

Uma avaliação da relação espacial do NDVI com áreas de soja mediante métricas da paisagem e técnica amostral

Gustavo Felipe Balué Arcoverde¹
Eduardo Eiji Maeda¹
Dênis de Moura Soares¹
Corina Costa Freitas¹
Sidnei João Siqueira Sant'anna¹

¹ Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12201-970 - São José dos Campos - SP, Brasil
{gustavo, eduardo, denis}@dsr.inpe.br , {corina, sidnei}@dpi.inpe.br

Abstract. The investigation of the relationship between spatial-temporal variables from remote sensing images with spatial landscape patterns is necessary in the agriculture in order to improve and make operational the agriculture areas estimation. This work suggest the relationship between a multi-temporal NDVI (NDVI-MT) stratum and areas classified as soybean fields in Tuneiras do Oeste county (Paraná-Brasil), under two approaches: using linear regressions from landscape metrics extracted from the stratum with the soybean areas; and using direct expansion sampling techniques, to evaluate the potential of this stratum in order to estimate soybean areas. The regression between the stratum and the soybean areas achieved high correlations, however a moderate and specific result was found with the landscape metrics set. The estimation of soybean areas using the NDVI-MT stratum got a good performance.

Palavras-chave: NDVI, agricultural area, landscape metrics, multiple linear regression, sampling, NDVI, área agrícola, métricas da paisagem, regressão linear multivariada, amostragem

1. Introdução

A agricultura brasileira tem sido um dos pilares na manutenção da economia brasileira. Desde a sua função primordial de suprir alimentos, como também nos setores da agroindústria e energético a mesma vem sendo chave para o desenvolvimento de vários países. Para tanto, esforços institucionais têm sido tomados no intuito de monitorar e planejar tal produção, onde a estimativa de safras tem tido uma demanda primordial nas exportações agrícolas, controle de preços, estocagem agrícola, políticas sócio-econômicas, entre outros.

A previsão de safras no Brasil, implementada desde 1944, é uma atividade inerente ao funcionamento de duas instituições brasileiras: Companhia Nacional de Abastecimento (CONAB) e ao Instituto Brasileiro de Geografia e Estatística (IBGE), Figueiredo (2005). Os levantamentos feitos são a respeito da quantidade produzida, áreas, rendimento médio e valor da produção.

A amostragem tem sido a maneira mais eficaz para a estimativa de safras a nível nacional, sendo a forma não-probabilística mais utilizada, contudo a abordagem probabilística tem tido um crescente destaque devido à sua melhor quantificação de erros e variâncias associadas, além da crescente demanda por resultados menos subjetivos.

O projeto GeoSafras, coordenado pela CONAB, visa o aperfeiçoamento metodológico do sistema de previsão de safras no Brasil, em que nas estimativas de área, emprega uma estatística probabilística amostral por pontos, auxiliado pelo sensoriamento remoto. As particularidades de resoluções espaço-temporais, associadas, por exemplo, à presença de nuvens, tamanho de *pixel* e frequência de imageamento, tem sido limitações nestes tipos de projetos para a aplicação de metodologias exclusivamente de mapeamento. Por outro lado, pesquisas vêm contornando esta situação e aprimorando o papel do sensoriamento remoto à medida que realizam inferências e assimilam técnicas e abordagens de diversos ramos da

ciência. Por exemplo, Soares (2006) estabeleceu uma simulação de uma resolução média espacial a partir de um sensor de baixa resolução utilizando-se de métricas da paisagem.

A paisagem abarca fatores estruturais que se retroalimentam e que possuem funcionalidades – dinâmicos. Uma das formas das métricas da paisagem extraírem estas estruturas é através de padrões baseando-se em aspectos espaciais, ou seja, valores calculados que têm a aptidão de quantificar a composição e configuração espacial de um cenário, o qual pode ser apresentado por um mapa ou uma imagem de sensoriamento remoto, Frohn e Hao (2006).

Para tanto, no ramo agrícola são necessárias investigações da relação entre variáveis espaço-temporais de imagens de sensoriamento remoto orbital, com padrões da paisagem que estão relacionados à estrutura espacial, por exemplo, o comportamento de formas e texturas de alvos agrícolas. Para que futuros estudos possam fazer uso de um maior número de variáveis extraídas de uma imagem orbital, para uma estimativa de área mais aprofundada e operacional.

Neste sentido, este trabalho propõe uma relação da estrutura espacial do índice de vegetação por diferença normalizada (NDVI) com a área de soja mediante duas abordagens. Uma através de regressões lineares, utilizando-se das métricas da paisagem obtidas a partir de um estrato de NDVI; com o intuito de explicar a distribuição da classificação de áreas de soja. E concomitantemente, através de uma amostragem aleatória por expansão direta, através de pontos, avaliar a estimativa de área de soja através deste mesmo estrato.

2. Área de Estudo

Escolheu-se uma região onde houvesse representatividade da produção de soja e facilidade na obtenção de dados. O Paraná possui uma grande produção de soja: na safra de 2004/05 chegou à cifra de 9.541,3 mil toneladas com uma área de 4.148,4 mil hectares (CONAB, 2006). Foi escolhido o município de Tuneiras do Oeste (**Figura 1**), com 69.965 hectares (ha). Possui um clima subquente-úmido com 3 meses de seca, com Argissolos e Latossolos, ambos vermelhos e distróficos. De acordo com o IBGE na safra de 2004/05 a produção de soja foi de 30.000 toneladas, com uma área de 15.000 ha.

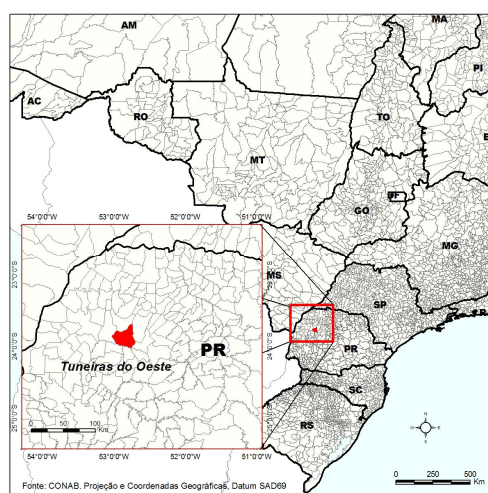


Figura 1: Localização da área de estudo, município de Tuneiras do Oeste (PR)

3. Material e Métodos

Foram obtidas imagens do sensor TM que estivessem livres de nuvens, da data de plantio: 23 de novembro de 2004; e de florescimento e maturação da soja: 11 de fevereiro de 2005. As

bandas 3 e 4 foram transformadas para reflectância aparente, para otimizar e ajustar os valores biofísicos espectrais, necessários para gerar os NDVI's de ambas as datas. Para a confecção do Estrato de NDVI, decorreram os seguintes passos: (a) diferença aritmética entre a imagem de NDVI de fevereiro com a de novembro, resultando em um NDVI multitemporal (denominando-se neste trabalho de NDVI-MT), informando o incremento vegetativo no município; (b) verificação de padrões do NDVI-MT a partir de fatiamentos do histograma, para discriminar espacialmente estratos de incremento vegetativo no período analisado, ou seja, imagens de estratificação do NDVI-MT, onde estas iriam ser posteriormente avaliadas com a classificação de soja, em termos de área (**Figura 2**).

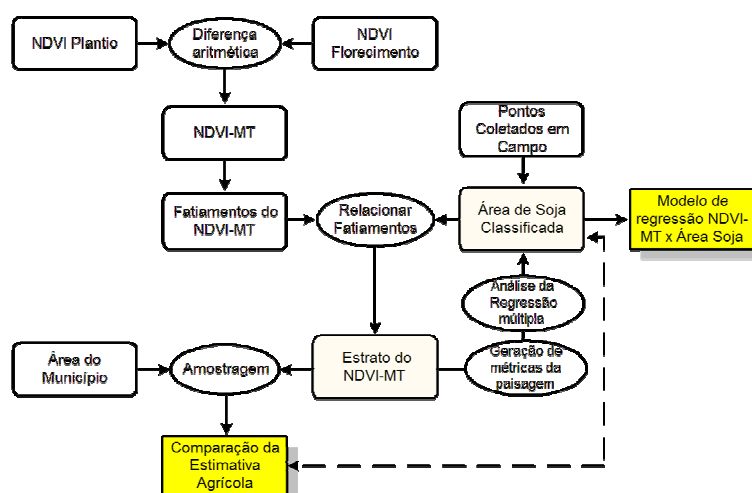


Figura 2: Organograma do trabalho

Para escolher a classificação mais adequada para a área de soja, foram analisadas duas classificações supervisionadas a partir de imagens do sensor ETM+ (bandas 1 a 5, e 7) com relação à um Mapa Verdade – confeccionado por Luiz (2003) no município de Ipuã (São Paulo) da época de florescimento da soja –, uma por *pixel*, a máxima verossimilhança (Maxver), e outra por segmentos, a *Battacharya*. Para o cultivo de soja a Maxver apresentou uma taxa de acerto de 70,64% contra 65,67% do outro classificador, com índices kappa muito próximos, o primeiro com 0,6552 e o segundo com 0,6611. Desta forma, a imagem de fevereiro do município de Tuneiras do Oeste, do sensor TM com as bandas 1 a 5, e 7, foi classificada por Maxver-ICM como áreas de soja e não-soja, tendo como auxílio para as regiões de treinamento, os pontos de coleta de soja em campo do projeto GeoSafras. A área de soja classificada correspondeu a 17179,920 ha, ou 24,5 % do município.

Para avaliar as estratificações das imagens de NDVI-MT que foram mais correlacionadas com a classificação de soja, foram obtidas as áreas em hectares (ha) de cada uma daquelas e da classe soja da classificação; extraídas por segmentos de quadrículas de 50 x 50 *pixels* nas imagens correspondentes, totalizando 292 amostras operacionais (**Figura 3**). Escolhendo-se assim o Estrato do NDVI-MT mais apropriado para a relação de um conjunto de métricas da paisagem com a área de soja classificada.

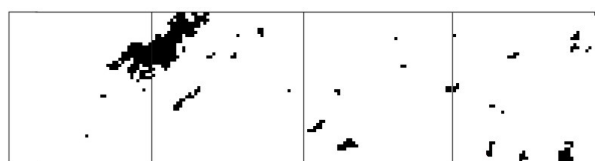


Figura 3: Exemplo de 4 quadrículas de 50 x 50 *pixels* sobre uma estratificação do NDVI-MT (em preto)

As métricas da paisagem foram confeccionadas através do programa *Fragstats 3.3*, McGarigal et al. (1995), (**Tabela 1**), nas quadriculas correspondentes à imagem do NDVI-MT estratificado escolhido. Nesta etapa, considerou-se a abordagem por classes em vez da poligonal, uma vez que a classificação por Maxver apresenta bastantes *pixels* isolados. A escolha das métricas da paisagem levou em consideração as mais adequadas em cada tipologia métrica. Na métrica de Área Nuclear, NDCA; e na de Conectividade, CONNECT, adotou-se uma faixa de 60 metros; na primeira sendo como uma faixa de profundidade e na segunda como um limiar para considerar uma conexão, de acordo com McGarigal et al. (1995).

Tabela 1: Relação das tipologias métricas e as métricas em si, consideradas a partir do Estrato de NDVI-MT para a regressão linear múltipla

TIPOLOGIA	MÉTRICA ESPACIAL DA PAISAGEM	UNIDADE
ÁREA/DENSIDADE/BORDA	LSI (Landscape Shape Index)	-
	PD (Patch Density)	número por 100 ha.
	GYRATE-CV (Coeficiente de variação:Radius of Gyration)	metros
FORMA	SHAPE-CV (Coeficiente de variação:Shape Index)	-
ÁREA NUCLEAR	NDCA (Number of Disjunct Core Areas)	-
ISOLAMENTO/PROXIMIDADE	ENN-CV (Coeficiente de variação:Euclidean Nearest Neighbor Distance)	metros
CONTÁGIO/ESPALHAMENTO	CLUMPY (Clumpiness Index)	-
	AI (Agregation Index)	porcentagem
	IJI (Interspersion and Juxtaposition Index)	porcentagem
	PLADJ (Percentage of Like Adjacencies)	porcentagem
CONECTIVIDADE	CONNECT (Connectance Index)	porcentagem

A partir das métricas da paisagem (variáveis independentes) e da área de soja classificada (variável dependente), ambas extraídas da dimensão das quadriculas supracitadas, foi realizada uma análise de regressão múltipla entre as mesmas. Primeiramente foi observada a normalidade da variável dependente, como uma prerrogativa na avaliação de testes posteriores no modelo. Foram observadas as melhores transformações para este objetivo através do teste de Shapiro-Wilk. Nas observações das variáveis independentes, também foram analisadas algumas transformações, a fim de aproximar a relação linear com a variável dependente e melhorar o coeficiente de determinação (r^2) com esta. As transformações testadas em todas as variáveis foram: a raiz quadrada, o logaritmo na base 10 e o neperiano, a exponenciação e a razão.

Na seleção das variáveis independentes, primeiramente verificou-se as redundâncias de altas correlações com a variável dependente, tratando de eliminar aquelas com um r^2 inferior. Em seguida foi feita uma análise de multicolinearidade, avaliada pelo Fator de Inflação de Variância (VIF) que avalia a variância entre os coeficientes do modelo, Neter et al. (1996), eliminando aquelas com maior multicolinearidade. Com as variáveis restantes, foi realizado um teste *t-Student* que avalia a significância dos parâmetros de forma individual, e testes que avaliam a exclusão de um conjunto destes, como o coeficiente de determinação múltipla (R^2_p), o coeficiente ajustado de determinação múltipla (R^2_a), o C_p , e o StepWise. Adotou-se um teste-F para verificar a possibilidade de exclusão de pelo menos duas variáveis (modelo reduzido), de acordo com Neter et al. (1996).

Na análise de resíduos do modelo, primeiramente avaliou-se os *outliers*, baseando-se em critérios que indicam quais destes são influentes no modelo. Posteriormente foi avaliada a variância dos resíduos por gráficos e pelo teste *Levene*, um teste *t-Student* que se baseia na separabilidade dos dados em duas bandas para a avaliação de heterocedasticidade dos resíduos. Em caso de rejeição deste é necessário o ajuste dos mesmos através do artifício do método dos mínimos quadrados ponderado, que estipula pesos para as amostras.

Para a validação do Modelo utilizou-se um método de comparação de erros de um subconjunto de amostras do modelo gerado, através do cálculo do Erro Quadrático Médio de Predição (MSPR), que deve resultar em um valor próximo ao Quadrado Médio dos Resíduos (MSE) do modelo resultante, Neter et al. (1996).

A análise de estimativa de área de soja com o Estrato do NDVI-MT, seguiu uma abordagem por amostragem de expansão direta por pontos. Para obter as respectivas proporções amostrais de área de soja, em todo o município e no Estrato do NDVI-MT, partiu-se aleatoriamente de 200 pontos para aquele e 42 pontos para o Estrato, que é proporcional à área deste. Em cada ponto houve uma informação binária, de soja ou não-soja, de acordo com a classificação realizada anteriormente. Foram realizadas as estimativas da proporção real de soja para as duas amostragens, assim como um teste de hipótese normal padrão “z” para averiguar se houve diferença significativa em adotar a estratificação do NDVI-MT como base de amostragem. Todas as estatísticas supracitadas foram consideradas com uma confiança de 95%.

4. Resultados

4.1 Análise de Regressão Linear

Foram analisados 5 fatiamentos do histograma NDVI-MT, de acordo com a **Figura 4** observa-se o melhor fatiamento com a respectiva regressão com a área de soja classificada. Assim como, a **Figura 5** ilustra a distribuição espacial deste fatiamento escolhido, denominado Estrato do NDVI-MT, com as áreas de soja classificadas.

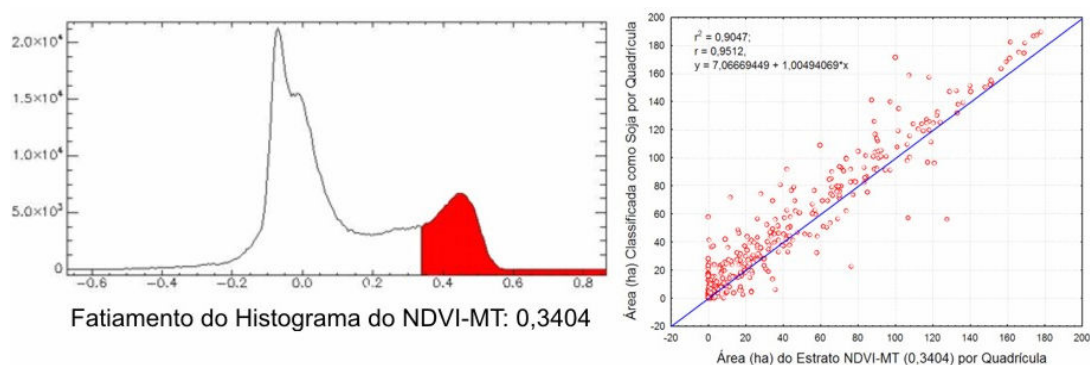


Figura 4: Fatiamento escolhido do NDVI-MT e respectiva regressão linear com a classificação de área de soja

Na regressão linear multivariada, a melhor transformação na variável dependente foi a sua raiz quadrada, aumentando o seu valor de Shapiro-Wilk para 0,9735.

A variável IJI possuía boa parte de seus valores nulos, o que não permitiu a utilização desta no andamento do modelo. Na maioria das variáveis independentes foram necessárias transformações, a matriz de correlação na **Tabela 2** mostra esses resultados e as correlações com a variável dependente.

De acordo com esta matriz foram eliminadas as seguintes variáveis independentes: Logaritmo na base 10 de PD, CLUMPY, e logaritmo da base 10 de AI. De acordo com a análise de multicolineariedade as variáveis mais problemáticas foram as raízes quadradas de PLADJ, NDCA e SHAPE, porém, verificou-se posteriormente que eliminar a raiz quadrada de PLADJ acarretaria um “ r^2 ” baixo para o modelo, por tanto eliminou-se somente as últimas duas, o que favoreceu o andamento da análise, onde as 5 variáveis resultantes apresentaram valores de VIF’s por volta de 1.

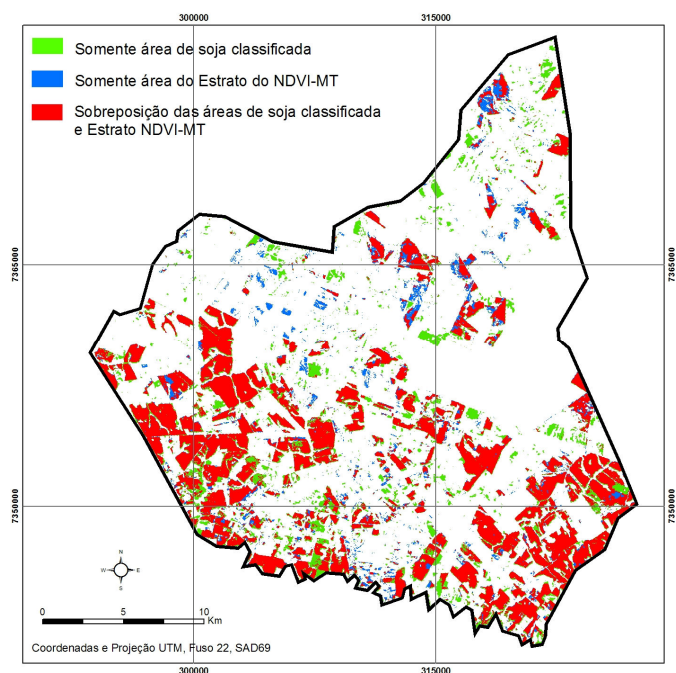


Figura 5: Relação espacial entre as áreas classificadas como soja, o Estrato do NDVI-MT, e a sobreposição de ambos

Tabela 2: Matriz de Correlação

	Area ^{1/2}	PD_log	LSI_log	Gyrate-CV ^{1/2}	Shape-CV ^{1/2}	NDCA ^{1/2}	ENN-CV_log	CLUMPY	Pladj ^{1/2}	Connect ^{1/2}	AI_log
Area ^{1/2}	1,00	0,04	0,15	0,52	0,37	0,67	-0,20	0,55	0,69	0,23	0,59
PD_log		1,00	0,81	0,38	0,30	0,20	0,19	-0,20	-0,03	-0,12	-0,06
LSI_log			1,00	0,32	0,59	0,39	0,12	-0,20	0,03	-0,06	-0,01
Gyrate ^{1/2}				1,00	0,72	0,49	-0,05	0,54	0,64	0,18	0,59
Shape ^{1/2}					1,00	0,50	-0,05	0,27	0,43	0,17	0,37
NDCA ^{1/2}						1,00	-0,10	0,61	0,73	0,12	0,65
ENN_log							1,00	-0,12	-0,13	-0,04	-0,10
CLUMPY								1,00	0,96	0,19	0,95
Pladj ^{1/2}									1,00	0,23	0,98
Connect ^{1/2}										1,00	0,23
AI_log											1,00

De acordo com as abordagens de exclusão de variáveis de forma conjunta, a maioria apresentou a melhor opção para 4 variáveis, foram elas: as raízes quadradas de PLADJ e CONNECT e os logaritmos na base 10 de LSI e ENN-CV. Não obstante, de acordo com o P-valor do teste *t-Student*, para cada parâmetro, apenas 3 variáveis foram significativas: A raiz quadrada de PLADJ e os logaritmos na base 10 de LSI e ENN-CV. Segundo o teste-F sobre aquelas 5 variáveis, permitiu-se a exclusão das variáveis: raiz quadrada de GYRATE-CV e CONNECT; com um F crítico de 3,05 e F calculado de 2,61.

No tratamento dos resíduos, foram eliminadas 17 amostras que foram identificadas como *outliers* significativos. O teste *Levene* rejeitou o modelo como homocedástico, o que levou ao artifício do método dos mínimos quadrados ponderado.

Para a validação do Modelo foram extraídas 40 amostras, conseguindo um MSPR de 3,17 que é próximo ao MSE do modelo definido de 1,52. Sendo definidos os coeficientes de cada parâmetro na **Tabela 3**. Na **Figura 6** estão os valores da área de soja observada versus a área de soja estimada, por quadrículas. Percebe-se que o modelo foi mais adequado para valores entre 6 e 10 dos valores observados da raiz quadrada de área em hectares de soja, o que corresponde a áreas em torno de 36 a 100 ha. Para valores inferiores a estes a estimativa possui muita variância, enquanto que para áreas maiores há uma subestimação de áreas de soja observada.

Tabela 3: Coeficientes e estatísticas do Modelo definido

	Coeficientes	Desvio-padrão	Teste-t	P-valor
Intercepto	-19,8559	2,007334	-9,89167	0,000000
LSI log	6,6510	1,053762	6,31166	0,000000
ENN log	-1,5259	0,529473	-2,88194	0,004304
PLADJ_ ^1/2	2,9464	0,163916	17,97529	0,000000

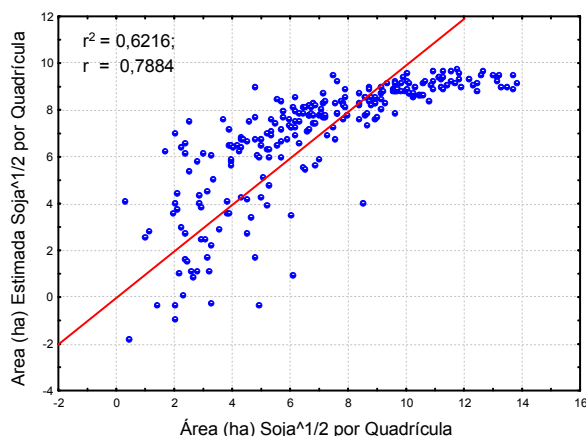


Figura 6: Área de soja estimada (Y) versus área de soja observada pela classificação Maxver-ICM (X).

4.2 Análise Amostral

As proporções de soja encontradas nos 200 pontos alocados em todo município e nos 42 pontos no Estrato do NDVI-MT foram de 0,68 e 0,8571, respectivamente. Na **Tabela 4** encontram-se os dois tipos de amostragem, com os respectivos intervalos de confiança para as proporcionalidades e as áreas de soja estimadas para o município.

Tabela 4: Estimativas das duas amostragens adotadas

Amostragem no Município		Amostragem no Estrato NDVI-MT	
0,615349	< P <	0,74465	0,75126 < P <
43053,06	< Área <	52099,639	11063,918 < Área <
			14385,48

Segundo o teste de hipótese, de que as diferenças de proporções poderiam ser iguais, obteve-se um “z” calculado de 2,3020 contra um “z” crítico de 1,96, por tanto conclui-se que não há razões para rejeitar que o parâmetro estimado pela amostragem na área fatiada seja o mesmo da amostragem realizado na área total do município.

5. Aspectos Conclusivos

As regressões lineares indicaram que apesar de que o Estrato do NDVI-MT e a classificação de soja fossem altamente correlacionados em termos de área, a relação de outras métricas daquele com a variação de área de soja foram moderadamente correlacionados.

Das métricas logradas no modelo podemos sintetizar: (a) o LSI traduz o grau de complexidade de uma forma, ou seja, indica o quão o perímetro do polígono-classe é maior na hipótese de que este fosse um polígono compacto; (b) o Coeficiente de Variação ENN denota as distâncias mínimas entre cada borda de polígono-classe, indica isolamento de feições; e (c) o PLADJ explica o grau de agregação, através de uma medida chamada auto-adjacência, ou seja, uma proporção de células adjacentes da mesma classe. De acordo com as análises de resíduos e regressões entre as variáveis independentes e a área de soja, à medida que esta aumenta, o Estrato do NDVI-MT corresponde a este incremento com polígonos mais

agregados, com menores distâncias entre si, mas com formas mais complexas. Como foi observado, este modelo foi mais adequado para áreas médias neste estudo.

Na técnica amostral o Estrato respondeu eficazmente, estimando uma área de soja próxima àquela pelo IBGE. Este procedimento poderia otimizar estudos e observações de culturas agrícolas de verão em campo para aperfeiçoar estimativas de áreas, para tanto estudos para operacionalizar o fatiamento do histograma de um NDVI-MT devem continuar.

6. Referências

- Adami, M.; Moreira, A. M.; Rudorff, T.; Costa Freitas 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. **Agric. São Paulo**, v. 5, n. 2, p. 5-13, jul/dez 2004. Disponível na biblioteca digital URLib: <<http://iris.sid.inpe.br:1908/>>. Acesso em: 15 out. 2006.
- Companhia Nacional de Abastecimento (CONAB). **Tabela da série histórica de grãos – Safra 1976/77 a 2005/06: Brasil por produtos**. 2006. disponível em <http://www.conab.gov.br/conabweb/index.php?PAG=134>>. Acesso em: 05 ago 2006.
- Figueiredo, D. C. Projeto GeoSafras, sistema de previsão de safras da Conab. **Revista de Política Agrícola**. ano XIV, n. 2, p. 110-120, 2005.
- Frohn, R. C.; Hao Y. Landscape metric performance in analyzing two decades of deforestation in the Amazon Basin of Rondonia, Brazil. **Remote Sensing of Environment**, v. 100, n. 2, p. 237-251, Jan. 2006.
- Frohn, R. C. **Remote Sensing for Landscape ecology: new metrics indicators for monitoring, modeling, and assessment of ecosystems**. Cincinnati: Lewis Publishers, 1998. 99p.
- Instituto Brasileiro de Geografia e Estatística (IBGE). **Sistema IBGE de Recuperação Automática (SIDRA)**. Disponível em: <<http://www.sidra.ibge.gov.br>>.
- Ippoliti-Ramilo, G. A.; Epiphanyo, J.C.N.; Shimabukuro, Y.E. LandSat-5 thematic mapper data for pre-planting crop area evaluation in tropical countries. **Int. J. Remote Sensing**. V. 24, n 7, p. 1521-1534, 2003.
- Luiz, A.J.B **Estatísticas agrícolas por sensoriamento remoto**. 2003. 98p. (INPE-10462-TDI/929). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, 2003.
- McGarigal, K.; Marks, B. J. **FRAGSTATS: spatial pattern analysis program for quantifying landscape structure**. Technical Report PNW-GTR-351. Portland: U.S. 149 Department of Agriculture, Forest Service, Pacific Northwest Research Station, 1995. 122 p. Disponível em: <http://www.fs.fed.us/pnw/pubs/gtr_351.pdf>. Acesso em 10 out. 2006.
- Meyer, P.L. **Probabilidade: aplicações à estatística**. Rio de Janeiro: Livros Técnicos e Científicos Editora S.A., 1977. 391 p.
- Neter, J.; Kutner, M. N.; Nachtsheim, C. J.; Wasserman, W. **Applied linear statistical models**. Boston: WCB/McGraw-Hill, 4ª Ed. 1996, 791 p.
- Périco, E.; Cemin, G.; Lima, D. F. B.; Rempel, C. Efeitos da fragmentação de habitats sobre comunidades animais: utilização de sistemas de informação geográfica e de métricas de paisagem para seleção de áreas adequadas a testes. In: Simpósio Brasileiro de Sensoriamento Remoto, 12., 2005, Goiânia. **Anais...** São José dos Campos: INPE, 2005. p16-21.
- Soares, D. de M. **Áreas agrícolas em sensores com resolução espacial de 30 m estimadas a partir de dados originais e simulados MODIS e métricas da paisagem**. 152 p. Dissertação (Mestrado em sensoriamento remoto) – INPE, São José dos Campos, 2006.