Anais XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, Brasil, 16-21 abril 2005, INPE, p. 891-898.

Análise de regressão linear múltipla para simulação da banda do SWIR com outras bandas espectrais

Natalia de Almeida Crusco[1]
Camila Souza dos Anjos[1]
Corina da Costa Freitas[1]
Camilo Daleles Renno[1]
José Carlos Neves Epiphanio[1]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12201-970 - São José dos Campos - SP, Brasil
{natalia, csa, epiphani}@dsr.inpe.br
{corina, camilo}@dpi.inpe.br

Abstract. The main objective of this paper is the simulation of a spectral band in the shortwave infrared wavelenght through the metodology of multiple linear regression analysis. It was done all the steps for a complete analysis, and it was chosen the better moldel that fits the premises. The dependent variable was the shortwave infrared band and the independent variables were the bands 1, 2, 3 and 4 of the Landsat satellite. All this variables represents the digital number on the image chosen. The final model contemplates only the values of band 4 and the logaritmic values of band 3. It was done the simulation on CBERS, but the equation have to pass through a calibration in relation to equation gotten for the Landsat. This calibration was not made on this analysis.

Palavras-chave: SWIR, CBERS, regressão múltipla, simulação.

1. Introdução

Simulações com vistas a diversas aplicações têm sido realizadas na área de sensoriamento remoto. Freire e Bastos (1998) simularam a reflectância orbital de água e vegetação no sertão nordestino considerando a influência de fatores atmosféricos, com a finalidade de fazer inferência sobre as propriedades da superfície. A simulação da banda pancromática do sensor ETM+ através da combinação linear das bandas multiespectrais do sensor ETM foi realizada com sucesso e pode ser utilizada nos satélites Landsat-5 e Spot (Boggione et al., 2003). Neste mesmo trabalho, os autores afirmam que a simulação de uma ou mais bandas de um sensor multiespectral pela combinação de outras bandas é uma ferramenta útil para o processamento de imagens.

A banda SWIR (shortwave infrared 1,55-1,75μm) guarda relação com o conteúdo de água na estrutura celular das folhas. Este tipo de informação é importante em análises de comunidades vegetais, principalmente quando relacionadas ao seu desenvolvimento fenológico e fisiológico.

Os sensores TM (Thematic Mapper) e ETM (Enhanced Thematic Mapper) do programa Landsat proporcionaram substancial aumento de informação quando as configurações espectrais passaram a incorporar a região espectral do infravermelho de ondas curtas (1,55-1,75μm). A relação especial dessa banda com o teor de umidade tem sido relevante na distinção entre vegetações e avaliação do estresse hídrico (Gao, 1996). A banda no infravermelho de ondas curtas tem por característica responder à diferença de umidade que ocorre entre espécies de vegetações como reflexo da disponibilidade hídrica do ambiente (Moran et al., 1997). É importante considerar que o comportamento espectral da água não é o que define a resposta na banda do infravermelho de ondas curtas, mas sim seu arranjo nos diferentes tipos de tecidos vegetais, o que torna essa resposta função do tipo de vegetação (Knipling, 1970).

O sensor CCD (Câmera Imageadora de Alta Resolução) presente no satélite CBERS (Satélite Sino Brasileiro de Recursos Terrestres) não contempla esta banda do SWIR. Assim, surgiu a necessidade de realizar este trabalho, que tem como principal objetivo a simulação das respostas dos alvos a esta faixa de comprimento de onda a partir das bandas existentes: visível e infravermelho próximo.

2. Materiais e Métodos

O sensor TM do satélite Landsat-5 apresenta algumas bandas com intervalos de comprimento de onda muito similares às bandas presentes no CCD/CBERS, como apresentado na **Tabela 1**.

Tabela 1: Bandas espectrais do sensor TM (Thematic Mapper) do Landsat e do CCD (Câmera de Alta Resolução) do CBERS-2.

	TM	CCD
Banda 1	0,45 – 0,52 µm	0,45 – 0,52 µm
Banda 2	0,52 – 0,60 µm	0,52 – 0,59 µm
Banda 3	0,63 – 0,69 µm	0,63 – 0,69 µm
Banda 4	0,76 – 0,90 µm	0,77 – 0,89 µm
Banda 5	1,55 – 1,75 µm (SWIR)	0,51 – 0,73 µm (pancromática)
Banda 7	2,08 – 2,35 µm	
Resolução espacial	30 metros	20 metros

Como evidenciado na **Tabela 1**, as bandas espectrais são muito similares nos dois sensores. A maior diferença é a ausência de uma banda na faixa do infravermelho de ondas curtas na câmera CCD/CBERS.

Utilizou-se uma imagem do satélite Landsat-5, órbita/ponto 220/76, referente à região de Piracicaba datada de 22/04/2004, que é uma cena que apresenta grande heterogeneidade de alvos.

Foi feita uma análise de correlação entre os níveis de cinza de todas as bandas do TM, a fim de analisar se existia alguma relação entre tais bandas com vistas a substituir a resposta espectral presente na banda 5. Para evitar a alta correlação entre os pixels adjacentes na imagem foi aplicado um algoritmo, desenvolvido no software ENVI/IDL versão 3.5, para um processamento de decorrelação entre os pixels que seriam utilizados. Os parâmetros utilizados foram 100 pixels para coluna e 100 pixels para linha, ou seja, da imagem original, a cada 100 pixels de coluna e de linha um seria amostrado.

Todas as variáveis utilizadas neste trabalho, tanto a variável dependente quanto as variáveis preditivas, representam os valores de nível de cinza presentes nas bandas em uma imagem do satélite Landsat. A variável dependente (Y) corresponde ao nível de cinza referente à banda 5 (SWIR), e as variáveis explicativas representam os valores de nível de cinza presentes nas outras bandas (1, 2, 3, 4 e 7). Os índices de vegetação NDVI (Normalized Difference Vegetation Index) e o NDWI (Normalized Difference Water Index) foram inicialmente adicionados como variáveis explicativas.

Como o objetivo final do trabalho era a simulação da resposta do SWIR para o sensor CCD/CBERS, optou-se pela retirada das seguintes variáveis: B7, NDVI, NDWI. As variáveis B7 e NDWI foram retiradas, pois não apresentam correspondentes espectrais no CBERS. Já o NDVI foi retirado da análise, pois a informação contida neste índice está presente nas bandas do vermelho e do infravermelho próximo. Assim, restaram quatro variáveis independentes. A partir destas foi feita a análise exploratória para a redução de variáveis

3. Resultados e Discussão

O primeiro passo da análise foi a obtenção de uma matriz gráfica que relaciona a variável dependente com todas as variáveis independentes. Uma análise gráfica prévia sugeriu que alguma transformação deveria ser aplicada em algumas das variáveis presentes no modelo. Pela análise individual da variável dependente (banda 5) com cada variável independente, optou-se pela transformação logarítmica, que diminuiu a multicolinearidade entre as variáveis (**Tabela 2**).

Tabela 2. Matriz de correlação após as transformações

	B4	**log B1**	**log B2**	**log B3**	**B5**
B4	1,0000	0,4061	0,4809	0,3279	**0,7654**
log B1	0,4061	1,0000	**0,9604**	**0,8935**	0,5296
log B2	0,4809	0,9604	1,0000	**0,9399**	0,6522
log B3	0,3279	0,8935	0,9399	1,0000	0,6216
B5	0,7654	0,5296	0,6522	0,6216	1,0000

Pela transformação logarítmica das variáveis dependentes, todos os R^2 tiveram seus valores aumentados, exceto a variável B4. Assim, optou-se por utilizar nas análises futuras os valores logaritmizados das variáveis referentes às bandas 1, 2 e 3, e, para a banda 4, utilizaram-se os valores sem a transformação.

Como o número de variáveis é pequeno, foram realizadas todas as regressões possíveis para a análise de retirada das variáveis. A princípio foi executado o algoritmo de Best Subsets presente no software Minitab versão 13.0, para a verificação prévia de quais seriam as variáveis com maior probabilidade de serem retiradas.

Para a confirmação dos resultados obtidos no método Best Subsets, foram analisados os valores de R^2 obtidos nos modelos de regressão gerados a partir da combinação de todas as variáveis e também pela análise do Cp (Neter et al., 1996). Pela análise do Cp, o modelo com todas as variáveis apresenta o menor valor e o único valor aceitável. Isso ocorre quando Cp é próximo de p, onde p é o número de parâmetros, ou seja, número de variáveis mais 1 (neste caso, p=5). Mas a alta multicolinearidade entre as variáveis logB1, logB2 e logB3 limita a utilidade do modelo.

Na sequência da análise, observou-se que o maior valor de R^2 referia-se ao modelo que continha quatro variáveis. Entretanto, as bandas 1, 2 e 3 apresentam altos valores de correlação, como mostra a **Tabela 2**. Da mesma forma ocorre com os modelos que englobam, respectivamente, as variáveis B1, B3 e B4; e B2, B3 e B4, onde as bandas 1 e 3, e 2 e 3 apresentam alta correlação. Assim, optou-se por gerar um modelo que contivesse apenas a informação das bandas 3 e 4. Tal modelo minimiza a correlação entre as bandas e mantém o valor de R^2 aceitável (0,74). Mesmo apresentando um R^2 ajustado menor (0,69), o modelo de regressão contendo as variáveis logB2 e B4 também foi considerado, já que a primeira variável apresenta maior correlação com Y.

Para a verificação de qual modelo melhor se ajustaria aos propósitos de simulação da banda 5, optou-se pela análise de três modelos. O primeiro é aquele que contém todas as variáveis, o segundo apenas com as variáveis logB3 e B4, que apresentou maior valor de R^2, sem conter as variáveis dependentes correlacionadas, e o terceiro contendo as variáveis logB2 e B4, sabendo que a variável representando a banda 2 apresenta a maior correlação com a variável dependente entre as demais variáveis.

3.1. Análise dos Resíduos

A análise dos resíduos permite a visualização de quais conjuntos de variáveis apresentam maior ajuste ao modelo de regressão. Os resíduos devem apresentar as seguintes propriedades: linearidade, normalidade e variância constante.

A linearidade dos modelos pode ser avaliada através do teste "lack of fit" (Neter et al., 1996), pelo valor do teste F. Neste teste de falta de ajuste, todos os modelos analisados apresentaram linearidade, ou seja, os valores de F calculados foram maiores que os valores de F tabelados.

O gráfico de "Normal Probability Plot" dos resíduos, onde cada resíduo é plotado contra o seu valor esperado sob normalidade, apresenta os indícios de normalidade. Além da análise visual do gráfico, um teste formal para avaliação da normalidade dos erros pode ser feito através do coeficiente de correlação (r) do modelo, entre os resíduos e os seus valores esperados sob normalidade. Um alto valor do coeficiente de correlação indica a normalidade dos erros. Esta afirmação é válida para os três modelos analisados, cujos valores de r foram maiores que 0,99.

Para analisar se a variância dos erros era constante, foram realizados alguns testes. Na observação visual dos gráficos dos resíduos com as variáveis preditoras, as variâncias não apresentavam um padrão constante. Para fazer a análise de constância da variância optou-se pelo método modificado de Levene. A primeira tentativa em relação a este teste foi realizada no software Statistica 6.0, mas descartaram-se os resultados, pois a definição da divisão dos grupos a serem analisados não era clara. Assim, optou-se pela execução do teste através do software Microsoft Excel, utilizando-se as fórmulas disponíveis em Neter et al. (1996).

Com os resultados obtidos pelo teste modificado de Levene, observou-se que apenas o modelo contendo as variáveis logB3 e B4 (modelo 2) apresenta a variância dos resíduos constante para as mesmas. Em relação ao modelo com todas as variáveis, estes resultados só vieram a reforçar a necessidade da retirada de variáveis altamente correlacionadas, que geram problemas de multicolinearidade. Assim, para a etapa de validação, foi utilizado apenas o modelo 2, que apresenta todas as características desejadas para tal.

3.2. Outliers

Vários métodos são utilizados para a verificação da existência e da influência dos "outliers" na reta de regressão ajustada. Neste trabalho tal análise foi realizada pela observação dos valores de DFFITs e Distância de Cook.

Em relação ao método DFFITs, Neter et al.(1996) afirmam que, quando as amostras são grandes, como é o caso (4158 amostras), um ponto pode ser considerado influente se o DFFIT exceder $2*\sqrt{p/n}$. Para os dados utilizados o valor calculado foi de: $2*\sqrt{3/4158}$ = 0,053722. Como havia valores muito próximos a este limite, analisaram-se somente os casos que apresentavam os valores mais altos de DFFIT (acima de 0,09).

Para os valores de DFFITs calculados, 38 casos foram considerados influentes de acordo com este método. Para a confirmação da influência destes valores no modelo de regressão, foi realizado o método da Distância de Cook. Este método avalia a influência do "iésimo" caso sobre os demais casos, analisando a diferença entre os outros casos. É diferente do método DFFIT, que avalia a influência de cada caso separadamente. O valor da distância de Cook (D) deve ser analisado em função da probabilidade deste na função de distribuição F, correspondendo ao valor do percentil em que se encontra.

Através da distribuição F, têm-se as seguintes suposições: se o valor do percentil for menor do que 10 ou 20% do iésimo caso, este não apresenta influência sobre os valores ajustados. Se o valor do percentil estiver próximo a 50% ou mais, os valores ajustados com estes casos diferem substancialmente, ou seja, estes valores são considerados *outliers*.

Obtidos os resultados da aplicação do método da distância de Cook, foi verificado em quais casos este valor ultrapassava 30%. Nenhum dos casos presentes no modelo ficou acima de 10% na distribuição F. Assim, constatou-se que não existem *outliers* que influenciem o ajuste dos valores da regressão.

3.3. Características do modelo selecionado

Após todas as análises realizadas, o modelo selecionado contém duas variáveis independentes: o logaritmo dos valores da banda 3 e os valores de nível de cinza da banda 4, com um R^2 ajustado de 0,7397. O gráfico dos valores observados com os valores ajustados (**Figura 1**) mostra que existe uma relação linear entre os valores.

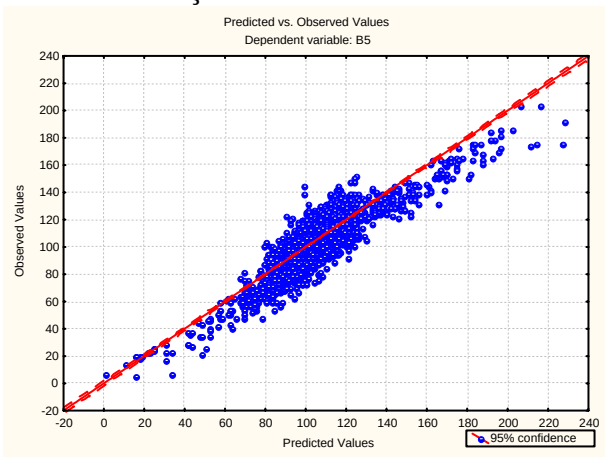


Figura 1. Gráfico dos valores observados e valores ajustados.

A equação de regressão do modelo selecionado é:

$$Y = -122{,}15 + 0{,}963B4 + 84{,}372 \log B3$$

Pela matriz de correlação do modelo final (**Tabela 3**), observa-se que não existe alta correlação entre as variáveis independentes e que estas apresentam, individualmente, uma correlação considerável com a variável dependente.

Tabela 3. Matriz de correlação do modelo final.

	B4	**logB3**	**B5**
B4	1,0000	0,3279	0,7654
logB3	0,3279	1,0000	0,6216
B5	0,7654	0,6216	1,0000

Os fatores analisados no modelo, como normalidade dos resíduos, homocedasticidade da variância, aleatoriedade dos resíduos, análise da interação e presença de outliers, indicam ser este o melhor ajuste entre todas as possibilidades encontradas para a geração de um modelo de regressão linear eficiente para simular a banda 5 a partir de outras bandas do TM.

3.4. Validação do modelo

Existem vários métodos para examinar a validade do modelo de regressão. O método utilizado neste estudo avalia a capacidade preditiva do modelo selecionado. Calcula-se o quadrado médio do erro para os valores da validação (MSPR) e se compara com o valor do quadrado médio do erro (MSE) do modelo ajustado.

$$MSPR = \left[\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2 \right] / n^*$$

onde,

Y_i é o valor da variável dependente no iésimo caso da validação

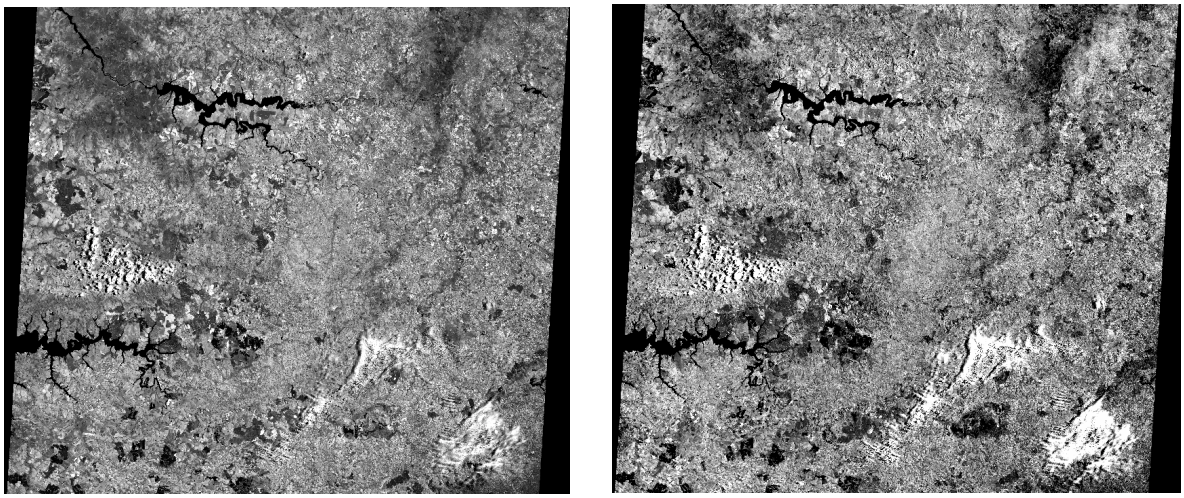
$\hat{Y}_i$ é o valor ajustado para o iésimo caso da validação baseado no modelo selecionado

n^* é o número de casos da validação

O valor do MSPR encontrado foi de 90,32804. Se o valor do MSPR encontrado for próximo do valor do MSE (Mean Square Error) do modelo de regressão construído com os dados iniciais, então este MSE do modelo selecionado não é seriamente tendencioso e indica uma boa habilidade de predição do modelo. Como o valor do MSE encontrado no modelo foi de 95,6, conclui-se que o modelo selecionado está bem adequado quanto aos parâmetros analisados.

3.5. Aplicação da equação de regressão sobre a imagem

Após a validação do modelo, foi realizada a simulação da banda 5 através das bandas selecionadas (bandas 3 e 4) pela aplicação da equação B5 = -122,15 + 0,963B4 + 84,372logB3. Na imagem com órbita/ponto 220/76, a mesma utilizada para a extração dos níveis de cinza referentes às variáveis do modelo, foi realizada as operações indicadas pela equação acima. O resultado da aplicação na imagem é mostrado na **Figura 2**, em que a primeira é a imagem da banda 5 original e a segunda é a imagem da banda 5 simulada.



(a) (b)

Figura 2.(a)imagem original da banda 5 do TM/Landsat-5 e (b)imagem TM/Landsat-5 simulada pelo modelo.

Em uma avaliação visual da imagem simulada, observou-se que alvos com resposta espectral alta e baixa, como as nuvens e a água, apresentaram comportamentos semelhantes aos presentes na imagem original. As culturas agrícolas apresentaram maior contraste na imagem simulada, mas estas variações são pequenas quando comparadas à imagem original. Pela análise do resultado obtido a simulação apresentou um bom resultado e esta abordagem mostrou-se factível para o sensor TM.

Para a simulação da banda do SWIR no satélite CBERS, foi necessária a realização de um mosaico de quatro imagens. Procuraram-se imagens que apresentassem um menor intervalo temporal possível em relação à imagem do sensor TM utilizada. Estas imagens foram

adquiridas em 24/07/2004 para a órbita 155 pontos 125 e 126, e em 16/08/2004 para a órbita 156 pontos 125 e 126. Após mosaicadas, as imagens CBERS foram recortadas de forma que englobassem apenas a área contemplada na imagem Landsat.

A equação obtida para a imagem TM foi aplicada para a imagem CBERS (**Figura 3**). O resultado obtido apresentou bons resultados para alvos com baixa resposta espectral, já o mesmo não pode ser afirmado para alvos com altas respostas, como nuvens, por exemplo. Pelo fato de a aquisição das imagens mosaicadas serem de diferentes datas, acredita-se que uma correção radiométrica entre as imagens deve ser realizada, afim de evitar a diferença de tonalidade entre elas. De forma geral, os alvos escuros responderam melhor à simulação quando comparados aos alvos claros, como pode ser observado na parte inferior da imagem CBERS simulada. No entanto ficou claro que correções devem ser realizadas. Uma das transformações está relacionada à escala de valores representadas pelos níveis de cinza. Ou seja, um certo valor de nível de cinza da banda 3 do sensor TM não necessariamente representa o mesmo valor de nível de cinza da banda 3 do sensor CCD. Para a correção deste problema, poderia ser feita uma regressão entre estas bandas e, a partir daí, aplicá-la à equação do modelo.

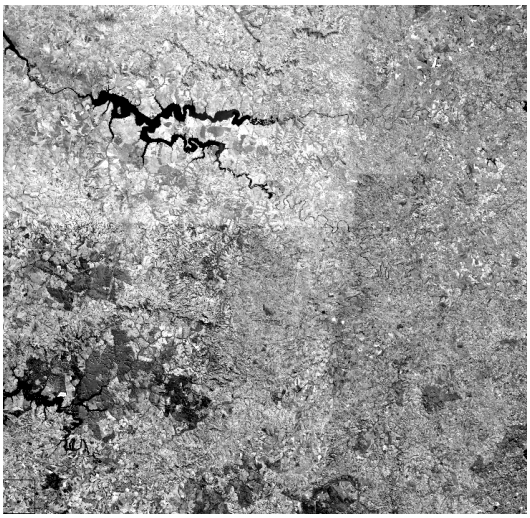


Figura 3. Imagem SWIR simulada com as bandas do sensor CCD

Para a obtenção de uma equação que possa ser utilizada em qualquer imagem, visando à simulação da banda SWIR em outro sensor, é necessário que sejam feitas correções radiométricas nas imagens, ou seja, é indicada a transformação dos valores de nível de cinza em radiância ou reflectância. Assim, haverá correspondência entre as grandezas físicas representadas no TM e no CCD. Embora a simulação da CCD/CBERS realizada neste trabalho tenha sido gerada com valores de nível de cinza, o objetivo principal foi o de verificar a possibilidade da simulação de uma banda espectral SWIR e de indicar um roteiro metodológico para tal simulação.

4. Conclusão

As análises iniciais contendo todas as variáveis selecionadas mostrou que algumas variáveis necessitavam de transformação, assim como a presença de outras variáveis não faziam sentido para a execução de um modelo de regressão. Após a retirada e as transformações de algumas variáveis, verificou-se a existência de relação linear entre as variáveis modificadas.

A validação do modelo mostra que todas as técnicas utilizadas foram úteis na sua construção, assim como também na simulação da banda 5 através de outras bandas presentes no satélite Landsat. A imagem gerada pelo modelo guardou muita semelhança com a imagem

original. É evidente que a imagem simulada precisa de ajustes e correções, e estes poderão ser conseguidos com maior rigor analítico e com um maior número de parâmetros. A correção radiométrica e a obtenção dos coeficientes com base nos valores de radiância ou reflectância resultaria em uma "equação mais universal" que poderia ser usada numa maior gama de situações. As variáveis para a geração de um modelo de regressão linear múltipla devem representar os valores de reflectância, já que assim, as imagens obtidas pelo sensor TM e pelo sensor CCD se tornam equivalentes.

A simulação de uma suposta banda referente ao SWIR para a câmera CCD/CBERS a partir de um modelo derivado de bandas do TM mostrou-se factível. Mas a avaliação do resultado não é preciso, pois os dados utilizados não passaram por uma correção radiométrica e não há dados que possibilitem a comparação com outra imagem semelhante, tanto na questão espacial, quanto na questão espectral. Tais aspectos deverão ser levados em conta num futuro trabalho.

Trabalhos futuros podem ser realizados, tanto na geração de um modelo que compreenda as imagens corrigidas, quanto na avaliação destas imagens, levando em consideração diferentes tipos de alvos, épocas do ano, latitudes, etc.

5. Referências

Boggione, G.A.; Pires, E.G.; Santos, P.A.; Fonseca, L.M.G. Simulation of a panchromatic band by spectral combination of multispectral ETM+ bands. In: International Symposium on Remote Sensing of Environmental, 30. (ISRSE), Nov. 2003, Hawai, HA. **Proceedings**... Honolulu, 2003. Papel, On-line. Publicado como: INPE-10573-PRE/6036.

ENVI 3.5, West Lafayette, Colorado: RSI. 2000.

Freire, M.L.F.; Bastos, E.J.B. Simulação da reflectância espectral planetária de alvos. **Revista Brasileira de Geofísica**, v.16, n.2-3, p.181-190, 1998.

Gao, B. A normalized difference water index for remote sensing of vegetation liquid water from space. **Remote Sensing of Environment**, v.58, p.257-266, 1996.

Knipling, E.B. Physical and physiological basis for the reflectance of visible and near infrared radiation from vegetation. **Remote Sensing of Environment**, v.1, n.3, p.155-159, 1970.

MINITAB 13.0, State College, PA: Minitab Inc. 2000.

Moran, M.S.; Inoue, Y.; Barnes, E.M. Opportunities and limitations for image-based remote sensing in precision crop management. **Remote Sensing of Environment**, v.61, p. 319-346, 1997.

Neter, J.; Kutner, M.H.; Nachtssheim, C.J., Wasserman, W. **Applied linear statistical models**. Boston: WCB/McGraw-Hill, 4th edition, 1996.

Statistica 6.0, Tulsa, OK: Statsoft. 2001. 1 CD-ROM.