

Análise espacial da colheita da cana-de-açúcar no Estado de São Paulo: a influência da precipitação

Daniel Alves de Aguiar¹
Wagner Fernando da Silva¹
Flávia de Fonseca Feitosa²
Fábio Guimarães Gonçalves¹
Rodrigo Rizzi³
Bernardo Friedrich Theodor Rudorff¹

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

2. Center for Development Research - ZEF
Walter-Flex-Str. 3 D-53113 – Bonn - Alemanha.
flavia@dpi.inpe.br

3. Universidade Federal de Pelotas – UFPel
Caixa Postal 354 · 96010-900 – Pelotas – RS, Brasil.
rodrigo.rizzi@ufpel.edu.br

Abstract Geographic information about the sugarcane harvest development is an essential tool for crop estimation systems. In Brazil the sugarcane's harvest occurs between April and November and the long-term rains prevent the usual progress of harvest activities. As a result, the raw material availability for sugar and alcohol production becomes affected almost every crop. In this work we evaluated three regression models to identify the correlation degree between harvest and rain. We applied the classical linear model and the models with global and local spatial effects in two periods between September and October 2004's crop (PI and PII). Results from statistical and spatial analyzes (Moran, LISA map, and clustering maps) suggested the model with local spatial effects as the best alternative.

Palavras-chave: simple linear regression, spatial regression, geoprocessing, agricultural monitoring, regressão linear simples, regressão espacial, geoprocessamento, monitoramento agrícola.

1. Introdução

O Brasil é um dos líderes mundiais na produção e exportação de vários produtos agropecuários, sendo o maior produtor de cana-de-açúcar e maior exportador de açúcar e álcool do mundo. Mais de 5 milhões de hectares de cana são cultivados no Brasil em duas regiões distintas: Centro-Sul e Norte-Nordeste (UNICA, 2006). O estado de São Paulo é o principal produtor nacional de cana-de-açúcar, sendo responsável por aproximadamente 60% de toda produção de cana, álcool e açúcar e por 70% das exportações nacionais (UNICA, 2006).

O período de colheita da cana-de-açúcar na região Centro-Sul ocorre entre os meses de abril e novembro. A informação sobre a variação da quantidade de cana colhida ao longo desse período é importante dentro de um sistema de estimativa de safra, pois está diretamente ligada à disponibilidade de matéria prima para a produção de açúcar e de álcool e, conseqüentemente, aos estoques destes produtos para comercialização no mercado interno e externo.

Atualmente, a avaliação periódica da evolução da colheita da cana é feita pelo Centro de Tecnologia Canavieira (CTC), por meio de informações amostrais fornecidas pelas unidades produtoras. Uma outra metodologia para o monitoramento da colheita foi apresentada por Aguiar et al. (2006), em que imagens *Normalized Difference Vegetation Index* (NDVI) do

sensor *MODerate Resolution Imaging Spectroradiometer* (MODIS) foram utilizadas. De acordo com ambas as metodologias a safra de 2004 apresentou aproximadamente 4% de cana bisada¹.

Com base na hipótese de que o principal impedimento para o andamento normal do corte da cana é a ocorrência de chuvas de longa duração, o presente trabalho teve como objetivo identificar, por meio de técnicas estatísticas, análise espacial e geoprocessamento, o grau de relação entre a precipitação e a colheita da cana-de-açúcar no estado de São Paulo para safra de 2004.

2. Área e período de estudo

O estado de São Paulo localiza-se aproximadamente entre os paralelos 19° 05' e 25° 00' sul e os meridianos 44° 00' e 53° 05' oeste e está inserido na região sudeste do Brasil, com uma área total de 248.209,426 km² (**Figura 1a**). O período analisado está compreendido entre os meses de setembro e outubro de 2004 e sua escolha justificou-se pela ocorrência de uma queda brusca da colheita (**Figura 1b**).

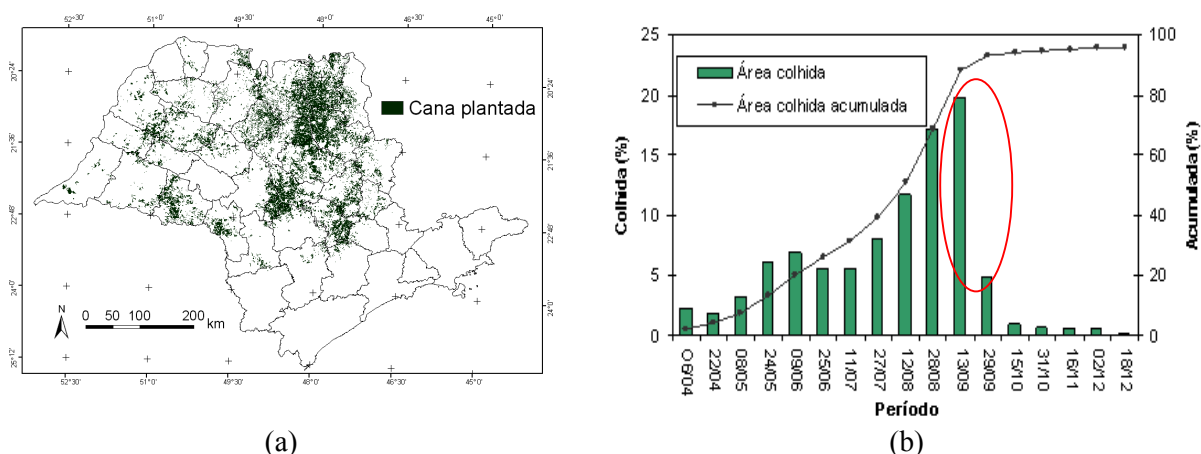


Figura 1 - Localização da área de estudo (a) e gráfico da evolução da colheita da cana para a safra 2004 (b). A elipse vermelha destaca a queda da colheita no mês de setembro.

2. Dados

▪ **Colheita:** um mapa temático com a distribuição espacial da área total de cana colhida em cada Escritório de Desenvolvimento Regional (EDR) foi obtido por meio da metodologia desenvolvida por Aguiar et al. (2006). Este mapa apresenta a área colhida em dois intervalos, a saber: de 13/09 a 28/09/2004 e 29/09 a 14/10/2004, denominados PI e PII.

▪ **Precipitação:** neste estudo foram utilizados dados de previsão de chuva cedidos pelo Centro de Previsão de Tempo e Estudos Climáticos (CPTEC). Os dados são gerados pelo modelo ETA (Black, 1994), que fornece previsão de chuva a cada 6 horas. Seu uso justificou-se pela facilidade de tabulação e ausência de inconsistências. Os dados têm formato de grade regular, com células de 40 por 40 km que cobrem todo o centro-sul do país.

Os valores diários acumulados de precipitação para cada um dos períodos supracitados foram interpolados a fim de ajustá-los aos dados de colheita de cana. O método adotado foi a Krigagem ordinária, que compreende um conjunto de técnicas de estimação e predição de superfícies baseada na modelagem da estrutura de correlação espacial (Bailey e Gatrell, 1995; Druck et al., 2004). As superfícies resultantes foram recortadas, mantendo-se apenas as áreas do estado recobertas com cana-de-açúcar.

¹ Cana não colhida ao final da safra.

2.3 Modelos de regressão

Com o intuito de estabelecer o grau de relação entre a precipitação e a colheita, foram comparados três métodos de regressão: regressão clássica, modelos que incorporam efeitos espaciais globais e modelos que incorporam efeitos espaciais locais.

▪ **Regressão linear clássica:** o objetivo desta técnica estatística é prever e relacionar uma variável dependente a partir de valores conhecidos de uma variável explicativa ou independente (Hair Jr. et al., 1998). Nesta técnica é pressuposto que os dados apresentam independência espacial. Dessa forma, os componentes da equação de regressão não levam em conta a estrutura espacial dos dados (Neter et al., 1996).

A equação que descreve o modelo é apresentada a seguir:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

Em que:

Y = variável dependente (área de cana colhida);

X = variável independente (precipitação);

β_0 e β_1 = parâmetros da regressão;

ε = erro aleatório

▪ **Modelo com efeitos espaciais globais:** este modelo agrega efeitos espaciais globais e supõe que é possível capturar a estrutura de correlação espacial num único parâmetro adicionado ao modelo de regressão clássico (Anselin, 1988; Bailey e Gatrell, 1995; Druck et al., 2004). Um desses modelos é o *Spatial Lag*, que atribui a autocorrelação espacial à variável dependente Y, ou seja, incorpora o efeito espacial ao modelo. Formalmente esse modelo é expresso pela seguinte equação:

$$Y = \rho WY + X\beta + \varepsilon \quad (2)$$

Em que:

W = matriz de correlação espacial;

WY = expressa a dependência espacial em Y ;

ρ = coeficiente espacial autoregressivo;

ε = erro aleatório

Outro modelo com efeitos espaciais globais é o *Spatial Error*, que atribui a autocorrelação espacial ao erro, ou seja, os efeitos espaciais são considerados ruídos ou perturbações que precisam ser removidos (Anselin, 1988).

▪ **Modelos de regressão com efeitos espaciais locais:** neste estudo foi utilizado o modelo de regressão com regimes espaciais que divide a região de estudo em sub-regiões, cada uma com um padrão espacial próprio, considerando as particularidades de cada regime. Uma variável indicadora é utilizada para separar as observações em dois ou mais subconjuntos. Desta forma, cada regime espacial apresenta uma equação de regressão, embora todo o conjunto de dados seja utilizado na estimativa dos coeficientes de regressão (Bailey e Gatrell, 1995; Druck et al., 2004). O modelo geral de regressão está representado a seguir:

$$Y = \beta_0 + \beta_1 X + \beta_2 R1 + \beta_3 R1X + \varepsilon \quad (3)$$

Em que:

β_1 = parâmetros da regressão;

X = variável independente;

R1 = variável indicadora (0 ou 1);

ε = erro aleatório

Portanto, se considerarmos a existência de dois regimes espaciais (0 ou 1), a equação de regressão desmembra-se em outras duas equações, apresentadas a seguir:

$$Y1 = \beta_0 + \beta_1 X + \varepsilon \text{ (indicador = 0)} \quad (4)$$

$$Y2 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon \text{ (indicador = 1)} \quad (5)$$

Um dos modos para identificar os regimes espaciais é por meio do LISA *map* (Índice Local de Associação Espacial), no qual é avaliada a significância dos valores do Índice de Moran Local obtido para cada objeto em relação à hipótese nula, i.e, de não existência de autocorrelação espacial (Druck et al., 2004).

3. Metodologia

As unidades de análise consideradas foram os EDRs para os quais foram calculados os valores de precipitação e área de cana colhida. Entretanto, aqueles que não são produtores de cana-de-açúcar foram excluídos da análise.

A primeira tentativa de estabelecer a relação entre precipitação (variável independente) e colheita (variável dependente) foi por meio da regressão linear clássica. As análises exploratórias dos dados e de aptidão do modelo foram realizadas, i.e, análise de dispersão, análise de *outliers* e análise de normalidade e de variância dos resíduos da regressão.

Para analisar a dependência espacial dos dados foram utilizados mapas de agrupamento pelo método quantil e o Índice de Moran (*I*) para os resíduos da regressão linear. Este índice é um indicador global da autocorrelação espacial e indica como os dados estão correlacionados no espaço (Bailey e Gatrell, 1995; Druck et al., 2004). Valores entre -1 e 0 indicam uma correlação inversa e valores entre 0 e 1 uma correlação direta².

A significância do Índice de Moran foi verificada pelo teste de pseudo-significância, onde foram geradas 999 permutações dos valores de atributos associados. Como cita Druck et al. (2005), se o valor do índice medido originalmente corresponder a um extremo da distribuição simulada, então este é um valor com significância estatística.

A escolha do modelo de regressão com efeitos espaciais globais foi realizada por meio do diagnóstico de dependência espacial gerado pelo *software* GeoDa (<http://geoda.uiuc.edu/>) e baseado na matriz de vizinhança criada. Este diagnóstico indicou o modelo *Spatial Lag* como sendo o mais apropriado para o conjunto de dados.

Posteriormente, a fim de identificar regimes espaciais nos dois períodos de análise, foram gerados LISA *maps*. Os regimes indicados foram ajustados levando-se em consideração os valores de cana colhida por EDR. Após a definição dos regimes, pôde-se aplicar o modelo com efeitos espaciais locais.

4. Resultados e discussões

Quanto à regressão linear clássica, foi possível observar que as variáveis apresentaram uma correlação inversa nos dois períodos: para PI a correlação (*R*) é de -0,33, enquanto para PII de -0,30. A qualidade da regressão foi avaliada pelo coeficiente de determinação (*R*²) e ainda pelo fator Akaike³ (**Tabela 1**).

Por meio de análise gráfica, alguns *outliers* foram detectados para os dois períodos. Entretanto, os testes Cook's Distance e DFFITS confirmaram que esses *outliers* não eram influentes na determinação do modelo e por esse motivo não foram excluídos. O Teste de Breush-Pagan confirmou que a variância dos resíduos, ao nível de confiança de 95%, era

² Em correlações inversas o objeto analisado tem valor baixo e seus vizinhos apresentam valores altos ou vice-versa, já em relações diretas o objeto e seus vizinhos possuem valores altos ou baixos concomitantemente.

³ Quanto menor o valor, melhor a regressão.

constante para PI e não constante para PII. Gráficos de probabilidade normal e o Teste de Jarque-Bera confirmaram a normalidade dos resíduos para PI a 95% de confiança. Todavia, para PII os resíduos não estavam normalmente distribuídos.

O modelo resultante da regressão linear apontou a existência de correlação indireta entre as duas variáveis, corroborando a hipótese do trabalho. Entretanto, os testes de aptidão do modelo indicaram que a relação entre as variáveis não está sendo bem explicada, sendo assim, o modelo não é confiável para se fazer inferências. Os mapas de agrupamento da área colhida e da precipitação para os dois períodos (**Figura 2**) também atestam para este fato, já que apresentam a correlação espacial dos dados.

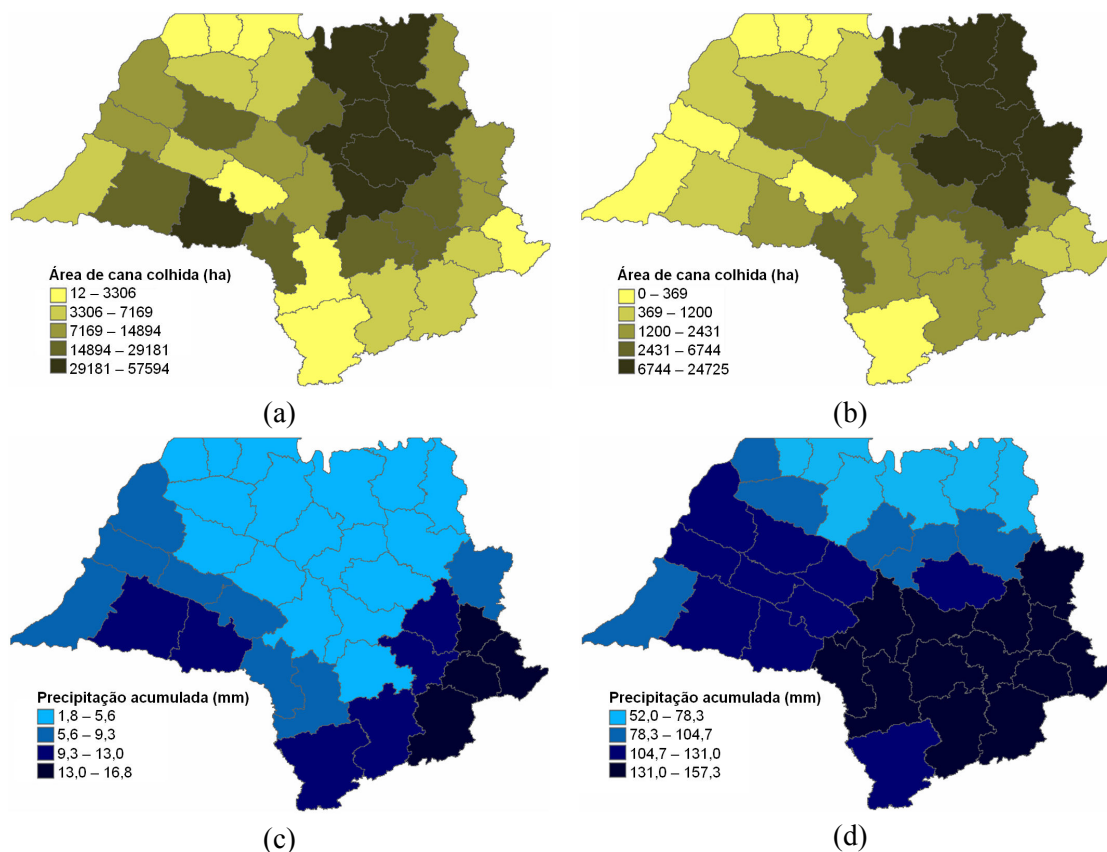


Figura 2 – Mapas de agrupamento da área de cana colhida em PI (a) e PII (b) e precipitação total acumulada em PI (c) e PII (d).

Nota-se, em relação aos dados de colheita, a presença de agrupamentos em ambos os períodos. O agrupamento localizado a nordeste do estado reuniu EDRs com grande produção de cana, p. ex., Ribeirão Preto, Jaboticabal, Barretos e Orlândia. O restante dos EDRs formaram agrupamentos menos homogêneos.

Quanto à precipitação, os mapas de agrupamento indicaram a mudança no comportamento temporal da chuva. Em PI houve um grande grupo, com baixos valores, que ocupou todo o centro norte do Estado e em PII formaram-se três grandes grupos (um ao norte, um a leste e outro a oeste).

A análise dos diagramas de espalhamento de Moran, aplicado aos resíduos da regressão linear (**Figura 3a** e **3b**), possibilitou a visualização da dependência espacial de uma variável em relação a sua vizinhança. Em PI, a dependência foi assinalada em todos os quadrantes, mas com ligeira predominância para os quadrantes alto-alto e baixo-baixo. Para PII houve grande predominância para os quadrantes alto-alto e baixo-baixo.

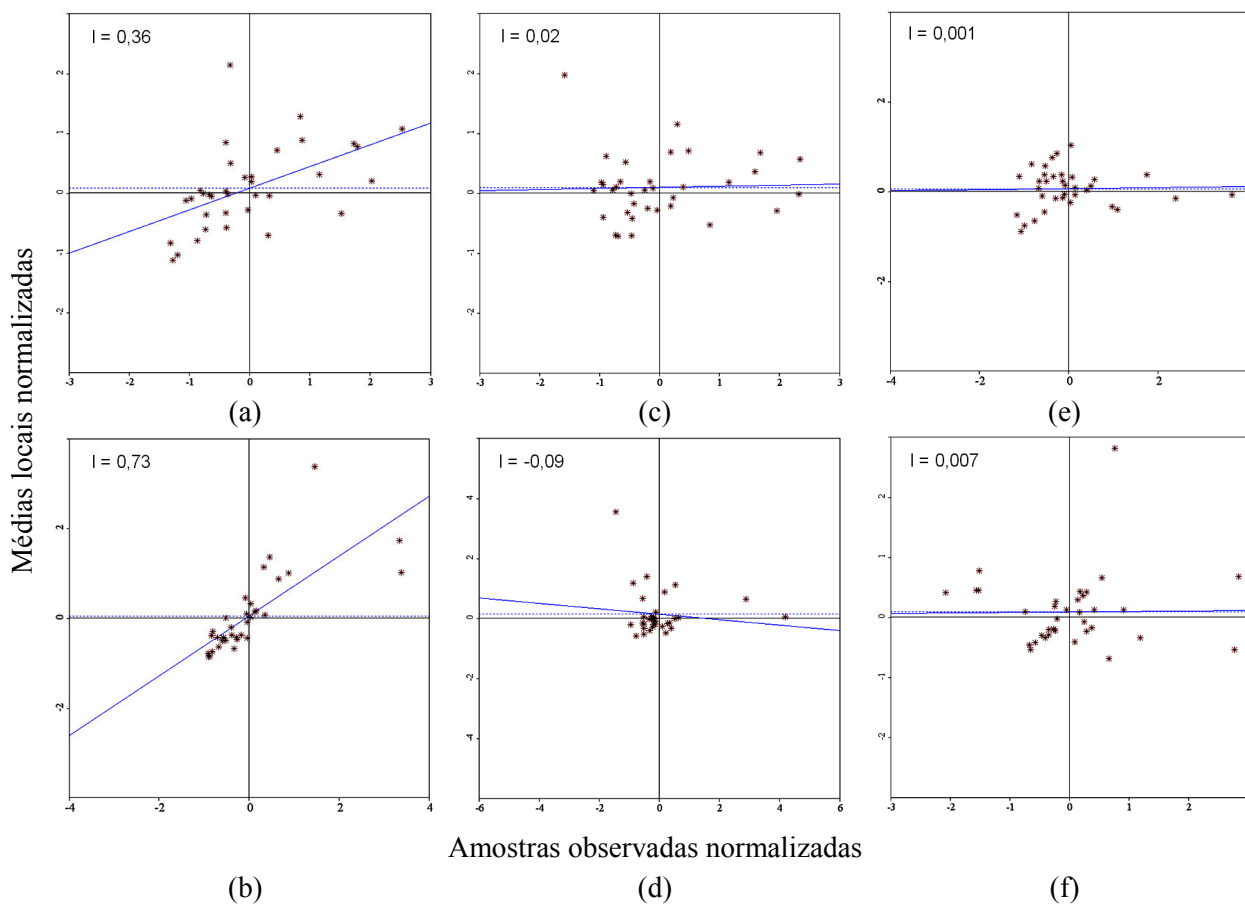


Figura 3 – Diagramas de espalhamento do Índice de Moran aplicado aos resíduos da regressão linear clássica em PI (a) e PII (b); aos resíduos da regressão Spatial Lag em PI (c) e PII (d); e aos resíduos do modelo de regressão com regimes espaciais locais em PI (e) e PII (f).

Ao aplicar o modelo de regressão *Spatial Lag*, tanto o coeficiente de determinação R^2 , quanto o Akaike melhoraram (**Tabela 1**), demonstrando que a incorporação de uma variável espacial auxiliou no ajuste do modelo. Como consequência da inserção desta variável espacial, os valores do Índice de Moran calculados para os resíduos dessa regressão ficaram próximos de zero (**Figura 3c e 3d**). Todavia, uma das premissas desse modelo é que as variáveis sejam estacionárias⁴. Sabe-se que o processo de colheita não é homogêneo em todas as regiões do estado, assim como a precipitação também não é, já que responde a fatores climáticos regionais e de meso escala.

A análise dos LISA *maps* (**Figura 4**) permitiu determinar os regimes espaciais para os dois períodos. Em PI apenas três classes são representadas, contudo a classe *alto-alto*⁵ se destaca e nela estão inclusos os EDRs de maior produção de cana-de-açúcar. Em PII notam-se dois agrupamentos, um a noroeste (*baixo-baixo*) e outro a nordeste (*alto-alto*) do Estado. O grupo *alto-alto* em PII difere de seu correspondente em PI apenas pela inclusão do EDR de Franca, classificado em PI como *alto-baixo*. Posto isso, foram considerados somente dois agrupamentos, apontados na **Figura 4**, a saber: 1 (vermelho e lilás em PI e apenas vermelho em PII) e 0 (branco e azul nos dois períodos). A inclusão da classe *baixo-baixo*, presente nos dois períodos, no regime 0 (quase todo caracterizado pela ausência de agrupamentos) justifica-se pela menor produção de cana-de-açúcar dessa área.

⁴ Processo em que não existem modificações nas variáveis do sistema.

⁵ Alto-alto: valor positivo e média local positiva; alto-baixo: valor positivo e média local negativa; baixo-baixo: valor negativo e média local negativa e baixo-alto: valor negativo e média local positiva.

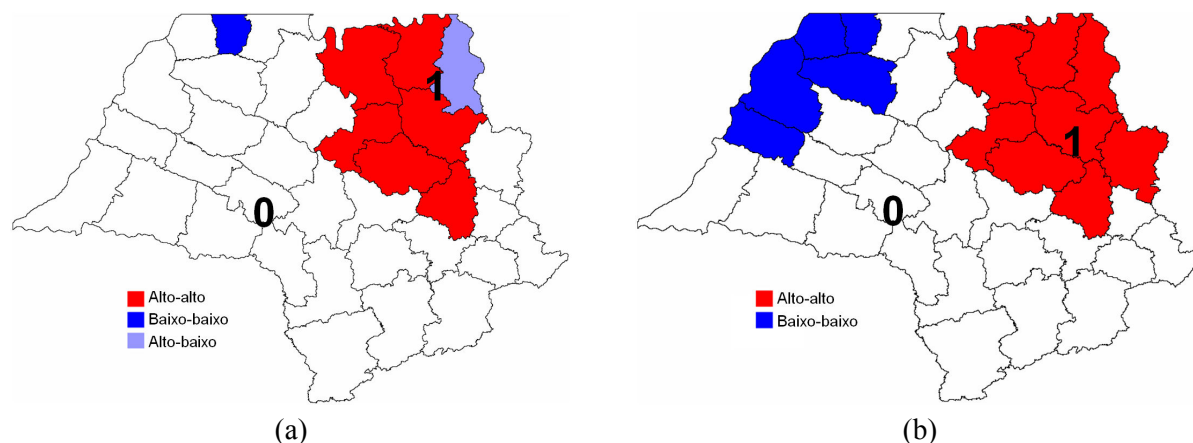


Figura 4 – LISA maps para PI (a) e PII (b), com destaques nos regimes selecionados.

Os valores do Índice de Moran, aplicado aos resíduos da regressão com modelos espaciais locais, são praticamente iguais a zero (**Figura 3e e 3f**), tornando evidente a melhora do modelo com a inclusão dos regimes espaciais.

A **Tabela 1** proporciona uma análise completa das regressões. A redução dos valores de I e do Akaike e o aumento do coeficiente de determinação (R^2), em relação à regressão linear, comprovam que as regressões com efeitos espaciais modelam melhor a relação entre a precipitação e a colheita.

Comparando os dois tipos de regressão espacial aplicados nesse estudo, nota-se que o modelo que leva em consideração os efeitos locais é ainda melhor. O valor de R^2 para PII é consideravelmente maior que o de PI, o que pode ser explicado pelos totais de precipitação maiores e de área de cana colhida menores neste período, em outras palavras, em PII a correlação inversa entre a precipitação e a colheita é mais explícita.

Tabela 1 – Valores de R^2 , Akaike e I para regressão linear e as regressões com efeitos espaciais globais e locais em PI e PII.

Período 13/09 a 28/09/2004 (PI)			
	Linear Clássica	Espacial Global	Espacial Local
R^2	0,11	0,33	0,57
Akaike	777	769	753
Moran (I)	0,360	0,020	0,001
Período 29/09 a 14/10/2004 (PII)			
	Linear Clássica	Espacial Global	Espacial Local
R^2	0,10	0,60	0,95
Akaike	707	684	610
Moran (I)	0,670	-0,090	0,007

O modelo escolhido, com base nos resultados apresentados, foi o com efeitos espaciais locais, a saber:

Para PI

$$C = 11837,94 - 471,21 * P \text{ (Regime 0)} \quad (6)$$

$$C = 35451,52 - 471 * P \text{ (Regime 1)} \quad (7)$$

Para PII

$$C = 2701 - 12,426 * P \text{ (Regime 0)} \quad (8)$$

$$C = 14728 - 12,75 * P \text{ (Regime 1)} \quad (9)$$

Em que:

C = Colheita e P = Precipitação

5. Conclusões

- Foi possível verificar que houve correlação inversa entre a precipitação e a colheita da cana-de-açúcar nos períodos analisados; e que a mesma é espacial.
- O modelo que melhor descreveu a correlação entre a precipitação e a colheita foi o modelo com regimes espaciais locais.

Referências

- Aguiar, D. A.; Silva, W. F.; Rizzi, R.; Rudorff, B. F.; Berka, L. M. S.; Paes, L. A. D. Monitoramento da Colheita de Cana. In Rudorff, B. T. F.; Shimabukuro, Y., E.; Ceballos, J. C. (Orgs.) **O sensor MODIS e suas aplicações ambientais no Brasil**. 1 ed. São José dos Campos: Bookimage, 2006, v. 1, p. 373-378.
- Anselin, L. **Spatial Econometrics, Methods and Models**. Dordrecht, Kluwer Academic, 1988. 304p.
- Bailey, T.C.; Gatrell, A. C. – **Interactive spatial data analysis** 1 ed. New York: Longman Scientific & Technical, 1995, v.1, 416 p.
- Black T.L., 1994: NMC Notes: The New NMC mesoscale ETA model: Description and forecast examples. **Weather and Forecasting**, 9, 256-278p.
- Centro de Previsão de Tempo e Estudos Climáticos (CPTEC). São José dos Campos, 2005. Disponível em: http://www.cptec.inpe.br/prevnum/exp_eta.shtml. Acesso em: 10 de Nov. de 2006.
- Druck, S.; Carvalho, S. C.; Câmara, G.; Monteiro, A. M. V. M. **Análise Espacial de Dados Geográficos**. Brasília, EMBRAPA, 2004. 209p.
- Hair Jr, J. F.; Anderson, R. E.; Tatham, R. L.; Black, W. C. **Multivariate data analysis**. Upper Sanddle River, New Jersey: Prentice Hall, 1998. 742p.
- Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. **Applied Linear Statistical Models** 3ª Ed.1996. 1400p.
- União da Agroindústria Canavieira de São Paulo (UNICA). São Paulo, 2006. Disponível em: <http://www.portalunica.com.br>>. Acesso em: 05 de Jul. de 2006.