

AVALIAÇÃO DE DOIS MÉTODOS DE AQUISIÇÃO DE ESTATÍSTICAS DE TREINAMENTO PARA
A ANÁLISE DIGITAL DE ÁREAS DE SOLO PREPARADO PARA PLANTIO

Getúlio Vargas de Assunção
Antonio Roberto Formaggio
Maurício Alves Moreira

Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq
Instituto de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12200 - São José dos Campos - SP - Brasil

RESUMO

O presente trabalho objetivou estudar dois métodos de obtenção de estatísticas de treinamento (vetor média e matriz de covariância) para a análise digital de áreas com solo preparado para plantio, em Latossolo Roxo. Para a realização desta pesquisa selecionou-se uma área de estudo de aproximadamente 400 km², localizada na Divisão Regional Agrícola de Ribeirão Preto, a nordeste do Estado de São Paulo. A metodologia empregada consistiu em uma análise digital onde foram utilizadas duas abordagens para a aquisição das estatísticas de treinamento. Na primeira foram obtidas áreas de treinamento sobre talhões conhecidos, representativas dos alvos de interesse (método supervisionado), para ser utilizadas no classificador de máxima verossimilhança (MAXVER). Na segunda abordagem foi utilizado um classificador não-supervisionado (K-Médias) para o agrupamento dos "pixels" de respostas espectrais homogêneas. Após o agrupamento, foi feita a identificação dos tipos de cobertura correspondentes às várias classes espectrais obtidas (classes informativas). Com base nos resultados do agrupamento, foram adquiridas áreas de treinamento contendo "pixels" puros ("pixels" sem influência de bordas dos talhões) para ser utilizadas no MAXVER. A avaliação do desempenho dos dois métodos foi feita através do teste t ($\alpha = 0.005$) para os fatores classificação correta (CC) e erro de inclusão (EI), obtidos de 8 segmentos selecionados aleatoriamente sobre a área de estudo com diferentes densidades de solo preparado.

ABSTRACT

The present work aimed at the study of two methods to obtain training statistics (mean vector and covariance matrix) to the digital analysis of areas featuring plowed land areas in Latosolic B (Terra Roxa). For the realization of this research a study area of approximately 400 km² localized in the Regional Agricultural Division of Ribeirão Preto, at the northeastern part of São Paulo State was selected. The applied methodology consisted of a digital analysis in which two approaches for the training

statistics acquisition were utilized. In the first one, training areas were obtained on known land lots which were representatives of interest targets (supervised method), to be input to a maximum likelihood classifier (MAXVER). In the second approach an unsupervised classifier (K-means) was utilized to group pixels with homogeneous spectral signatures. After the grouping an identification of the land cover types which corresponded to the obtained spectral classes (information classes) was accomplished. Basing on the results of the grouping, training areas with pure pixels (pixels without influence from the border of the land lot) were acquired to be utilized in the MAXVER. The evaluation of the performance of the two methods was made through the application of the t - test ($\alpha = 0.05$) to the factors of correct classification (CC) and commission error (EI) obtained from the 8 randomly scene segments of the study area with different densities of plowed land.

1. INTRODUÇÃO

A avaliação da quantidade de área ocupada com determinada cultura constitui um dos principais parâmetros envolvidos no cálculo das estimativas na previsão de safras. Esta área é avaliada, na maioria das vezes, por métodos convencionais, através de entrevistas junto a pessoas ligadas diretamente ao campo ou através de levantamentos aerofotográficos que fornecem uma boa aproximação da realidade.

No entanto, ambos os métodos apresentam desvantagens: o primeiro por sua alta subjetividade e o segundo por seu custo elevado.

As metodologias de utilização de dados do satélite LANDSAT surgiram então como uma forma de minimizar o custo e a subjetividade alcançados pelos métodos utilizados. Porém, um dos problemas encontrados na implantação destas metodologias é o fato de que o ciclo das principais culturas anuais coincide com o período de maior precipitação pluviométrica, restringindo assim a obtenção de imagens isentas de cobertura de nuvens.

Visando contornar este problema, Assunção e Duarte (1983) estabeleceram uma metodologia de obtenção de áreas de solo preparado para plantio, através de técnicas de sensoriamento remoto. Segundo estes autores, o conhecimento da quantidade de área de solo preparado para plantio é muito importante, sendo uma informação adicional que pode ser associada a informações de intenção de plantio para o fornecimento de uma primeira indicação das safras esperadas num determinado ano agrícola.

Para alvos como as áreas preparadas para plantio, que são encontradas em todo o território nacional e sujeitas aos mais variados tipos de solo, relevo, clima, etc, torna-se necessário o estabelecimento de metodologias para sua identificação para cada condição sobre áreas pilotos representativas. Posteriormente estas metodologias poderão ser extrapoladas para grandes áreas.

Dentre os métodos existentes para obter informações sobre áreas agrícolas, Hixson et alii (1981) comentam que a visão sinótica da Terra, fornecida por satélites de observação, como o LANDSAT, junto com o processamento digital dos dados, é muito útil para identificar e estimar áreas ocupadas com culturas.

Em se tratando de análise digital para o levantamento de áreas agrícolas, o procedimento de treinamento no computador é mais importante que a escolha do classificador para obter uma classificação satisfatória da cultura (Hixson et alii, 1980).

O desempenho da classificação, segundo Wharton e Turner (1981), Hoffer (1973) e Lima et alii (1982), parece ser mais satisfatório quando a definição dos parâmetros de cada classe isto é, a obtenção das estatísticas de treinamento, é obtida através da combinação de dois sistemas: o supervisionado e o não-supervisionado.

O presente trabalho tem por objetivo avaliar dois métodos de obtenção de estatísticas de treinamento para a análise digital de áreas com solo preparado para plantio.

2. MATERIAL

2.1 - ÁREA DE ESTUDO

A área de estudo localiza-se a nordeste do Estado de São Paulo (Figura 1), na Divisão Regional Agrícola de Ribeirão Preto, e é caracterizada por intensa atividade agrícola. Possui aproximadamente 400 km² e está compreendida entre os paralelos 20⁰45' e 21⁰00' de latitude sul e os meridianos 47⁰55' e 48⁰30' de longitude oeste.

Ocorre nesta área uma predominância de Latossolo Roxo (Figura 1) originário de materiais provenientes de rochas básicas, principalmente basaltos e diabásios. Possui um relevo suave ondulado e uma cor vermelha intensa (arroxeada) devido ao teor elevado de óxido de ferro (Fe₂O₃).

2.2 - FOTOGRAFIAS AÉREAS

Foram utilizadas fotografias aéreas infravermelhas coloridas em falsa-cor que recobrem toda a área de estudo, numa escala média de 1:20.000, obtidas em 06/09/79. Estas fotografias foram usadas como dados de referência para identificar os diferentes alvos nos dados do LANDSAT durante a análise digital e avaliar o desempenho dos dois métodos de aquisição de estatísticas de treinamento.

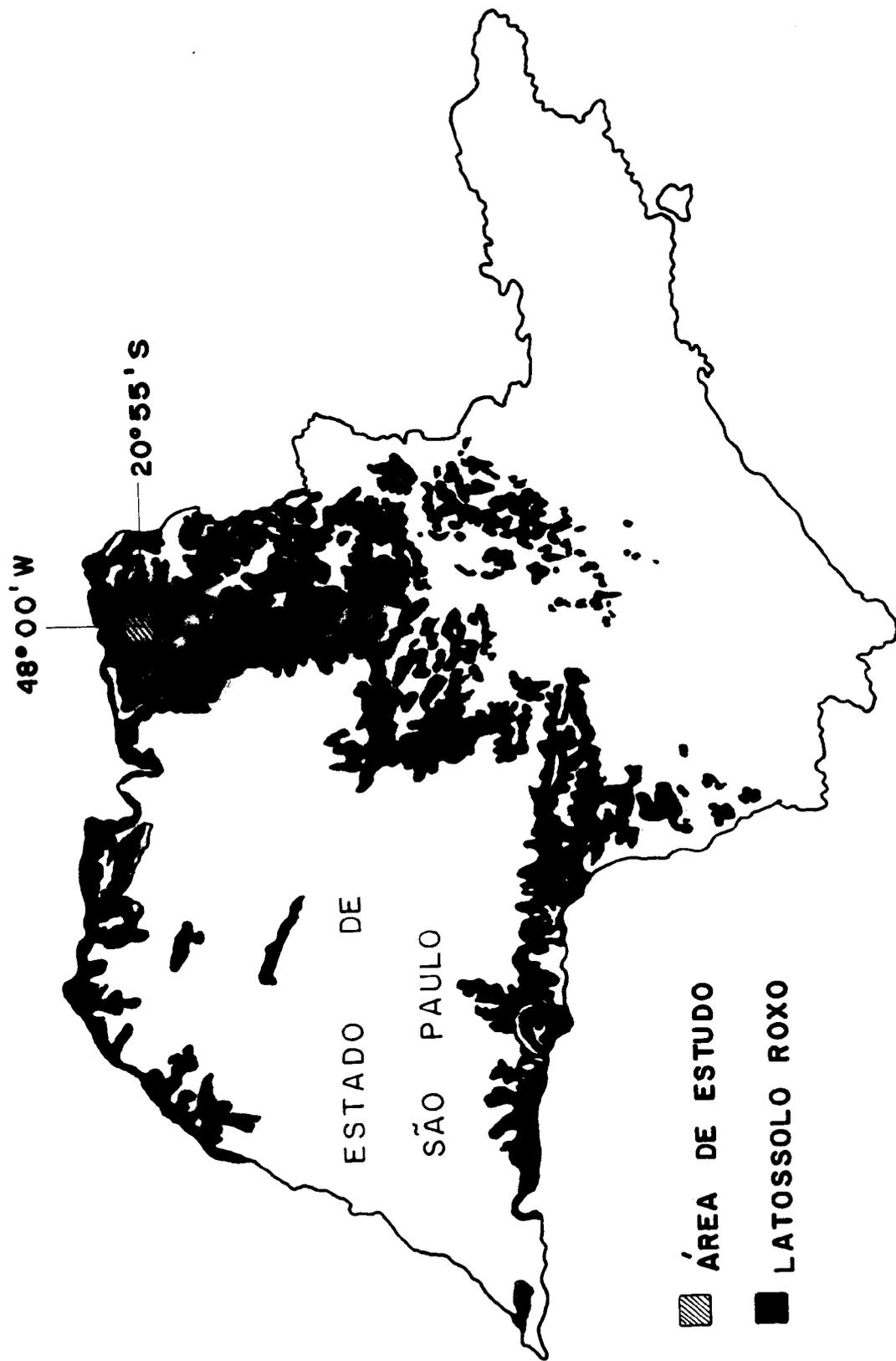


Fig. 1 - Localização da área de estudo.

2.3 - IMAGENS DE SATÉLITE

Foram utilizados, no tratamento automático das áreas preparadas para plantio, os dados digitalizados do sensor MSS do LANDSAT 2, armazenados em fitas compatíveis com computador.

Estes dados correspondem à órbita SRB 178, ponto 27, e foram obtidos em 01/09/79, apresentando assim uma defasagem de cinco dias em relação aos dados obtidos por aeronave.

3. METODOLOGIA

A metodologia foi baseada na análise automática dos dados MSS/LANDSAT da área de estudo, utilizando o Sistema Interativo de Análise de Imagem Multiespectral (Image-100).

A área de estudo foi delimitada e colocada na tela do Image-100 na escala 1:100.000, que é uma escala em que os alvos estudados se apresentam num tamanho adequado para a obtenção das amostras de treinamento.

3.1 - AQUISIÇÃO DAS ESTATÍSTICAS DE TREINAMENTO

As abordagens de aquisição das estatísticas de treinamento utilizadas foram denominadas TREINA 1 e TREINA 2.

3.1.1 - Abordagem TREINA 1 - nesta abordagem as áreas de treinamento necessárias para a obtenção do vetor média e da matriz de covariância (que caracterizam a distribuição de probabilidade das classes conhecidas e são essenciais na regra de decisão do algoritmo a ser utilizado na classificação automática) foram obtidas de maneira supervisionada. Para isso, posicionou-se o cursor do Image-100 sobre as classes de ocupação da área de estudo (solo preparado para plantio, cana nova, cana cortada, cana intermediária, cana adulta, água, áreas com drenagem deficiente, mata e pastagem), tomando como base o mapa temático obtido através da interpretação visual das fotografias aéreas. Após a obtenção das áreas de treinamento, estas foram arquivadas para posterior utilização na classificação de toda a área de estudo.

3.1.2 - Abordagem TREINA 2 - nesta abordagem foi feito, inicialmente, o agrupamento dos "pixels" de resposta espectral homogênea por meio de um algoritmo de classificação não-supervisionada (K-Médias). Após o agrupamento foi feita a classificação de toda a área de estudo em classes homogêneas. Dentro destas classes foram então obtidas as áreas de treinamento contendo "pixels" puros ("pixels" sem influência de bordas, localizados no centro dos talhões). Estas áreas de treinamento foram então arquivadas para serem utilizadas posteriormente na classificação de toda a área de estudo.

3.2 - CLASSIFICAÇÃO AUTOMÁTICA

Utilizou-se, na classificação automática da área de estudo, o algoritmo supervisionado MAXVER que utiliza o critério da máxima verossimilhança.

Para a fase de treinamento do algoritmo (determinação do vetor média e da matriz de covariância de cada uma das classes que compõem a área de estudo), foram utilizados, separadamente, os arquivos das abordagens TREINA 1 e TREINA 2. Na fase de classificação o algoritmo utilizou estas informações e determinou a densidade de probabilidade de cada ponto da área de estudo pertencer a cada uma das classes.

Feitas as classificações, os resultados, usando cada uma das abordagens, foram apresentados na tela do Image-100. Sobre esses resultados foi feita uma filtragem espacial utilizando o método de pós-processamento Uniformização de Temas - opção UNITOT, com limiar 2 e peso 2, conforme Moreira et alii (1982). Os resultados finais da área de solo preparado para plantio, obtidos após a filtragem espacial, foram apresentados na forma de mapas alfanuméricos na escala 1:20.000 compatíveis com a escala das fotos aéreas.

A área de cada um dos mapas alfanuméricos foi dividida em segmentos de 20 x 20 cm (16 km² no terreno). Posteriormente, foi feita uma amostragem aleatória de 8 segmentos, que correspondem a 32% da área de estudo; a localização destes segmentos foi fixada em ambos os mapas alfanuméricos. Sobre cada segmento foi delimitada a área de solo preparado obtida através das fotos aéreas e feita a contagem da área classificada corretamente e do erro de inclusão com a utilização de uma grade de pontos.

Para comparar quantitativamente o desempenho de classificação usando o TREINA 1 e o TREINA 2, aplicou-se um teste estatístico (teste t, $\alpha = 0,05$) sobre os dados de classificação correta (CC)¹ e erro de inclusão (EI)² obtidos de cada segmento amostrado.

4. RESULTADOS E DISCUSSÃO

Os temas interpretados através das fotografias aéreas foram: solo preparado para plantio, cana nova, cana cortada, cana intermediária, cana adulta, pastagem, mata e áreas com drenagem deficiente.

Durante a classificação automática da área de estudo, utilizando separadamente as duas abordagens de treinamento (TREINA 1 e TREINA 2), verificou-se que não foi possível separar as áreas com cana nova e cana cortada daquelas de solo preparado para plantio nos dados obtidos pelo satélite, muito embora pudessem ser facilmente discrimináveis nas fotos aéreas. Este problema já havia sido constatado por Assunção e Duarte (1983); para contorná-lo os temas cana nova e cana cortada foram incluídos ao tema solo preparado para plantio.

$${}^1CC(\%) = \left(\frac{\text{Área de solo preparado classificada corretamente}}{\text{Área total de solo preparado em fotos aéreas}} \right) \times 100$$

$${}^2EI(\%) = \left(\frac{\text{Área de outros alvos, classificada como solo preparado}}{\text{Área total de solo preparado classificada no Image-100}} \right) \times 100$$

A Tabela 1 apresenta a área de solo preparado obtida das fotos aéreas e a densidade de solo preparado, a área classificada corretamente e a área de outros alvos classificada como solo preparado (erro de inclusão) obtidas através dos dados do LANDSAT.

TABELA 1

ÁREA DE SOLO PREPARADO OBTIDA A PARTIR DAS FOTOGRAFIAS AÉREAS, DENSIDADE DE SOLO PREPARADO, ÁREA CLASSIFICADA CORRETAMENTE E ERRO DE INCLUSÃO PARA CADA SEGMENTO ESTUDADO UTILIZANDO AS DUAS ABORDAGENS DE OBTENÇÃO DAS ESTATÍSTICAS DE TREINAMENTO

SEGMENTO	SOLO PREPARADO EM FOTO AÉREA (ha)	DENSIDADE DE SOLO PREPARADO (%)*	ÁREA CLASSIFICADA CORRETAMENTE (ha)		ERRO DE INCLUSÃO (ha)	
			TREINA 1	TREINA 2	TREINA 1	TREINA 2
1	731,20	45,70	629,12	616,64	62,40	38,08
2	892,16	55,76	849,60	847,67	125,12	87,68
3	570,56	35,66	458,88	446,40	64,96	53,76
4	314,56	19,66	270,08	266,24	53,12	27,52
5	1.114,24	69,64	1.057,60	1.034,24	80,96	48,96
6	596,80	37,30	515,84	476,16	101,76	50,24
7	583,04	36,44	545,92	513,60	35,20	17,28
8	806,08	50,38	774,68	745,60	53,76	40,96

(*) - Densidade = $\frac{\text{Área de solo preparado obtida por Foto Aérea (ha)}}{\text{Área total do Segmento (1600 ha)}}$

A densidade de solo preparado variou de 19,66% a 69,64%, mostrando que a área de estudo, apesar de se situar numa região de intensa atividade agrícola, apresentou variações na quantidade de solo preparado por unidade amostrada.

A Tabela 2 mostra as percentagens de classificação correta, do erro de inclusão e da diferença relativa, obtidas ao utilizar separadamente as duas abordagens de treinamento (TREINA 1 e TREINA 2).

TABELA 2

PERCENTAGENS DE CLASSIFICAÇÃO CORRETA, ERRO DE INCLUSÃO E DIFERENÇA RELATIVA PARA OS SEGMENTOS ESTUDADOS UTILIZANDO AS DUAS ABORDAGENS DE OBTENÇÃO DAS ESTATÍSTICAS DE TREINAMENTO

SEGMENTO	CLASSIFICAÇÃO CORRETA (CC%)		ERRO DE INCLUSÃO (EI%)		DIFERENÇA RELATIVA (DR%)	
	TREINA 1	TREINA 2	TREINA 1	TREINA 2	TREINA 1	TREINA 2
1	86,04	84,33	9,02	5,82	- 5,43	- 10,46
2	95,23	95,01	12,84	9,37	+ 9,25	+ 4,84
3	80,43	78,24	12,40	10,75	- 8,19	- 12,34
4	85,86	84,64	16,44	9,37	+ 2,75	- 6,61
5	94,92	92,82	7,11	4,52	+ 2,18	- 2,79
6	86,43	79,78	16,48	9,54	+ 3,48	- 11,80
7	93,63	88,09	6,06	3,25	- 0,33	- 8,95
8	96,23	92,50	6,48	5,21	+ 3,39	- 2,42
\bar{X}	89,85	86,93	10,85	7,23	+ 0,89	- 6,32

Como pode ser visto na Tabela 2, os resultados de classificação correta obtidos pela utilização do TREINA 1 mostraram-se superiores em todos os segmentos, em relação ao TREINA 2; em compensação, os resultados do erro de inclusão obtidos utilizando o TREINA 2 foram melhores (mais baixos) em todos os segmentos, em relação ao TREINA 1.

A classificação correta de alguns segmentos foi prejudicada pela de faseagem de cinco dias entre a obtenção dos dados do satélite e da aeronave. Por ser setembro um mês em que a atividade agrícola é intensa naquela região, algumas áreas apresentaram-se diferentemente nas duas datas de aquisição dos dados. As áreas que foram preparadas para plantio, ou as áreas com cana adulta que foram cortadas neste intervalo, foram identificadas nos fotos aéreas como solo preparado ou cana cortada, ao passo que nos dados do satélite não foram classificadas como pertencentes a este tema. Estas áreas aumentaram o erro de omissão, diminuindo, em consequência, a classificação correta.

4.1 - ANÁLISE ESTATÍSTICA

Os dados de área classificada corretamente e do erro de inclusão da Tabela 1 foram submetidos a uma análise estatística com o objetivo de avaliar o desempenho dos dois métodos de aquisição das estatísticas de treinamento.

Inicialmente foi feito um testes de homogeneidade de variância (Zar, 1974) indicando que, ao nível de 5%, os dados de classificação correta e erro de inclusão dos dois métodos não apresentaram diferenças significativas entre suas variâncias. Posteriormente, foi aplicado sobre estes dados o teste t para pares de observações. Os valores t calculados foram

significativos ao nível de 5% para os dados de classificação correta ($t = 31,80$) e erro de inclusão ($t = 44,61$) das duas abordagens de treinamento. De posse desses resultados e observando a Tabela 2 pode-se inferir que o TREINA 1 teve uma classificação correta superior ao TREINA 2. Por outro lado, os dados de erro de inclusão obtidos ao utilizar o TREINA 2 foram melhores (mais baixos) que os obtidos pelo TREINA 1. Portanto, não se verificou uma superioridade de uma das abordagens; ambas apresentaram resultados médios de classificação correta superiores a 85% e o erro de inclusão girou em torno de 10%.

5. CONCLUSÕES

1. Nos levantamentos de solo preparado para plantio em regiões com intensa atividade agrícola, como é o caso da área de estudo, as informações coletadas no campo ou por aeronave devem ser obtidas na mesma data de aquisição dos dados do satélite. A defasagem na coleta dos dados pode acarretar:

- a) um aumento no erro de inclusão quando os dados de campo e/ou de aeronave forem obtidos depois da data da passagem do satélite,
- b) um decréscimo da classificação correta quando os dados de campo e/ou de aeronave forem obtidos depois da data da passagem do satélite.

2. Os temas que apresentaram maior confusão com as áreas de solo preparado para plantio foram as áreas com pastagem degradada e as de drenagem deficiente com vegetação esparsa.

3. Não foi possível separar as áreas com cana nova e cana cortada daquelas de solo preparado para plantio, utilizando dados do MSS/LANDSAT de uma única data.

4. A diferença relativa da Tabela 2 mostra que, para a maioria dos casos observados, os resultados obtidos pelo TREINA 2 apresentaram uma tendência de subestimar a área de solo preparado, enquanto os obtidos pelo TREINA 1 apresentaram uma tendência para superestimar. Por outro lado, a diferença relativa média obtida a partir das amostras indicam que o TREINA 1 se aproximou mais da área real de solo preparado quando comparado com o TREINA 2.

5. De maneira geral as duas abordagens apresentaram boa capacidade de identificar as áreas de solo preparado para plantio, com resultados médios de classificação correta superiores a 85%, e o erro de inclusão girando em torno de 10%.

6. As áreas de solo preparado para plantio em Latossolo Roxo apresentaram um alto contraste em relação aos alvos vizinhos. Assim sendo, as áreas de treinamento obtidas de maneira supervisionada (TREINA 1) definiram as classes tão bem quanto aquelas obtidas pela TREINA 2. Como esta abordagem utiliza um maior tempo computacional na aquisição das estatísticas de treinamento, recomenda-se então a utilização da abordagem TREINA 1.

REFERÊNCIAS BIBLIOGRÁFICAS

- ASSUNÇÃO, G.V.; DUARTE, V. *Avaliação de áreas preparadas para plantio (solonú) utilizando dados do satélite LANDSAT*. São José dos Campos, INPE, 1983, (INPE-2637-TDL/113).
- HIXSON, M.M.; SCHOLZ, D.; FUHS, N. Evaluation of several schemes for classification of remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 46 (12): 1629-1638, Dec. 1980.
- ; DAVIS, B.J.; BAUER, M.E. Sampling LANDSAT classification for crop area estimation. *Photogrammetric Engineering and Remote Sensing*, 47 (9): 1343-1348, Sept. 1981.
- HOFFER, R.M. ADP of multispectral scanner data for land use mapping. *The Laboratory for Application of Remote Sensing*, West Lafayette, IN, Purdue University, 1973. 27p. (information note).
- LIMA, A.M.de; CHEN, S.C.; MOREIRA, M.A. Avaliação de dois métodos de aquisição de estatísticas de treinamento para classificação automática. *Simpósio de Sensoriamento Remoto*, 2., Brasília, 10-14 maio, 1982.
- MOREIRA, M.A.de; CHEN, S.C.; LIMA A.M.de. Estudos do método uniformização de temas (UNITOT) e análise da correlação entre áreas estimadas utilizando dados do LANDSAT e fotografias aéreas. *Simpósio de Sensoriamento Remoto*, 2., Brasília, 10-14 maio, 1982.
- WHARTON, S.W.; TURNER, B.J. ICAP. An interactive analysis procedure for analyzing remote sensed data. *Remote Sensing of Environment*, 11 (4): 279-293, 1981.
- ZAR, J.H. Two-sample hypotheses. In: *Biostatistical analysis*. Englewood Cliffs, N.J.; Prentice-Hall, 1974. cap. 9, p. 101-120.