avançados – IEAv
Rod. Tamoios, km 5,5 – 12.228-840 – São José dos Campos – SP, Brasil
darcton@ieav.cta.br

[2] Universidade de Brasília – UnB
Campus Darcy Ribeiro – Brasília – DF, Brasil
nmsousa@unb.br

Abstract. Owner of the world largest rainforest, Brazil put forth its best effort in many different segments of the society looking forward to mitigating Amazon deforestation effects. This paper presents a prediction study of Amazon deforestation throughout multivariate analysis techniques, more specifically a logistic regression model. Given regional characteristics, usage of remote sensing data of medium resolution, organized and modeled in geographical information systems, constitutes an increasingly adopted *praxis*. In this case, the study was intended to raise space-temporal relationships that exist among biophysical variables identified in a single Landsat scene, located in São Félix do Xingu, Southern of Pará State, Brazil, observed along with six different dates from 1985 to 2004. Therefore, the "Multivariate Analysis Applied to Zoning" method (AMAZON) was conceived, tested, and validated. In order to achieve such goal, during the pre-processing stage, the data were divided in four categories – forest, non-forest, hydrograph, and deforestation – taking into account the methodological procedures developed to PRODES DIGITAL by the Brazilian National Institute for Space Research (INPE). Right after selection and formation of the variables to be used in the model, the following step was to generate the logistic regression coefficients in a four-year basis. This procedure played a key role in the understanding of the independent variables effects on the response-variable (environmental impact = deforestation).

Palavras-chave: deforestation, remote sensing, geographical information systems, multivariate analysis, logistic regression, land cover change, desmatamento, sensoriamento remoto, sistemas de informação geográfica, análise multivariada, regressão logística, mudanças na cobertura do solo.

1. Introdução

Há um consenso em torno do entendimento de que o conhecimento integrado dos processos de mudança de **uso** do solo e dos fatores que afetam tais processos possibilita a elaboração de modelos preditivos confiáveis de médio e de longo prazo, em diferentes escalas. Entretanto, é também consenso que a compreensão absoluta desses processos – sintetizada por meio de técnicas de modelagem – depende de uma abordagem transdisciplinar, o que implica uma alta complexidade de concepção e de implantação.

Em função disso, este artigo procura tão somente abordar mudanças na **cobertura** do solo, o que é obtido por meio da análise de uma única externalidade do uso do solo amazônico – o desmatamento – dado que "não há meios de se estudar diretamente as mudanças no uso do solo" (BRIASSOULIS, 2000). Tal abordagem é feita por meio de regressão logística e o método denominado, por conveniência, de 'Análise Multivariada Aplicada a Zoneamento (AMAZON)'.

No AMAZON, onze variáveis (uma dependente e dez independentes, sendo uma de exposição, sete de controle e duas de interação) têm seus efeitos verificados em um número de cerca de 2,5 milhões de eventos (células de floresta *versus* células de desmatamento) contidos em uma cena Landsat TM ao longo de duas décadas.

2. Procedimentos Metodológicos

2.1 Coleta de Dados

Os modelos de mudanças no uso e na cobertura do solo tratam do resultado de uma complexa rede de interações entre forças biofísicas e socioeconômicas sobre o tempo e o espaço, que é conseqüência dessas mudanças. No caso da Amazônia brasileira, tal complexidade atinge patamares críticos, não só pelas características intrínsecas à região, como extensão territorial e dificuldades de acesso, entre outras, mas pela extrema dificuldade em se obter dados consistentes, quando a proposição da pesquisa envolve uma abordagem dos aspectos socioeconômicos.

A partir da idéia de que é possível obter uma indicação apriorística dos fatores que exercem influência nas mudanças na cobertura do solo, optou-se pela utilização de variáveis de ordem biofísica, cuja obtenção é bem menos complexa relativamente às variáveis socioeconômicas.

A "Terra do Meio", denominação atribuída à região do estado do Pará situada "no meio" da faixa de terra que separa os rios Xingu e Araguaia, possui tais características, como de resto as demais regiões localizadas na vasta região que ficou conhecida por "Arco do Desmatamento".

Para que fosse possível estudar a evolução dessas variáveis de forma consistente e sistemática, seria necessária uma área grande o suficiente para abarcar as principais variáveis em todas as suas grandezas possíveis. A melhor relação custo-benefício para uma abordagem como esta foi encontrada no satélite Landsat 5. Em uma única cena, é possível cobrir uma área de 185km x 185km com um conjunto de resoluções que pode ser considerado excelente para aplicações florestais (pixel de 30m x 30m, sete bandas espectrais, revisita a cada 26 dias e vida útil superior a duas décadas).

Como forma de observar o fenômeno do desmatamento por um período de tempo longo o suficiente para inferir relações de dependência entre variáveis, foram selecionadas cenas do período que se estende de 1985 a 2004, via de regra espaçadas a cada quatro anos (exceção feita à primeira, separada em três anos da próxima cena).

Esse período foi amostrado, a partir da cena de 1988, a cada quatro anos, totalizando seis cenas: 1985 (10 de junho), 1988 (18 de junho), 1992 (29 de junho), 1996 (8 de junho), 2000 (30 de junho) e 2004 (26 de maio). A fim de minimizar efeitos típicos de sazonalidade, foram selecionadas cenas da estação seca, que apresentaram boa qualidade (ausência de cobertura de nuvens). O intervalo de quatro anos entre as cenas se deve à combinação de uma amostragem que permite a montagem de uma série temporal suficientemente extensa, com uma quantidade de dados razoavelmente reduzida.

Com vistas a facilitar as ações subseqüentes de interpretação e de classificação das imagens, um extensivo trabalho de campo foi planejado.

Para tanto, a Força Aérea Brasileira destacou dois de seus esquadrões baseados em Belém-PA – 1° Esquadrão de Transporte Aéreo (1° ETA) e 1° Esquadrão do 8° Grupo de Aviação (1°/8° GAV) – para apoiarem a condução do trabalho. O esforço aéreo despendido nas missões de helicóptero (1°/8° GAV) foi da ordem de 15 horas de vôo, distribuídas entre o deslocamento (ida e volta de Belém a S. F. do Xingu) e a missão de sobrevôo propriamente dita. Outras 7 horas foram voadas em aeronaves Bandeirante (1° ETA), em favor do reabastecimento dos helicópteros.

2.2 Organização dos Dados de Entrada

A área-teste escolhida para o estudo é delimitada pela "órbita/ponto" 225/64 do satélite Landsat 5, localizada no município de São Félix do Xingu, região Sul do estado do Pará.

Em sua preparação, o AMAZON leva em conta parâmetros que permitem verificar o relacionamento da variável dependente com as variáveis independentes, de exposição e de

controle. As operações que permitem conhecer tal relacionamento foram conduzidas no ArcGIS 9.2.

Os parâmetros mencionados acima são extraídos de planos de informação (PI) que compõem o projeto sobre o qual o modelo é processado. Parte desses PI advém de dados vetoriais colhidos junto a fontes secundárias, como IBGE e Imazon, por exemplo, e são alimentados diretamente no ArcGIS.

Outros PI têm que ser preparados a partir de um tratamento específico dado às imagens orbitais, responsáveis por retratar as mudanças biofísicas que ocorrem na área de estudo e que são o alvo deste artigo. Essa preparação foi conduzida no aplicativo SPRING 4.3.2.

Considerando os parâmetros necessários à geração do AMAZON, as imagens foram classificadas em 'floresta', 'não-floresta', 'desmatamento' e 'hidrografia'.

São inúmeros os procedimentos metodológicos existentes na literatura concebidos para possibilitar classificações temáticas mais precisas de uso e de cobertura do solo. De maneira geral, o estabelecimento desses procedimentos é dependente de dois fatores principais: o sistema sensor utilizado na coleta dos dados e a natureza do alvo classificado. Por exemplo, imagens de radar conduzem a procedimentos metodológicos bastante distintos daqueles desenvolvidos para imagens ópticas. O mesmo pode ser dito em relação à classificação de áreas urbanas e áreas florestais.

Esses fatores ainda possuem características subordinadas, por assim dizer. No caso dos sistemas sensores, cabem destacar: resolução e abrangência espectral, resolução espacial e polarização, por exemplo. No caso dos alvos, há que se considerar sua extensão, existência ou não de sazonalidade, as classes de uso e de cobertura do solo nas quais se tem interesse, entre outras coisas. Uma vez combinadas, essas características geram possibilidades virtualmente irrestritas de procedimentos de classificação.

No caso deste estudo, considerando o fato de que a classificação, ainda que importante, não passa de uma etapa intermediária para a formulação do modelo a ser testado, buscou-se um procedimento metodológico consagrado para o binômio 'sensor óptico multiespectral *x* área de floresta úmida'.

Assim sendo, foi escolhido o procedimento desenvolvido pelo INPE para o PRODES DIGITAL (CREPANI *et. al.*, 2001). Trata-se de um método operacionalmente consagrado para a classificação do desmatamento na região amazônica, composto de cinco etapas, a saber: (*i*) geração das imagens sintéticas; (*ii*) segmentação das imagens sintéticas; (*iii*) geração do arquivo de contexto e extração de regiões; (*iv*) classificação da imagem segmentada; e (*v*) mapeamento da imagem segmentada.

Para a escolha e a geração das variáveis, a concepção de um método relativamente simples como o AMAZON pressupõe um número limitado de variáveis, dado o volume computacional a ser gerado no processo de regressão. Não se deve perder de vista o fato de que uma cena Landsat possui um número superior a 40 milhões de pixels, que devem ser analisados para cada uma das variáveis escolhidas e para cada uma das datas consideradas.

Essa quantidade de dados é de tal monta que não foi possível computá-los em sua máxima resolução. Assim, esse número foi reduzido em 16 vezes, por meio do agrupamento dos pixels em células de formato 4 *x* 4, manobra que tornou possível executar a regressão logística do AMAZON no ambiente computacional empregado (*Pentium 4* de 3.4 GHz, 2 GHz de RAM).

Além da variável dependente 'impacto ambiental' IPT, que é binária (0,1), oito variáveis independentes – PAD, ROD, HID, EPA, DEC, ALT, IFR e APE, que são explicitadas a seguir – foram inicialmente criadas com base na percepção de quais fatores têm o potencial de exercer influência na dinâmica do desmatamento e, naturalmente, tendo em vista a literatura consultada. A partir dessas variáveis independentes, outras duas – W1 e W2 – foram geradas a partir da interação multiplicativa entre a variável independente de exposição e duas variáveis independentes de controle.

O método estatístico de regressão – linear ou logística – não pressupõe necessariamente uma relação temporal entre a variável dependente e as variáveis independentes. Em geral, busca-se relacionar, para uma dada amostra de tamanho 'n', alguns fatores "explicativos" – as variáveis independentes – para um outro fator "explicado" – a variável dependente. A partir do estabelecimento de uma relação estatística entre esses fatores, torna-se possível "predizer" o parâmetro explicado do evento 'n+1', tomando-se por base seus atributos explicativos.

Com o AMAZON, entretanto, tomando esse evento como sendo uma célula da imagem Landsat, não se pretende predizer o parâmetro explicado da célula 'n+1' da imagem, mesmo porque tal célula não existe, mas sim todas as 'n' células cobertas por essa imagem.

Dessa forma, a relação entre a variável dependente e as variáveis independentes só faz sentido se houver um lapso temporal entre IPT e as demais variáveis, após o qual o efeito das variáveis explicativas sobre a variável explicada possa ser verificado. Para tanto, aos "anos-base" foram atribuídos os parâmetros das variáveis independentes, cuja regressão logística permite predizer, em termos probabilísticos, a variável dependente da data subseqüente.

Uma vez que, das quatro classes inicialmente geradas – floresta, desmatamento, hidrografia e não-floresta – as duas últimas permaneceram virtualmente inalteradas ao longo de todo o período estudado, elas foram reclassificadas e passadas à categoria de *NoData*. Com isso, criou-se uma seqüência de imagens divididas nas classes 'floresta' e 'desmatamento'.

Como a variável-resposta deve ser verificada na data seguinte àquela do ano-base, é essencial que a área na qual essa variável é válida seja idêntica àquela do ano-base, onde são "interrogadas" as variáveis independentes. Assim, evita-se a comparação de áreas de tamanhos diferentes. Assim, a "máscara" do ano-base é composta somente pela classe 'floresta'. Na data seguinte, as células dessa área são confrontadas com as mesmas células, agora divididas em 'floresta' e 'desmatamento'. Quando essa data se torna o ano-base em relação à data subseqüente, a classe 'desmatamento' é reclassificada como *NoData* e assim por diante. Ao final, as seis datas consideradas no estudo (1985, 1988, 1992, 1996, 2000 e 2004) geram cinco modelos, referenciados nos anos-base.

No modelo inicial (Equação 1), todas as variáveis concebidas para o AMAZON, em um dado ano-base, são usadas na regressão logística que estabelece a probabilidade de ocorrência de impacto ambiental relativa àquele ano-base, a ocorrer na próxima data.

$$Logito\left[P\left(IPT_{pd}\right)\right] = \ln\left[\frac{P\left(IPT_{pd}\right)}{1-P\left(IPT_{pd}\right)}\right] = \beta_0 + \beta_1 PAD_{ab} + \beta_2 ROD_{ab} + \beta_3 HID_{ab} + \quad (1)$$

$$+ \beta_4 EPA_{ab} + \beta_5 DEC_{ab} + \beta_6 ALT_{ab} + \beta_7 IFR_{ab} + \beta_8 APE_{ab} + \beta_9 W_{1ab} + \beta_{10} W_{2ab}$$

onde $P(IPT_{pd})$ é a probabilidade do impacto ambiental provocado pelo desmatamento na próxima data;
PAD_{ab} é a proximidade da área desmatada no ano-base;
ROD_{ab} é a proximidade de rodovias no ano-base;
HID_{ab} é a proximidade de hidrovias no ano-base (não implementada nesta tese);
EPA_{ab} é a existência (ou não) de área protegida no ano-base;
DEC_{ab} é a declividade no ano-base;
ALT_{ab} é a altitude do terreno no ano-base;
IFR_{ab} é o índice de fragmentação no ano-base;
APE_{ab} é a área do polígono envolvente da célula de floresta no ano-base;
W_{1ab} é o produto *PAD* x *DEC* no ano-base;
W_{2ab} é o produto *PAD* x *ALT* no ano-base; e
β_i são os coeficientes das variáveis independentes.

3. Resultados e Discussão

O desempenho de cada uma das variáveis revela a primeira e mais marcante característica do AMAZON, que é a mudança na representatividade de suas variáveis, à medida que evolui a cobertura do solo. Uma percepção da mudança na cobertura do solo para a área de estudo pode ser obtida a partir da seqüência de imagens classificadas em hidrografia (azul), não-floresta (magenta), floresta (verde) e desmatamento (amarelo), apresentadas na Figura 2.

Figura 2. Evolução das classes ‘floresta’ e ‘desmatamento’, de 1985 a 2004.

Partindo do modelo inicial, as variáveis são eliminadas uma a uma (para tanto, basta fazer o respectivo parâmetro ‘β’ igual a ‘0’), no intuito de verificar qual é a configuração que adere melhor à predição de áreas desmatadas.

A configuração mais eficiente é obtida com base em tabelas de classificação, nas quais os valores esperados para 'floresta' e 'desmatamento' são confrontados com os valores observados para essas mesmas classes. Com isso, são gerados os termos de ajuste do modelo[1] (AM), de verdadeiros positivos[2] (VP) e de falsos positivos[3] (FP), que permitem escolher o modelo preservado para o ano-base em questão, com base no desempenho do modelo (DM).

A Tabela 1 sintetiza os parâmetros envolvidos na construção das tabelas de classificação empregadas em cada um dos modelos.

Tabela 1 – Síntese das tabelas de classificação utilizadas para avaliar o AMAZON.

NOME DO MODELO PRESERVADO PARA O ANO-BASE		
Total de células da máscara [(+ +) + (+ –) + (– +) + (– –)]	Desmatamento observado (+)	Floresta observada (-)
Desmatamento esperado (+)	Número de células '+ +'	Número de células '– +'
Floresta esperada (–)	Número de células '+ –'	Número de células '– –'
$AM = \frac{(++)+(--)}{(++)+(+-)+(-+)+(--)}$	$VP = \frac{(++)}{(++)+(+-)}$	$FP = \frac{(-+)}{(-+)+(--)}$

Lesschen *et. al.* (2005) recomendam as tabelas de classificação para comparar resultados de modelos. Embora não existam regras gerais para se julgar os valores considerados aceitáveis, ajustes de modelos com percentuais superiores a 50% são estatisticamente melhores do que a aleatoriedade.

Pontius e Schneider (2001), por sua vez, defendem que um valor superior a 70% é considerado aceitável para a modelagem de uso e de cobertura do solo, enquanto Hosmer e Lemeshow (2000) consideram percentuais superiores a 80% como excelentes, e superiores a 90% como excepcionais.

Para os fins deste estudo, o modelo a preservar é aquele que apresenta o maior índice resultante da operação abaixo (desempenho do modelo – DM), tomando-se o cuidado de não considerar modelos com valores de 'AM' e 'VP' inferiores a 70%, assim como com valores de 'FP' superiores a 30%. Enquanto o critério para os valores de 'AM' segue as orientações de Pontius e Schneider (2001), o critério de verdadeiros positivos superiores a 70% e de falsos positivos inferiores a 30% é adotado neste trabalho por similaridade de raciocínio. Em conseqüência, estipula-se que DM (Equação 2) seja superior a 0,28 [(0,7 * (0,7 -0,3)].

$$DM = AM \times (VP - FP) \quad (2)$$

[1] - Percentual resultante do somatório das predições corretas para as classes 'floresta' e 'desmatamento', dividido pelo total de células da máscara analisada.

[2] - Percentual resultante da divisão do número de células corretamente preditas para a classe 'desmatamento' pelo número total de células dessa classe.

[3] - Percentual resultante da divisão do número de células erroneamente preditas para a classe 'desmatamento' pelo número total de células da classe 'floresta'.

4 Conclusões

Tomando como exemplo o ano-base de 2000 e partindo do modelo gerado na ferramenta ArcGrid do ArcGIS 9.2 (Figura 3a), um mapa de probabilidade de desmatamento é gerado (Figura 3b) e posteriormente categorizado em duas classes (valores > 0,5 ⇨ '1'; valores < 0,5 ⇨ '0'). A essa imagem (Figura 3c), é somada a variável dependente 'IPT' (Figura 3d) da data subseqüente. O resultado, dividido em quatro classes (Figura 3e) – floresta (- -), desmatamento (+ -), floresta (- +) e desmatamento (+ +) – permite a geração da tabela de classificação (Figura 3f) que quantifica o desempenho do modelo. Ao modelo, são atribuídas as letras 'P-R-E-D-A', o que significa dizer que aquele modelo faz uso das variáveis 'PAD', 'ROD', 'EPA', 'DEC', e 'ALT'.

O critério para a escolha do melhor modelo referente a cada ano-base é arbitrado pela Equação 2. Para cada um dos anos considerados, também é possível observar outros resultados promissores, respeitados os critérios de 'DM' superior a '0,28', 'AM' e 'VP' superiores a 70% e 'FP' inferior a 30%.

De maneira geral, nos primeiros estágios do desmatamento (1985, 1988 e 1992), o AMAZON apresenta melhor desempenho somente com as variáveis 'PAD', 'DEC' e 'ALT'. À medida que a atividade de desmatamento é intensificada, o desempenho do AMAZON melhora com o incremento de novas variáveis, como 'EPA' e 'ROD', e até mesmo com as variáveis 'IFR', 'APE' e as multiplicativas ('W1' e 'W2'). A partir de 2000, as variáveis 'IFR' e 'APE' passaram a ter influência no modelo.

As conclusões obtidas a partir dos diversos modelos gerados ano a ano serviram de base de entendimento e de validação das variáveis escolhidas para a formulação do AMAZON. Para tanto, a variável dependente ('IPT') é obtida pela classificação 'floresta *x* desmatamento' da data subseqüente à do ano-base ao qual se aplica o modelo.

Como foi dito acima, isso permite um melhor entendimento do papel de cada uma das variáveis preservadas do modelo. Por outro lado, essa abordagem não torna possível a proposta original de predição[4] da atividade de desmatamento para datas futuras, uma vez que 'IPT' não está disponível nesses casos.

Referências Bibliográficas

Briassoulis, H. **Analysis of Land Use Change: Theoretical and Modeling Approaches**. The Web Book of Regional Science. 2000. Disponível em: http://www.rri.wvu.edu/WebBook/Briassoulis/contents.htm. Acesso em: 13 set 2004.

Crepani, E.; Duarte, V.; Shimabukuro, Y. E. **Sensoriamento Remoto e Geoprocessamento no Mapeamento Regional da Cobertura e Uso Atual da Terra**. São José dos Campos, INPE, 36p. (INPE -8478-NTC/346). 2001.

Hosmer, D. W.; Lemeshow, S. **Applied Logistic Regression**, 2nd Edition, Wiley, New York. 2000.

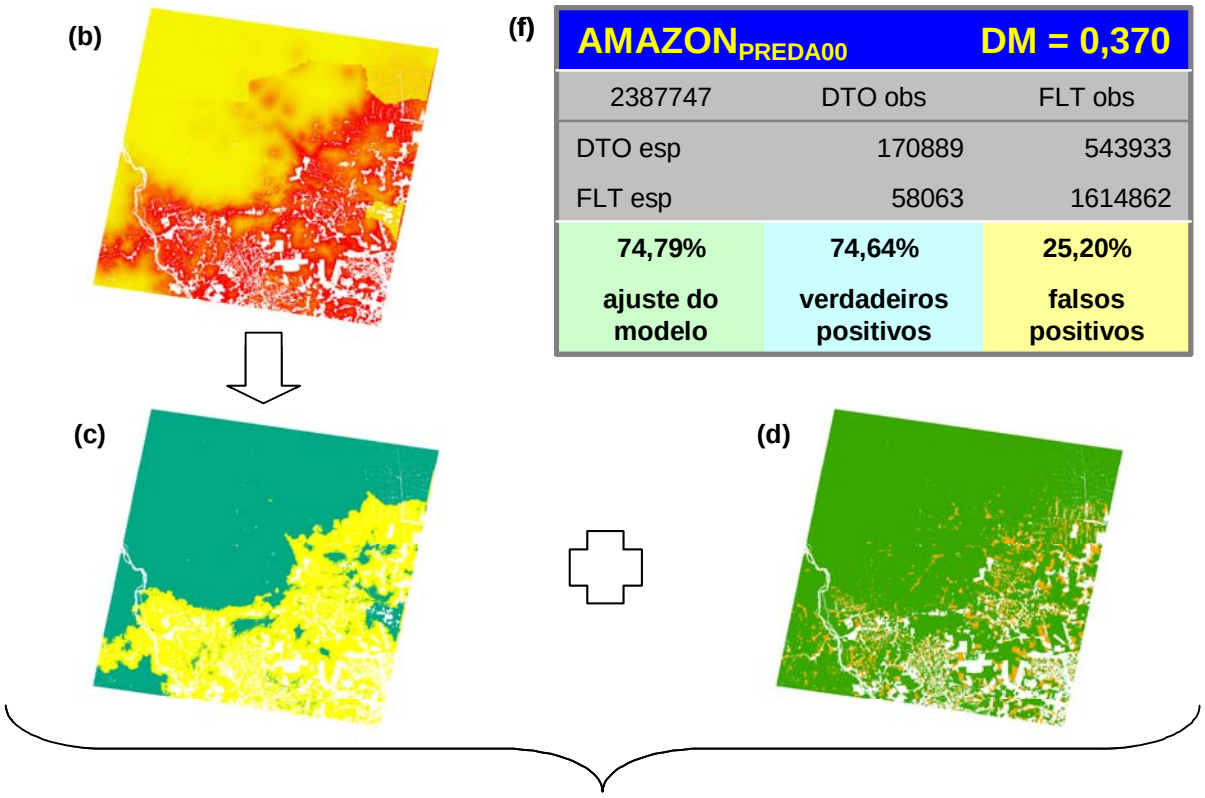
Lesschen, J. P.; Verburg, P. H.; Staal, S. J. **Statistical Methods For analyzing The Spatial Dimension of Changes in Land Use and Farming Systems - LUCC Report Series N° 7**, Wageningen, The Netherlands: Wageningen University. 2005.

Pontius Jr, R. G.; Schneider, L. C. **Land-Cover Change Model Validation by an ROC Method for the Ipswich Watershed**, Massachusetts, USA. Disponível em www.elsevier.com/locate/agee.

[4] - Nunca é demais enfatizar que a predição estatística não diz respeito a tempos futuros, mas sim a novos eventos.

(a)

$$Logito(IPT_{04}) = -0{,}703 - 0{,}208PAD_{00} - 0{,}067ROD_{00} - 2{,}145EPA_{00} - 0{,}210DEC_{00} + 0{,}087ALT_{00}$$



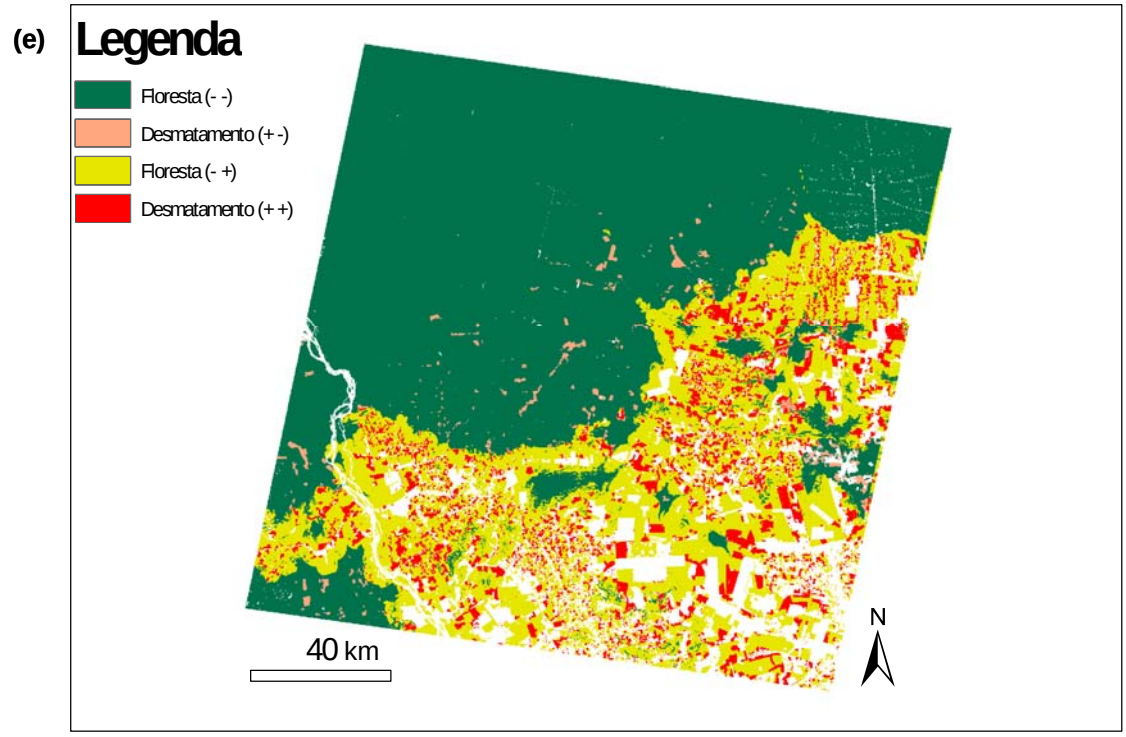


Figura 3. Melhor desempenho do AMAZON no ano-base de 2000.