

Estratificação espacial utilizando árvores de decisão para estimativa da área de culturas agrícolas de verão com imagens de pré-plantio

Gustavo Felipe Balué Arcoverde¹
José Carlos Neves Epiphânio¹

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

Abstract. The improvement of the estimative of area of summer crops are important for many users. In general, stratifications are applied for these estimates. In this study we evaluate the process of spatial stratification through tree decisions. The stratifications usually to improve the statistical variances associating when coupled to field verification. Pattern recognition of summer crop on pre-planting TM/Landsat images using tree decision was used to generate spatial stratifications. For this purpose, multi-features were extracted from the satellite images in Barretos municipality, São Paulo State, in Brazil. The tree decision was generate setting the minimum instances by leaves in order to reduce the size of the trees. As results, the variance and area estimates using stratifications has not a significant improvement in relation to the area estimates without stratification, based on a Monte Carlo method. This study showed that in small areas for crop areas to be estimated, the size of the defined stratum (a single stratum) should be nearly the half of whole area to be estimated. Otherwise, the tree decision had a good performance to recognize patterns.

Palavras-chave: crop area estimation, remote sensing, spatial stratification, pattern recognition, tree decision.

1. Introdução

O levantamento da área agrícola é uma informação fundamental no planejamento agrícola e no planejamento do território como um todo, seja na questão econômica, agrária, ambiental, ou social. A extensão da área agrícola, além de ser uma componente no cálculo da produção agrícola do território, é uma variável no cálculo de impostos, por exemplo. A amostragem tem sido a maneira mais utilizada pelo IBGE (Instituto Brasileiro de Geografia e Estatística) e CONAB (Companhia Brasileira de Abasatecimento) para a estimativa nacional de área agrícola, sendo a forma não-probabilística a mais utilizada. Contudo, a abordagem amostral probabilística pode ser uma alternativa para tais estimativa, pois oferece possibilidade de quantificação de erros e variâncias associadas, além de atender à crescente demanda por resultados menos subjetivos (Epiphânio et al., 2002; Adami et al., 2004). Neste sentido, o uso do sensoriamento remoto (SR) permite a aplicação de técnicas probabilísticas, amostrais, e de mapeamento, e pode oferecer uma maior operabilidade e menor custo de execução para as estimativas de área (González-Alonso et al., 1991; FAO, 1996).

As estimativas de áreas agrícolas por amostragem probabilística podem ser realizadas indiscriminadamente para toda a população (amostragem aleatória simples – ASS) ou de forma separada em sub-populações, que possuem características comuns entre si (Amostragem Aleatória Estratificada – AEE) (FAO, 1996; Luiz, 2003). A AEE tem o propósito de redução da variância da estimativa, e portanto, melhorar a estimativa como um todo (FAO, 1996). Também, Cochran (1977) justifica as AEE porque as estimativas geralmente cobrem uma população grande e, portanto, o conjunto de amostras pode não atender ao requisito de ser normalmente distribuída. Além desses pressupostos estatísticos, a AEE também pode contribuir na otimização de recursos humanos, econômicos e de logística, haja vista que a verificação da amostragem é dirigida em sub-populações.

Segundo Luiz (2003), pode haver duas possibilidades de estratificação no cenário brasileiro para estimativas de área agrícola: uma dentro do município e outra entre os municípios. A primeira é essencialmente do tipo espacial e varia de acordo com o aspecto da paisagem; a segunda varia segundo o histórico da área agrícola municipal. No que concerne à

estratificação baseada na paisagem, é reconhecido o potencial que as imagens de SR têm para a detecção de áreas com maior probabilidade de ocorrência de culturas agrícolas (FAO, 1996), e também com em relação à aplicação de técnicas probabilísticas, amostrais, e de mapeamento (González-Alonso et al., 1991). No entanto, dependendo da paisagem, estratificações espaciais em pequenas áreas podem apresentar-se com padrão fragmentado, dificultando o desempenho da AAE (Tsiligrídes, 1998).

Existem diferentes técnicas de reconhecimento de padrões em imagens de SR que podem auxiliar na identificação de áreas com maior probabilidade de ocorrência e auxiliar em estratificações espaciais. Para abastecer tais técnicas de reconhecimento de padrão em imagens de SR, o uso da extração de diferentes atributos que vão além do simples valor do pixel, tais como textura e forma (multi-atributos), têm agregado maior informação e assim melhorado tais técnicas em imagens de SR (Lock e Kirchner, 1997; Oliveira et al., 2004). O uso de multi-atributos também pode ser uma alternativa quando não se dispõe de um adequado número de imagens de SR, ou quando não é possível obter imagens do período de desenvolvimento vegetativo da cultura agrícola em questão quando há cobertura de nuvens. Para o tratamento e processamento de multi-atributos, são aconselháveis técnicas de mineração de dados.

Algoritmos de geração de árvores de decisão (AD) são considerados como de mineração de dados (Witten e Frank, 2005a). A geração de AD constitui uma via não-paramétrica e eficaz quando se tem grande quantidade de dados (Xu et al., 2005). A geração de árvores de decisão pode ser realizada de forma automática pelo algoritmo C 4.5 implementado no programa de uso público WEKA (Witten e Frank, 2005b). Este algoritmo baseia-se em equações de limiarização da entropia das variáveis e ganho das amostras (Quinlan, 1996).

De acordo com as considerações comentadas, este trabalho teve como objetivo avaliar as estimativas de área de agrícolas de verão utilizando AAS e AAE. Foram avaliadas as estratificações que possibilitariam as AAE. As estratificações foram fundamentadas em técnicas de geração de AD em imagens de pré-plantio em relação às culturas de verão.

2. Área de Estudo

Escolheu-se uma área que tivesse uma diversidade de cultivos e que tivesse tradição na produção agrícola, e ao mesmo tempo tivesse uma situação de relevo plano a suave movimentado. Desta forma, escolheu-se o município de Barretos, no Estado de São Paulo (Figura 1).

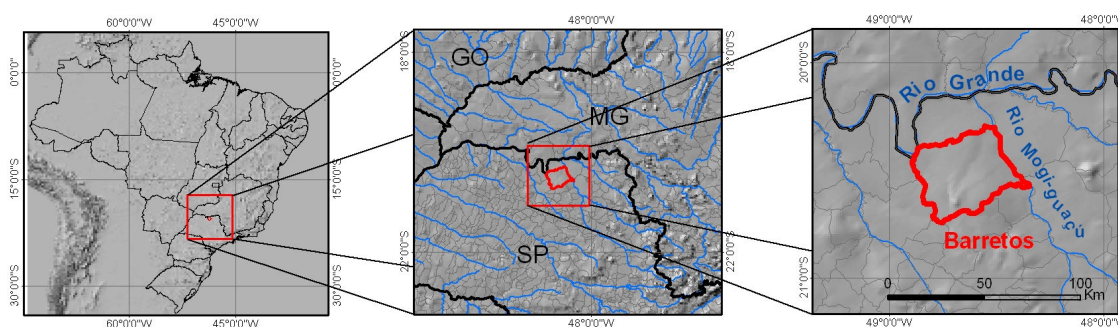


Figura 1. Área de estudo: Município de Barretos (SP).

3. Material e Métodos

Os dados de imageamento orbital utilizados compreenderam imagens do sensor TM a bordo do satélite Landsat-5. Foram coletadas imagens das datas de pré-plantio de 12 de setembro de 2006 e de 15 de novembro de 2006. As imagens foram georreferenciadas com base em uma imagem disponível gratuitamente pela Global Land Cover Facility – GLCF –

(<http://glcfapp.umiacs.umd.edu:8080/esdi/index.jsp>) para a projeção e coordenadas UTM, Fuso 22 e transformação para o Datum horizontal SAD 69. Para que as estratificações pudessem ser avaliadas em relação a todo o município de Barretos, foi gerado um Mapa de Referência do município que retratasse os usos da terra na safra de verão 2006/2007. Para tanto, foi realizada uma verificação de campo em março de 2007. Alguns usos agrícolas foram agrupados em culturas agrícolas de verão.

O processo de estratificação espacial foi realizado mediante classificação das imagens por árvore de decisão. O processo de classificação por árvore de decisão seguiu as seguintes etapas: geração de um plano de informação agregada, segmentação das imagens, extração de multi-atributos, seleção e treinamento de amostras de treinamento, geração da árvore de decisão, geração do estrato espacial.

Informações espaço-temporais condensadas em uma única imagem são de interesse para fins de processamento; assim, gerou-se tal plano de informação. Para tanto, verificou-se visualmente que a banda 3 (vermelho) possuía maior variedade de níveis de cinza em talhões agrícolas em épocas de pré-plantio. Desta forma, a banda 3 da data de novembro foi subtraída da banda 3 de setembro, aplicando-se realce linear, resultando em uma imagem de diferença de datas realçada. Esta nova imagem foi denominada “Dif-b3” e seria considerada como um novo plano de informação.

A etapa de segmentação foi realizada nos moldes do segmentador por multirresolução do programa DEFINIENS (2006). A segmentação foi realizada com a banda espectral 3 das imagens de setembro e novembro. Por se tratar de imagens de diferentes datas, foi realizado um realce gaussiano de desvio padrão 2 para cada imagem, visando manter a mesma variância entre as imagens das diferentes datas. Os parâmetros estabelecidos para a segmentação foram: 70 para escala; e entre os critérios de homogeneidade, adotaram-se 0,6 para cor e 0,4 para forma, sendo que a forma foi repartida igualmente em 0,5 em suavidade e compacidade. Na continuidade dos processamentos foram excluídos os segmentos relativos aos corpos d’água.

A extração de atributos foi realizada pelo programa DEFINIENS; desta forma os cálculos dos atributos de entrada foram realizados por objeto (segmento). A extração dos atributos foi obtida a partir do plano de informação Dif-b3 e das bandas espectrais 3 (b3), 4 (b4) e 5 (b5) das datas de 12/09/2006 e 15/11/2006. Foram extraídas dezenas de atributos; e aqueles atributos que foram selecionados pelas AD serão comentados na etapa de Resultados. As equações dos atributos podem ser consultadas em DEFINIENS (2006).

A seleção de amostras de treinamento - em forma de segmentos - foi baseada em critério visual do especialista. Os segmentos foram gerados nas imagens de datas de pré-plantio, com a pressuposição de que não haveria o conhecimento prévio do especialista em relação aos tipos de uso agrícola que ocorreriam na safra 2006/2007. Assim, a seleção das amostras baseou-se em padrões visuais nas imagens de pré-plantio que poderiam se transformar em futuras áreas de uso de culturas agrícolas de verão. Devido às características da área de estudo, optou-se por separar dois conjuntos de amostras de treinamento que tivessem maior representação da cobertura da terra nas datas de pré-plantio. Assim, foram consideradas amostras de áreas de uso agrícola para culturas verão (AUA) da seguinte forma: onde ocorreu solo nu nas imagens das datas do trabalho (AUA1); e onde ocorreu alguma evolução da vegetação nas imagens das datas do trabalho (AUA2). Por outro lado, para evitar que a AD classificasse as amostras somente em AUA1 e AUA2, e assim toda a área de estudo fosse classificada nessas duas classes somente, foi necessário abastecer o algoritmo com amostras desconhecidas das classes AUA1 e AUA2. O rótulo destas amostras foi denominado “RESTO”.

Com respeito à etapa de seleção de amostras da AD, primeiramente, a quantidade das amostras AUA1 e AUA2 deu-se conforme a disponibilidade de segmentos que pudessem indicar áreas destinadas para culturas de verão. Por outro lado, a seleção das amostras RESTO

foi realizada de forma aleatória e sem reposição, desconsiderando as amostras já selecionadas como AUA1 e AUA2. Como avaliação da quantidade das amostras RESTO, atribui-se três e cinco vezes o tamanho de amostras conhecidas (AUA1 e AUA2). Desta forma, seriam geradas duas classificações conforme a quantidade de amostras RESTO, as quais resultariam nas estratificações finais, sendo denominadas respectivamente de Est-C45-1 e Est-C45-2.

Quanto à geração das AD, é necessário o treinamento prévio das amostras. Esta etapa foi realizada pelo algoritmo de geração de árvores de decisão C 4.5, implementado no programa WEKA (Witten e Frank, 2005b). Para tanto, foram consideradas as configurações padrão deste programa, com exceção do ajuste da quantidade mínima de amostras por folha. Este ajuste visou à geração de AD de tamanhos menores, mas que preservassem as taxas de acerto de AUA. Foram observados os desempenhos das AD geradas pelos “índices kappa” gerados através de validação cruzada de fator 10 entre as amostras de treinamento. Finalmente, a geração de estratos espaciais, que serviriam para a AAE, seria obtida pela extrapolação das regras booleanas das AD geradas para todos os segmentos da imagem.

Em relação às estimativas das AUA, empregou-se o método amostral probabilístico de expansão direta por pontos, utilizando AAS e AAE. Estas estimativas deram-se pela ordem da proporção de AUA “ P ”, cujos pontos amostrais forneceriam a proporção estimada “ \hat{P} ”, ou seja, a informação bimodal de conter AUA (1) ou não conter AUA (0), de acordo com o Mapa de Referência. Os resultados dos parâmetros estimados foram fornecidos por valores médios conforme o método de Monte Carlo, em mil interações. Os cálculos das equações do método amostral empregado, bem como das variâncias e de intervalos de confiança de “ \hat{P} ”, estão descritos em Luiz (2003) e Arcoverde (2008), sob as premissas estatísticas de Cochran (1977) e Bolfarine e Bussab (2005).

Neste estudo, foram separadas as estimativas por AAE cuja proporção de AUA “ P ” estava conforme o Mapa de Referência e a “ P ” que foi observada visualmente nas datas de pré-plantio. A quantidade de pontos aleatórios para a ASS seguiu de acordo com cálculo estipulado por Bolfarine e Bussab (2005), em que se admite uma certa margem de erro e considera-se um grau de confiança baseado em uma distribuição normal padrão “ Z ”. A alocação da quantidade estabelecida nos estratos pela AAE foi realizada segundo a alocação ótima de Neyman, que considera a redução da variância total estimada de acordo com propriedades dos estratos e considera um custo para a estimativa, cujos cálculos seguiram conforme Adami (2007).

4. Resultados e Conclusões

Em relação ao georreferenciamento, os erros quadráticos médios (Root Mean Square – RMS) das respectivas imagens foram ao redor de 0,3 e 0,5. Quanto à geração do mapa de referência, foram consideradas 21 classes. As classes que representaram as áreas destinadas para o uso agrícola de verão (AUA) foram reunidas em: Amendoim, Soja, Cultivo de verão indeterminado, Crotalária e Milho (Figura 2).

Para a fase de treinamento de amostras, foram selecionadas 14 amostras de AUA1 e 10 amostras de AUA2. Desta forma, foram atribuídas como amostras RESTO a quantidade de 72 para o Est-C45-1 e 120 para o Est-C45-2. Quanto à geração das AD, a geração da AD do Est-C45-1 foi executada ajustando o parâmetro de quantidade mínima de instâncias por folha de duas instâncias para sete instâncias. Este procedimento melhorou a taxa de acerto nas amostras de treinamento AUA2 de 40% para 70%, enquanto a amostra AUA1 manteve-se a 64,30%, e o índice kappa passou de 0,4874 para 0,5712. A Figura 3 apresenta esta AD. A geração da AD do Est-C45-2 foi realizada ajustando o parâmetro de quantidade mínima de instâncias por folha de duas instâncias para três instâncias. Isto melhorou a taxa de acerto nas amostras de treinamento AUA1 de 57,14% para 71,43%, e manteve-se em 60% para a

amostra de treinamento AUA2, enquanto que o índice kappa passou de 0,59 para 0,65. A Figura 4 apresenta esta AD.

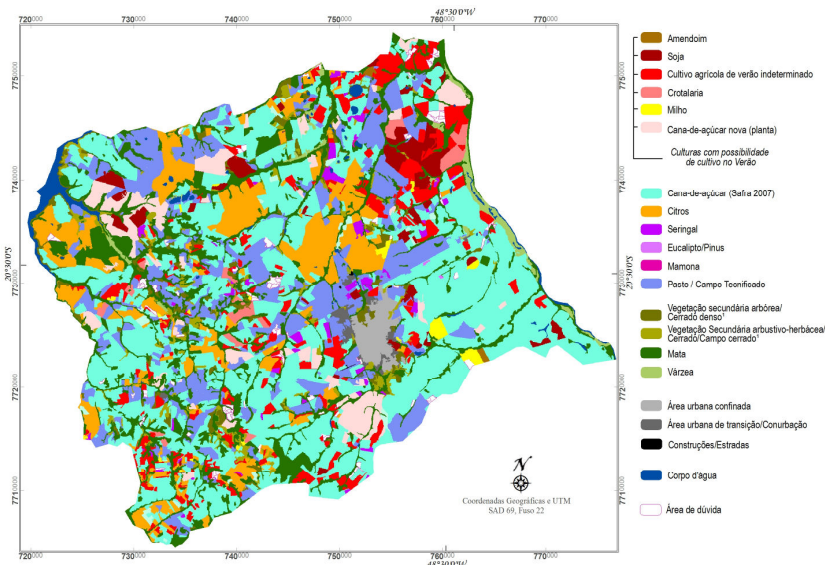


Figura 2. Mapa de Referência.

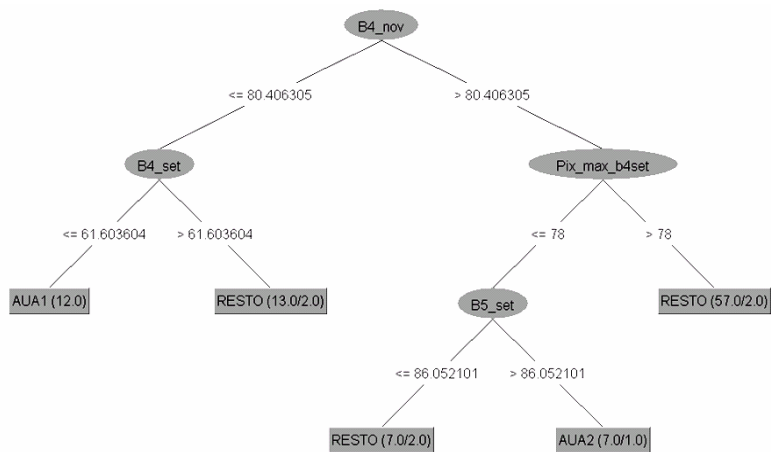


Figura 3. Árvore de Decisão para gerar o Est-C45-1.

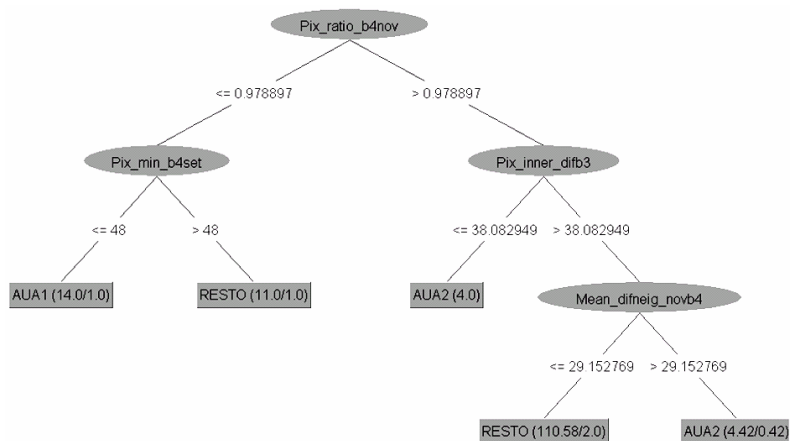


Figura 4. Árvore de Decisão para gerar o Est-C45-2.

Os atributos que ambas as AD selecionaram foram: (a) média das bandas b4 de setembro e novembro, e b5 de setembro (B4_nov, B4_set e B5_set, respectivamente); (b) valor máximo da banda b4 de setembro (Pix_max_b4set); (c) Razão de banda b4 de novembro (Pix_ratio_b4nov); (d) valor mínimo da banda b4 de setembro (Pix_min_b4set); (e) média de borda interna de dif-b3 (Pix_inner_difb3); e (f) a diferença média de objetos vizinhos da banda b4 de novembro (Mean_difneig_novb4).

Foi observado que o Est-C45-1 abrangeu área urbana, enquanto que o Est-C45-2 delimitou bem as AUA. Observou-se também que, apesar de serem AD com quantidade de nós parecidos, o desempenho de ambas foi bem desigual. Este fato comprova que para amostras desconhecidas deve-se atribuir uma quantidade grande de amostras, de forma a melhorar o desempenho global da AD. Haja vista que o desempenho do Est-C45-1 apresentou muitos erros de comissão, optou-se para o seguimento da estimativa de área agrícola, utilizando somente o Est-C45-2.

A proporção de AUA real no município (P), desconsiderando as áreas de corpos d'água, foi de 0,10922, o que significava 16.959,07 de ha. Realizado o método de Monte Carlo para as estimativas de AUA, a Tabela 1 exibe um resumo dos parâmetros de entrada e saída das estimativas por ASS e AEE. A partir desta tabela pode-se observar que: (a) o tamanho do estrato certo (definido) " $Nh1$ " foi menor que o restante da área do município, estrato " $Nh2$ "; (b) as proporções de AUA no estrato definido " $Ph1$ " e das proporções de AUA no estrato restante " $Ph2$ " mediante a " P " observada foram superestimados em relação à " P " real; (c) de acordo com a alocação de Neyman, a quantidade de pontos amostrais do estrato certo $nh1$ deveu-se ser menor que a do estrato restante " $nh2$ "; (d) as estimativas de " P " (\hat{P}) foram bastante semelhantes, assim como as variâncias de " P " ($Var(P)$). A partir desta última observação, através de testes de hipótese, com parâmetro " Z " normal a 5% de significância, de acordo com Mendenhall (1985), pôde ser deferido que todas as " \hat{P} " foram significativamente semelhantes entre si e que todas as " \hat{P} " foram significativamente iguais a " P ".

Tabela 1. Parâmetros de entrada e saída das estimativas de AUA por ASS e AEE.

	$Nh1(ha)$	$Nh2(ha)$	$Ph1$	$Ph2$	$nh1$	$nh2$	\hat{P}	$Var(P)$
AAS	155271		0,10		419		0,10948	0,000224
AAE - Neyman P observada	13788	141483	0,50	0,10	112	307	0,10920	0,000203
AAE - Neyman P real			0,45	0,08	167	252	0,10879	0,000209

A partir desses resultados, observou-se que a estimativa de AUA utilizando estratos não alcançou a pressuposição conjunta de redução da variância com a associação de concentração de pontos aleatórios no estrato certo. Verificou-se que ao atribuir mais pontos no estrato certo " $h1$ ", a variância tendeu a aumentar, o que contradiz os benefícios de diminuição da variância e otimização de logística ao utilizar estratos em estimativas. Este fato foi investigado em simulações realizadas com diferentes tamanhos de estrato certo $Nh1$ em relação ao tamanho da população " N " ($Wh1$), e diferentes proporções de AUA por estrato. A Tabela 2 apresenta os parâmetros utilizados para a simulação, e a Figura 5 apresenta o comportamento das variâncias da P . Neste trabalho o " $Wh1$ " foi de aproximadamente 0,12.

Tabela 2. Parâmetros da Simulação da Variância de P e Figuras associadas.

$Wh1$	$Ph1$	$Ph2$	Figuras
0,12	0,9	0,1	Figura 5 - A
	0,9	0,01	
0,12	0,5	0,1	Figura 5 - B
	0,5	0,01	
0,5	0,9	0,1	Figura 5 - C
	0,9	0,01	
0,5	0,5	0,1	Figura 5 - D
	0,5	0,01	

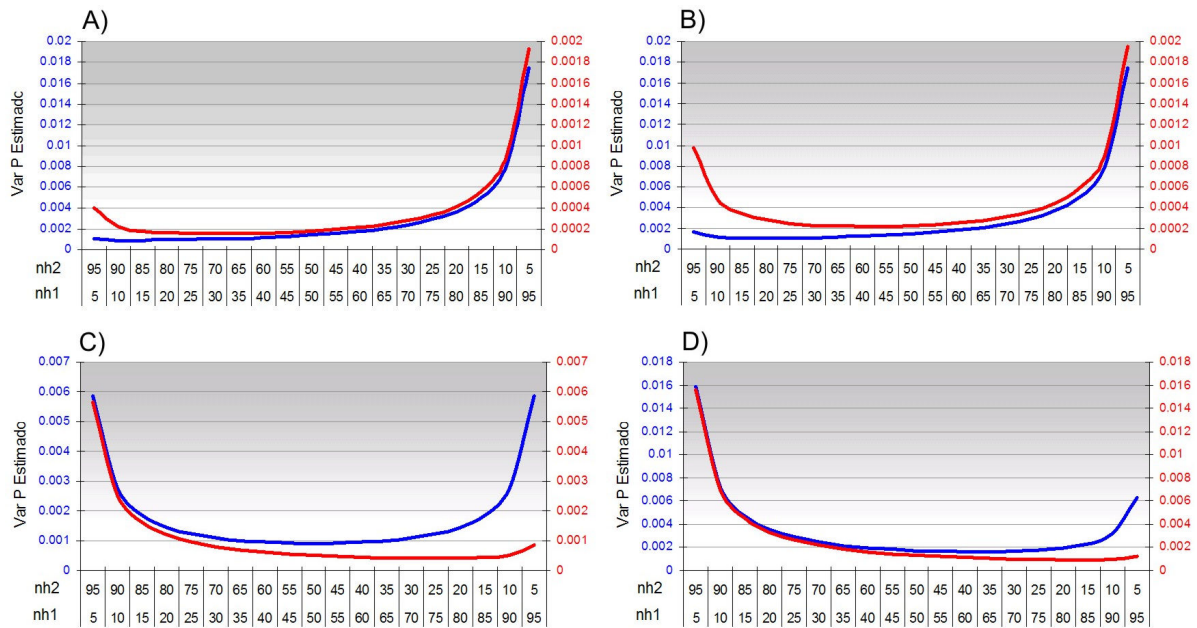


Figura 5. Simulações da Variância de P sob diferentes parâmetros descritos na Tabela 2.

Percebe-se que nos casos com “ $Wh1$ ” igual a 0,12, independentemente das proporções “ $Ph1$ ”, a variância sempre aumenta quando são atribuídos mais pontos aleatórios no Estrato certo. Entretanto, esta situação é alterada nos casos com “ $Wh1$ ” igual a 0,5, principalmente quando “ $Ph2$ ” é considerado muito pequeno (0,01) associado a um “ $Ph1$ ” não necessariamente maior, mas com a maior variância, de 0,5. Neste caso, até certo ponto a variância total é reduzida ao atribuir mais pontos sobre o Estrato-1. Desta forma, constata-se que o comportamento da variância estimada total está bastante atrelado ao tamanho da área dos respectivos estratos, além da proporção que cada estrato representa. Portanto, geralmente se faz necessário alocar mais pontos amostrais em áreas maiores, para que a variância total não seja prejudicada.

Associado à alocação amostral, o fator Custo é um parâmetro estipulado pelo especialista, em que é fornecida uma alternativa para situações em que a quantidade de pontos pode interferir na parte operacional e econômica da estimativa. No entanto, ao atribuir maior quantidade de amostras em estratos de custo mais elevado pode haver interferência na alocação ótima de amostras entre estratos, prejudicando a variância da proporção estimada. Este fato é explicado porque o custo está associado (na maioria dos casos) aos estratos com maiores áreas, e portanto mais pontos seriam alocados nestas áreas, o que aumenta a variância. Portanto, o custo é um parâmetro que pode direcionar a estimativa a ser mais econômica e operacional mas que pode incluir um prejuízo estatístico da estimativa global; assim, deve ser ajustado com cautela. Estas discussões podem ser fonte de estudos para verificar as áreas e municípios que fossem adequados para a estratificação espacial. Este

estudo poderia ser auxiliado, por exemplo, por métricas da paisagem e, assim, permitir estabelecer uma relação entre a componente variância da “ \hat{P} ” e o custo da estimativa com o grau de coesão e tamanho dos possíveis estratos.

5. Referências Bibliográficas

Adami, M.; Moreira, M.A.; Rudorff, T.; Freitas, C.C. Análise da eficiência dos estimadores de expansão direta e de regressão para áreas cultivadas com café, milho e soja no município de Cornélio Procópio, Estado do Paraná. **Agricultura em São Paulo**, v. 5, n. 2, p. 5-13, jul/dez 2004.

Arcoverde, G.F.B. **Estratificação espacial para estimativa da área de culturas agrícolas de verão com imagens de pré-plantio**. 2008. 180p. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, 2008.

Bolfarine, H.; Bussab, W. **Elementos de amostragem**. 1 ed. São Paulo: Edgard Blücher, 2005.

Cochran, W.G. **Técnicas de amostragem**. 2 ed. Rio de Janeiro: Fundo de Cultura, 1977. 555p.

DEFINIENS. **Definiens professional 5**: referente book. Munique: Definiens AG, 2006. p. 122. Disponível em: < http://www.pcigeomatics.com/products/ecog_applications/ecog5_ReferenceBook.pdf>. Acesso em: 16 jun. 2007.

Epiphany, J.C.N.; Luiz, A.J.B.; Formaggio, A. R. Estimativa de áreas agrícolas municipais utilizando sistema de amostragem simples sobre imagens de satélite. **Bragantia**, v. 61, n. 2, p. 187-197, 2002.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS (FAO). **Multiple frame agricultural surveys: current survey based on area and list sampling methods**. Roma, 1996. v. 1.

Gonzalez-Alonso, F.; Soria, S. L.; Gozalo, J. M. C. Comparing two methodologies for crop area estimation in Spain using Landsat TM images and ground-gathered data. **Remote Sensing of Environment**, v.35, n.1, p 2 9-35, 1991.

Lock, R. E. N.; Kirchner, F. F. Classificação de imagens multiespectrais Landsat TM e feições de textura: mapeamento da cobertura da terra. **Revista Floresta**. V. 27, n. 1/2, p. 41-56, 1997.

Luiz, A.J.B **Estatísticas agrícolas por sensoriamento remoto**. 2003. 98p. (INPE-10462-TDI/929). Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, 2003.

Mendenhall, W. **Probabilidade e estatística**. 1 ed. Rio de Janeiro: Campus, 1985. 325 p.

Oliveira, J.A.; Dutra, L.V.; Rennó, C.D. Seleção e extração de atributos para classificação de regiões. In: WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE, 4. (WORCAP), 2004, São José dos Campos. **Anais...** São José dos Campos: INPE, 2004.

Quinlan, J.R. Improved Use of Continuous Attribute in C4.5. **Journal of Artificial Intelligence Research**, v. 4, p. 77-90, 1996.

Tsiligirides, T.A. Remote sensing as a tool for a agricultural statistics: a case study of area frae sampling methodology in Hellas. **Computers and Electronics in Agriculture**, v. 20, p. 45-77, 1998.

Xu, M.; Watanachaturaporn, P.; Varshney, P.K.; Arora, M.K. Decision tree regression for soft classification of remote sensing data. **Remote Sensing of Environment**, v. 97, p. 322-336.

Witten, I.H.; Frank, E. **Data Mining: Practical machine learning tools and technique**. 2 ed. 3.4.12. São Francisco, EUA: Morgan Kaufmann, 2005. 558 p.

Witten, I.H.; Frank, E. **WEKA**. 3.4.12. São Francisco, EUA: Morgan Kaufmann, 2005. Disponível em:< <http://www.cs.waikato.ac.nz/ml/weka/>>. Acesso em: 17 nov 2007.