Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 1103-1110.

Análise de regressão linear múltipla para estimativa do índice de vegetação melhorado (EVI) a partir das bandas 3, 4 e 5 do sensor TM/Landsat 5

Silvia Cristina de Jesus [1]
Adalberto Koiti Miura [1]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brasil
{silviac, miura}@ dsr.inpe.br

Abstract. This work has as main objective to obtain an Enhanced Vegetation Index, EVI, from an incomplete set of orbital imagery of the Landsat 5 TM from the year of 1986, containing merely the 3 (red), 4 (infrared near) and 5 (short wave infrared) bands. The mains purpose of this study was to provide data for use in another study, concerning land use and land cover dynamics. Linear regression model analysis techniques was choose for estimate the EVI for the year of 1986, from a model tested and validated with datasets of spectral bands dated 1992, considering as explanatory variable just the bands 3, 4, 5 and the NDVI. The estimated EVI was evaluated on the basis of the residual analysis and by interpretation of the image of difference between the observed and estimated EVI. When calculated EVI (1992) was compared to the estimated one to same date, the result was very satisfying, which can be proved with residuals analysis and the image of residuals. Sites of underestimate and overestimate are located in regions which EVI usually presents higher values as dense trees and clouds, while lower values in areas of bare soil and water. Estimated EVI just potentialize maximum and minimum values and practically does not modify the index in relative values. When regression model based on 1992 data was applied to the validation of 1994 data, time lag and seasonality effects can be perceived. That bias was not observed when the model was generated through 1994 data and applied to the same year.

Palavras-chave: vegetation index, EVI, NDVI, multiple regression, Landsat 5 TM, índices de vegetação, regressão múltipla.

1. Introdução

Os índices de vegetação são tipos específicos de transformação radiométrica (lineares ou não), comumente utilizadas no intuito de avaliar os recursos naturais e monitorar a cobertura vegetal, em especial na detecção de mudanças de uso e cobertura das terras, pois apresentam uma relação direta e satisfatória com a fitomassa foliar verde. O princípio básico da aplicação dos índices de vegetação reside na premissa que os dosséis verdes e a vegetação saudável possuem interações distintas em regiões do espectro eletromagnético correspondente às faixas do visível e do infravermelho próximo, assim como o comportamento da água e dos solos. Dentre os muitos índices de vegetação existentes, o NDVI, índice de vegetação da diferença normalizada , e o EVI, índice de vegetação melhorado (HUETE, et al. 1994), podem ser relacionados como uns dos mais utilizados (Equações 1 e.2).

$$NDVI = \left(\frac{\rho_{(\mathrm{IVP})} - \rho_{(Vermelho)}}{\rho_{(IVP)} + \rho_{(Vermelho)}} \right) \quad (1)$$

$$EVI = G\left(\frac{\rho_{(IVP)} - \rho_{(Vermelho)}}{\rho_{(IVP)} + C_1\rho_{(Vermelho)} - C_2\rho_{(Azul)} + L} \right) \quad (2)$$

Onde: $\rho_{(\mathrm{Azul})}$ = Fator de reflectância bidirecional na banda do azul; $\rho_{(\mathrm{Vermelho})}$ = Fator de reflectância bidirecional na banda do vermelho; $\rho_{(\mathrm{IVP})}$ = Fator de reflectância bidirecional na banda do infravermelho próximo; C_1 *e* C_2 = Coeficientes de ajuste para

efeito de aerossóis da atmosfera para as bandas do vermelho e azul, respectivamente ,6[*] e 7,5[*]; *L* = Fator de ajuste para o solo, 1[*]; *G* = Fator de ganho, 2,5[*].

O NDVI e o EVI são muito semelhantes e ambos são utilizados no monitoramento de espaço-temporal da vegetação com o propósito de estudar padrões da atividade fotossintética em uma consistente base global, podendo ser utilizado usado para comparações de variações sazonais, interanuais e de longo prazo da estrutura da vegetação, fenologia e parâmetros biofísicos, bem como no monitoramento da dinâmica de uso e ocupação das terras. O EVI melhora algumas respostas do NDVI, por ser mais sensível às variações estruturais e arquitetônicas do dossel, de fitofisionomias com maior densidade de biomassa, além de reduzir as influências atmosféricas e do solo (HUETE et al., 2002).

O presente trabalho surgiu da necessidade de se comparar diferenças na cobertura das terras em dados multitemporais, por meio do uso de índices de vegetação. Como nem sempre é possível adquirir conjuntos completos de bandas espectrais para as datas consideradas, em função do custo e disponibilidade dos mesmos, tornou-se necessário estimar o índice de vegetação melhorado (EVI) a partir de um modelo de regressão linear que considerasse o conjunto de bandas disponíveis, em geral aquelas referentes ao vermelho, infravermelho próximo e infravermelho de ondas curtas (bandas 3, 4 e 5 do Landsat 5 TM).

2. Materiais e Métodos

2.1 Aquisição dos dados de reflectância

Os dados utilizados para a construção do modelo de regressão linear foram extraídos de três conjuntos de imagens Landsat 5 TM, do Município de Lucas do Rio Verde (MT), de órbita/ponto 221/69, datadas, respectivamente, de 20.12.1986, 24.04.1992 e 21.09.1994. As cenas, compostas por sete bandas espectrais, abrangem comprimentos de onda relativos às faixas: 1) azul (0.45 – 0.52 μm); 2) verde (0.52 - 0.60 μm); 3) vermelho (0.63 - 0.69 μm); 4) infravermelho próximo - NIR (0.76 – 0.90 μm) ; 5) Infravermelho de ondas curtas – SWIR (1.55 – 1.75 μm); 6) infravermelho termal – TIR (10.4 – 12.5 μm); 7) e uma segunda banda posicionada no infravermelho de ondas curtas – SWIR (2.08 – 2.35 μm). Apenas a data de 1986 apresentou-se incompleta, contando com as bandas 3, 4 e 5.

As imagens foram registradas com base em uma cena Landsat 7 ETM, ortorretificada, proveniente do acervo do *Global Land Cover Facility* (NASA, 2003). Adicionalmente, foi realizada uma correção atmosférica, com base na subtração do pixel escuro ou *DOS (Dark Objects Subtraction)* em todas as imagens, para minimizar os efeitos de espalhamento, absorção e refração da energia eletromagnética. Isto permite compatibilizar os dados de grandes extensões territoriais ou levantamentos multitemporais, uniformizando-os na mesma escala radiométrica (CHAVEZ, 1988). Os valores digitais de nível de cinza foram transformados em valores reflectância aparente, permitindo a comparação do comportamento espectral entre diferentes datas e locais. Esta transformação compensa as diferenças entre ganhos e "*off-set*" de cada banda espectral, bem como concilia as diferenças quanto a irradiância solar no topo da atmosfera e o ângulo de incidência da radiação sobre o alvo, no momento da aquisição da imagem.

[*] coeficientes para o sensor MODIS (HUETE et al., 1994).

2.2 Amostragem, preparação dos dados e seleção de variáveis para o modelo

Para a análise de regressão, foram amostrados de forma aleatória dois conjuntos distintos de 1000 coordenadas de pixels, utilizados para extração dos valores de reflectância aparente de todas as bandas, em todas as imagens. O primeiro conjunto foi utilizado na obtenção do modelo de regressão de seu teste, enquanto o outro forneceu dados para validação. Para a seleção de variáveis e proposição do modelo de regressão a ser utilizado foi implementado o procedimento *best subset* (NETER et al, 1996) por meio do critério R^2 e do R^2 ajustado, com as variáveis de imagem RTM3, RTM4, RTM5 e NDVI, sendo este último excluído da formulação do modelo final em virtude da multicolinearidade com as demais variáveis preditoras. Isto se deve ao fato de que o NDVI é um índice obtido por uma relação não linear entre as bandas 3 (vermelho) e 4 (infravermelho próximo). Para avaliar se a reta de regressão ajustada pode ser empregada na estimativa do EVI em outras imagens, foi conduzida a avaliação de sua adequação, por meio da análise dos resíduos, ao nível de significância de 5% . Os testes de Levene e Shapiro-Wilk (NETER et al., 1996) foram aplicados para verificar, respectivamente, a homocedasticidade e a normalidade dos dados. De acordo com Neter et al.(1996), quando o número de observações é muito maior do que o número de parâmetros, o efeito da dependência de resíduos é relativamente sem importância e pode ser descartado. A condição de normalidade dos resíduos não está relacionada, necessariamente, à obtenção dos estimadores de mínimos quadrados, mas é imprescindível para a definição de intervalos de confiança e testes de significância, isto é, os estimadores são não-tendenciosos, mas os testes não têm validade (SOUZA et al., 2005). Os limites de confiança para β foram obtidos pela Equação 3.

$$b - t_{1-\alpha/2;\, n-2} s(b) \leq \beta \leq b + t_{1-\alpha/2;\, n-2} s(b) \quad (3)$$

2.3 Validação do modelo de regressão linear

Quando um modelo de regressão é desenvolvido a partir de um determinado conjunto de dados, é inevitável que o modelo selecionado seja escolhido, pelo menos em parte, porque ele se ajusta bem ao conjunto de dados disponíveis. Para um grupo diferente de resultados aleatórios é possível que um modelo diferente seja obtido, em termos de variáveis preditivas selecionadas e/ ou suas formas funcionais e termos de interação presentes no modelo. Um resultado do desenvolvimento deste modelo é que a média do quadrado dos erros (MSE) apresente a tendência de suavizar a variabilidade inerente nas predições futuras do modelo selecionado. A mensuração da capacidade preditiva real do modelo de regressão formulado pode ser feita empregando o próprio modelo para predizer cada caso no novo conjunto de dados e, então, calcular a média do quadrado dos erros de predição (MSPR), conforme a Equação 4:

$$MSPR = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y})^2}{n^*} \quad (4)$$

Onde Y_i é o valor observado da variável no i-ésimo caso de validação; $\hat{Y}$ é o valor predito para o i-ésimo caso de validação baseado no conjunto de dados do modelo de regressão; e n^* é o número de casos no conjunto de dados de validação

Caso o valor de MSPR seja próximo do valor de MSE do conjunto de dados usados no ajuste do modelo de regressão, então o MSE para o modelo de regressão formulado não é viciado de forma grave e fornece uma indicação apropriada da capacidade preditiva do modelo. Se o MSPR é muito maior do que o MSE, é possível utilizar o MSPR como um indicador da aderência do modelo de regressão em predições futuras (NETER et al., 1996).

3. Resultados

3.1 Análises prévias

A seleção das variáveis independentes apontou as bandas 3, 4 e 5 como o melhor conjunto para a determinação do modelo de regressão, seguidos pela combinação das bandas 3 e 4 e pelo índice vegetativo NDVI. Os conjuntos de variáveis explicativas que contivessem o NDVI associado às bandas espectrais concebidas para a elaboração do modelo foram desconsiderados devido à grande correlação existente. Isto se deve ao fato de que o NDVI é um índice obtido por uma relação não linear entre as bandas 3 e 4. Uma análise preliminar não indicou a necessidade de transformação das variáveis. O modelo de regressão linear selecionado é representada pelas equações 5 e 6:

$$Y = 0{,}11673 - 2{,}64706 * X_1 + 2{,}49091 * X_2 + 0{,}53735 * X_3 \quad (5)$$

$$\hat{EVI} = 0{,}11673\ \ 2{,}64706 * RTM3 + 2{,}49091 * RTM4 + 0{,}53735 * RTM5 \quad (6)$$

Com o objetivo de quantificar o relacionamento linear entre X_i e Y, visualizado a partir da análise do diagrama de dispersão, o coeficiente de correlação amostral R = 0,98433036, revelando a existência de uma forte correlação positiva entre as variáveis. Foi calculado também o coeficiente de determinação que mede a proporção da variação que é explicada pelas variáveis independentes no modelo de regressão, que foi R^2 = 0,96890626. Para os dados utilizados o valor calculado para o teste DFFITS foi de 0,1265, sendo eliminados três pixels, que apresentaram valores anômalos por se localizarem sobre defeitos da imagem. Os novos valores de R e R^2 foram, respectivamente, 0,98468861 e 0,96961166, e o novo modelo é representado pela Equação 7.

$$\hat{EVI} = 0{,}12569 - 2{,}72194 * RTM3 + 2{,}45483 * RTM4 - 0{,}51716 * RTM5 \quad (7)$$

Assumindo um nível de confiança igual a 0,05 para uma amostra de tamanho n=997, o teste de Shapiro-Wilk resultou um valor de W igual a 0,73980, 0,92530, 0,78099, para as bandas 3, 4 e 5, respectivamente. Sendo este um valor superior ao tabelado e *p* nulo. A aplicação dos testes de normalidade e o gráfico de probabilidade normal para os resíduos (Figura 1a) mostraram que os dados de todas as variáveis estudadas tiveram distribuição que não difere da normal. O modelo foi submetido ao teste de Levene, que apontou a variância constante dos resíduos. Analisando os gráficos de resíduos contra os valores preditos (Figura 1b), verifica-se que os pressupostos de linearidade e homocedasticidade não foram violados.

a) b)

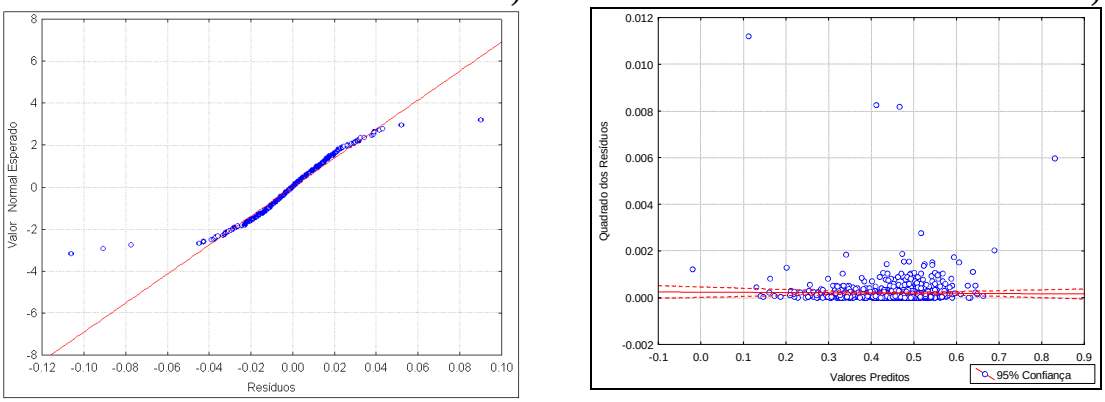

Figura 1. Análise da linearidade e homocedasticidade dos dados: a) Normalidade dos resíduos; b) Diagrama de dispersão dos resíduos a partir do $\hat{EVI}$.

3.2 Validação do modelo de regressão linear

Assumindo um valor de significância de 0.05, os intervalos de confiança para β_0 , β_1, β_2 e β_3 no modelo e para as amostras de validação são apresentados na Tabela 1.

Tabela 1. Intervalos de confiança para β_0 , β_1, β_2 e β_3

Modelo	Amostra de validação (1992)	Amostra de validação (1992)	Amostra de validação (1994)
$0{,}10921 < \beta_0 < 0{,}124255$	$0{,}121718 < \beta_0 < 0{,}136386$	$0{,}122331 < \beta_0 < 0{,}132768$	$1{,}969882 < \beta_0 < 2{,}151859$
$-2{,}76365 < \beta_1 < -2{,}53048$	$-3{,}0082 < \beta_1 < -2{,}71241$	$-2{,}89045 < \beta_1 < -2{,}68658$	$-0{,}00624 < \beta_1 < -0{,}0046$
$2{,}455326 < \beta_2 < 2{,}526493$	$2{,}395172 < \beta_2 < 2{,}463447$	$2{,}416849 < \beta_2 < 2{,}465609$	$-0{,}00635 < \beta_2 < -0{,}00513$
$-0{,}59193 < \beta_3 < -0{,}48277$	$-0{,}51255 < \beta_3 < -0{,}38422$	$-0{,}52837 < \beta_3 < -0{,}43929$	$-0{,}00304 < \beta_3 < -0{,}00246$

A comparação dos possíveis valores de β_0 , β_1, β_2 e β_3 mostra que os coeficientes linear e angular para o modelo de regressão e para as amostras de validação de 1992 (n=1000 e n=1997) são semelhantes. Por outro lado, os intervalos de confiança calculados para a amostra de validação de 1994 não são condizentes com aqueles do modelo de regressão.Nenhum dos coeficientes pode assumir valor zero e, portanto, o $\hat{EVI}$ deve ser explicado, necessariamente, pelas três variáveis consideradas. Os coeficientes de regressão estimados, seus desvios padrões, e outras estatísticas do modelo são mostrados na Tabela 2.

Tabela 2. Resultados dos modelos de regressão a partir das amostras de teste e de validação

	Amostras de teste	Amostras de validação		
Parâmetro	**1992**	**1992**	**1992**	**1994**
n	997	1000	1997	1000
MSE	0,000196	-	-	-
MSPR	-	0,000183	0.000179	0,014631
R^2	0,98392594	0,98433036	0,98448730	0,82649329

Para validar o modelo de regressão, foram selecionados aleatoriamente outros mil pixels da imagem de 1992 e mil pixels da imagem de 1994. A amostra de teste (1992, n = 997) foi adicionada à amostra de validação do mesmo ano. Os coeficientes de regressão estimados, seus desvios padrões, e algumas estatísticas do modelo de regressão ajustado aos dados de validação também são exibidos na Tabela 01. Os valores mostram a concordância entre os coeficientes das amostras de validação de 1992, ao passo que os dados obtidos em 1994 não forneceram um modelo de regressão com coeficientes similares aos do modelo formulado. O mesmo acontece com os valores de R^2. Por outro lado, os valores de MSPR são similares ao MSE do modelo de regressão escolhido. Estes resultados de validação mostram que o modelo selecionado é apropriado apenas para a estimativa do EVI na imagem obtida em 1992, não sendo adequado para análises multitemporais da cobertura vegetal. A Figura 2 apresenta a dispersão dos resíduos da aplicação do modelo de regressão estabelecido nos dados obtidos a partir da imagem de 1994. O mesmo conjunto de validação gerou uma nova regressão. A configuração dos pontos aponta que os dados de reflectância da imagem de 1994 não se ajustam ao modelo estabelecido, dado que é verificado um padrão na distribuição dos pontos.

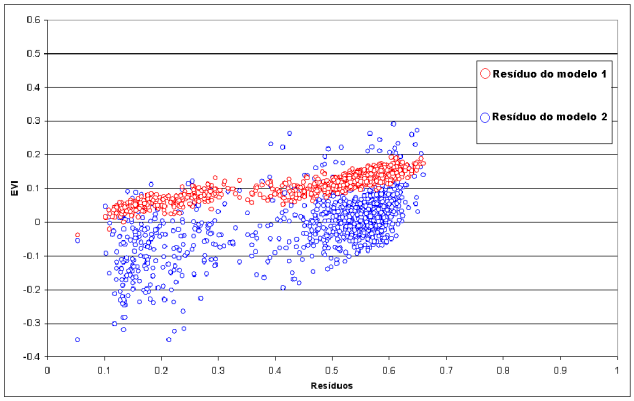

Figura 2. Comparação entre a relação EVI_1994 com os resíduos do Modelo 1 (1992) e do Modelo 2 (1994);

3.3 Aplicação do modelo de regressão sobre a imagem

A partir da validação do modelo, partiu-se para sua aplicação sobre o conjunto de 3 bandas TM (3, 4 e 5) de 1992, de forma a obter um produto EVI estimado que pudesse ser comparado com o índice de vegetação EVI tradicional, obtido pela transformação proposta por Huete et al. (1994). As imagens resultantes podem ser apreciadas e comparadas na Figura 3. De maneira geral, as diferenças entre as duas imagens são pouco perceptíveis, exceto, talvez, pelo ligeiro contraste maior apresentado pelo EVI tradicional, justificado pela maior variância dos dados. A imagem diferença entre o EVI para a data de 1992 (EVI_1992) e o EVI estimado para o mesmo período (EVI_est_1992) destaca os resíduos da análise de regressão com o modelo completo (Figura 4). Observa-se, de forma geral, um bom ajuste para os dados e uma tendência a superestimação nas áreas onde há presença de nuvens e de algum resquício de névoa. Os locais onde o solo encontrava-se desprovido de vegetação ou preparado para o plantio, também tiveram valores de EVI ampliados na imagem estimada pelo modelo de regressão.

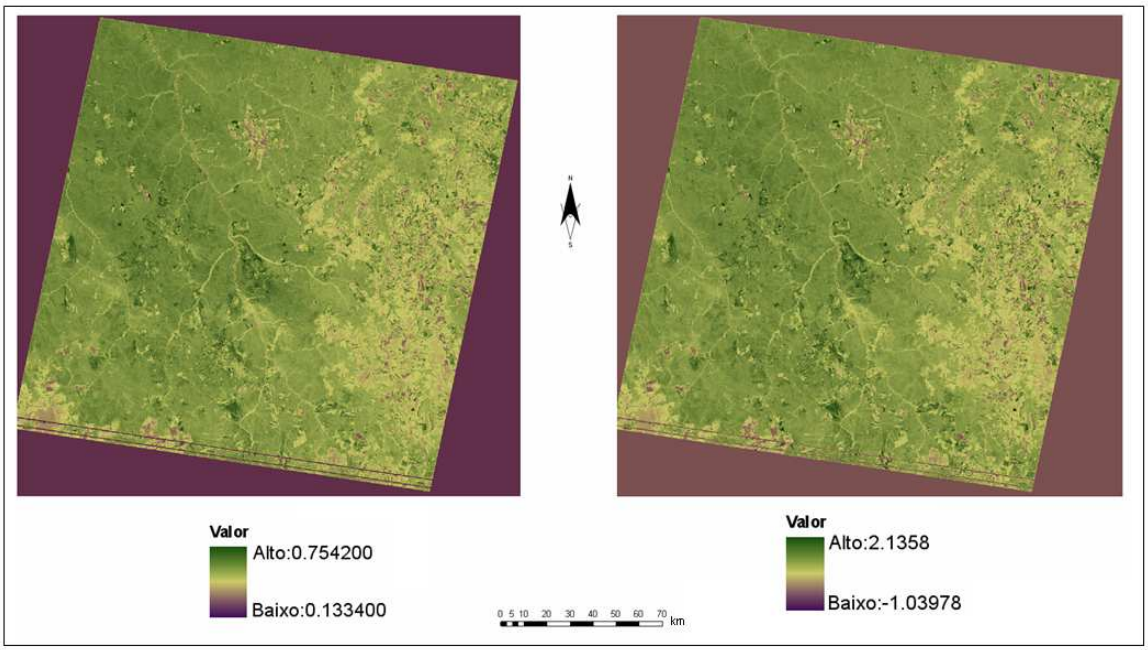

Figura 3. a) Índice de Vegetação Melhorado para 1992 (EVI_1992); b) Índice de Vegetação Melhorado Estimado para 1992 (EVI_Est_1992).

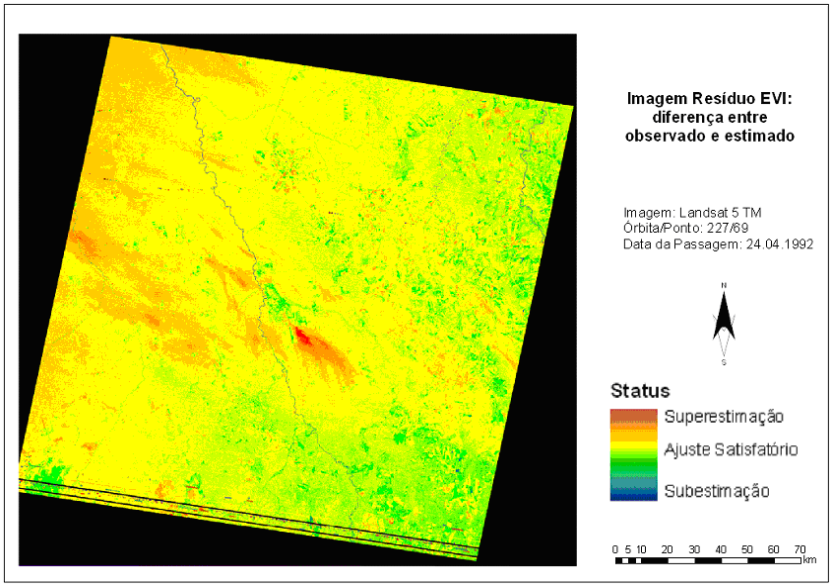

Figura 4. Resíduos da Análise de Regressão: Imagem diferença entre EVI_1992 e EVI_Est_1992.

De maneira análoga, as regiões que normalmente apresentam baixos valores de EVI, como corpos de água, áreas de vegetação pioneira de influência fluvial e hidro-seres associadas, tiveram um relativo decréscimo em seus índices. Porém, estes ambientes ocorrem em proporções muito baixas na imagem, sempre associadas a cursos de água. O mapa de EVI estimado para a imagem adquirida em 1986 a partir do modelo de regressão linear é exibido na Figura 5.

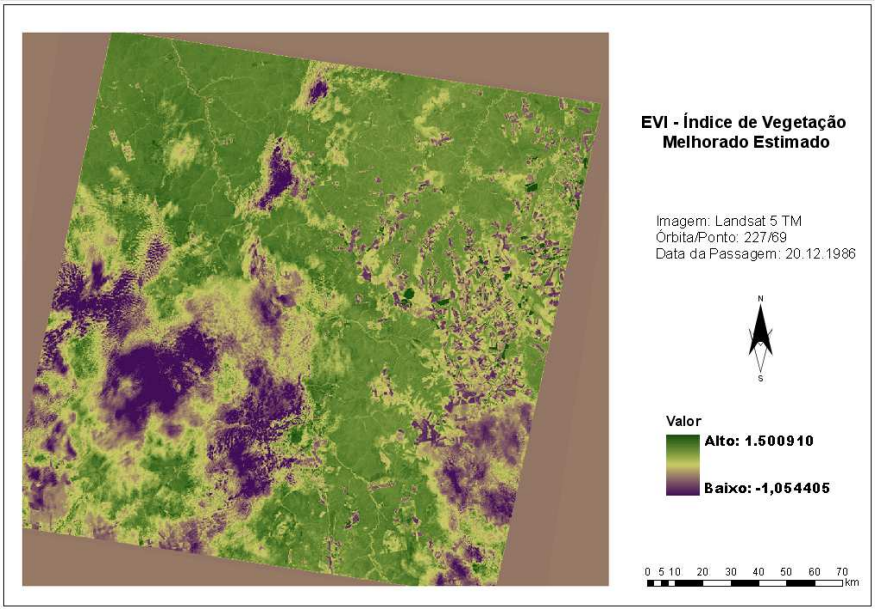

Figura 5. EVI estimado para a imagem adquirida em 1986.

4. Conclusões e considerações finais

Quando o EVI do ano de 1992 foi comparado com o seu correspondente estimado para a mesma data, o ajuste observado foi bastante satisfatório, o que pode ser comprovado na análise dos resíduos e da própria imagem resíduo. Os locais de subestimação e superestimação situaram-se nas regiões onde o EVI comumente apresenta valores altos, como nas áreas de vegetação arbórea densa e nuvens, enquanto os baixos, nas regiões de solo exposto e água. O EVI estimado apenas potencializa os

valores máximos e mínimos do índice, praticamente não alterando significativamente no contexto global, o que permite sua aplicação. Quando o modelo baseado nos dados de 1992 foi aplicado para validação sobre os dados de 1994, pode-se perceber o efeito da temporalidade (diferença entre anos) sazonalidade (diferença entre estações do ano), e do uso das terras no ajuste do modelo. Este viés não foi observado quando um modelo foi concebido a partir dos dados de 1994 e aplicados para aquele ano.

O índice de vegetação melhorado (EVI) estimado para a data de 1986 foi obtido de forma bastante rápida, sem problemas de implementação. Porém, o viés encontrado na estimação dos dados de 1994 sugere cautela. A estimação de índices de vegetação a partir de bandas multiespectrais só se justifica quando da impossibilidade de obtenção dos conjuntos de dados orbitais originais e completos.

Os resultados de validação mostram que o modelo de regressão selecionado é apropriado apenas para a estimativa do EVI na imagem obtida em 1992, não sendo adequado para análises multitemporais da cobertura vegetal, pelos fatores já mencionados. As estimativas de índices de vegetação a partir de bandas multiespectrais só se justifica quando da impossibilidade de obtenção dos conjuntos de dados orbitais originais e completos, como no caso deste trabalho. Possivelmente, a estimação do EVI para o ano de 1986 seria bem melhor sucedida se o conjunto de dados preditores fosse de um ano antes ou depois e no mesmo período do ano (mês ou estação), para que mudanças de uso e cobertura do solo, calendário agrícola e influências sazonais não influenciem o modelo.

Referências

Chavez., P.S. Jr. An improved dark-object subtraction technique for atmospheric scattering correction of multiespectral data. **Remote Sensing of Environment**, v. 24, n. 9, p. 459-479, set. 1988.

Huete, A.R. Didan, K., Miura, T.; Rodriguez, E. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v.83, v.1-2, p.195-213, 2002.

Huete, A.R.; Justice, C.; Liu, H. Development of vegetation and soil indices for MODIS EOS. **Remote Sensing of Environment**, v. 49, n. 3, p. 224-234, 1994.

NASA. Landsat Program. **Landsat TM scene p227r069_7t20010730**, Sioux Falls: USGS, 2003. Disponível em: < http://glcf.umiacs.umd.edu/data/landsat/>. Acesso em 15.ago.2007.

Neter, J.; Kutner,N.H.; Nachtssheim, C.J.; Wasserman, W. **Applied Linear Statistical Models**. 4 ed. Boston: McGraw Hill, 1996.

Souza, A.M.; Jacobi, L.F.; Pereira, J.E. **Gráficos de controle de regressão usando o *Statistica***. Florianópolis: Visual Books, 2005. 112 p.