Análises de Agrupamento para Modelos de Rendimento de Soja

José Victor B. Rodrigues Viviana Giampaoli

Departamento de Estatística, Universidade de São Paulo

1 Introdução

Devido a grande importância do setor agrícola na economia de um país, o controle das variáveis relacionadas a uma boa produção vem impulsionando pesquisadores de diversas áreas no intuito de compreender a estrutura de relações de um cultivo com o meio ambiente. O foco deste projeto foi desenvolver e melhorar os modelos de predição do rendimento de soja existentes na literatura, utilizando análises de agrupamentos, incorporando aos modelos existentes variáveis indicadoras dos grupos de munícipios formados, que possuem caracterísiscas comuns com respeito à variáveis climáticas. Para a realização deste trabalho foram utilizadas informações de 27 municípios do estado do Paraná.

2 Análise de agrupamento

Análise de Agrupamentos é o nome dado a um conjunto de técnicas utilizadas na identificação de padrões de comportamento em banco de dados através da formação de grupos homogêneos de casos, ou seja, encontrar e separar objetos em grupos similares. Neste caso, separamos os municípios em grupos similares com respeito à variáveis climáticas: temperatura mínima, temperatura média, temperatura máxima, precipitação acumulada e precipitação máxima. O critério de parecença utilizado foi o método das k-médias e o coeficiente de parecença adotado foi a distância euclidiana n-dimensional. Para maiores informações sobre análise de agrupamentos assim como critérios e coeficientes de parecença, o leitor pode consultar Kaufman and Rousseeuw (1990).

O uso dos métodos de partição pressupõe também o conhecimento do número k de partições desejadas, sendo que a escolha do número ótimo de grupos a serem formados na análise de agrupamentos é um problema difícil e existem poucas propostas para solucionar este problema, neste trabalho foi utilizado um índice (DHC_k), baseado na diferença de entropias condicionais, proposto por Souza (2007).

Uma maneira de compararmos os agrupamentos formados é apresentada em Meilă (2002), onde a autora apresenta um índice de comparação, chamado de variação da informação (VI). Considere A o conjunto dos índices de uma amostra de n observações e r variáveis. Então $VI(\boldsymbol{C}, \boldsymbol{C'})$ dá uma medida da variação total da informação induzida pelos agrupamentos $\boldsymbol{C}$ e $\boldsymbol{C'}$ com respeito a A. Além disso, em uma mudança do primeiro para o segundo agrupamento, VI também permite que se quantifique quanta informação sobre $\boldsymbol{C}$ será perdida e quanta informação sobre $\boldsymbol{C'}$ será ganha quando se consideram dois agrupamentos diferentes.

3 Modelo usual de predição

Para as plantas, seu rendimento depende principalmente dos processos de fotossíntese e respiração. E esses processos são diretamente relacionados com a espécie vegetal, nutrição da planta, energia disponível, população de plantas, de plantas daninhas e parasitos. A *evapotranpiração real* (perda de água do solo por evaporação e perda de água da planta por transpiração que ocorre numa superfície vegetada, independente de sua área, de seu porte e das condições de umidade do solo) (Pereira et al, 1975), é dependente dos mesmos atributos e processos citados. Portanto, segundo Dourado-Neto et al. (1999), é adotada a hipótese de que o rendimento pode ser previsto satisfatóriamente a partir da estimativa da evapotranspiração, deficiência e excedente hídrico, possibilitando formar um modelo de predição de rendimento. A relação funcional entre o rendimento agrícola (Y) e os diversos fatores de produção pode ser representada por:

$Y = Y(Energia, CO2, H2O, Nutrientes, O2, Microorganismos, Doenças, Pragas,$
$Ervas\, daninhas, etc.)$

e, considerando que a evapotranspiração real (ETR) depende dos mesmos fatores obtém-se que:

$ETR = ETR(Energia, CO2, H2O, Nutrientes, O2, Microorganismos, Doenças, Pragas,$
$Ervas\, daninhas, etc.)$

Portanto, para estimar o rendimento (Y), adota-se a hipótese:

$$Y = Y(ETR).$$

Um dos modelos mais difundidos na literatura é o modelo de predição relativo de rendimento proposto por Doorembos and Kassan (1979), dado por:

$$\frac{Y}{Ym} = 1 - \beta = [1 - \frac{ETR}{ETm}], \tag{1}$$

Em que:

Y: rendimento da cultura;
Ym: o rendimento máximo da cultura;
β: o coeficiente angular da regressão;
ETR: a evapotranspiração real;
ETm: a evapotranspiração máxima;

Onde a evapotranspiração máxima é a perda de água por evaporação e por transpiração para a atmosfera de uma cultura em qualquer fase do ciclo de desenvolvimento e sem deficiência de água no solo.

Para a cultura de soja, Matzenauer et al. (1998) verificaram que o período com maior consumo de água ocorre entre o início da floração e o início de enchimento de grão. Segundo estes autores, o ciclo de desenvolvimento da soja pode ser dividido em quatro períodos fenológicos, são estes:

1. O primeiro período compreende da semeadura até vinte dias após a emergência, coincidindo com a primeira folha trifoliada desenvolvida, denominado período de estabelecimento da cultura (S-V2).

2. O segundo período compreende de vinte dias após a emergência até o início do florescimento, denominado período vegetativo da cultura (V2-R1).

3. O terceiro período compreende do início do florescimento ao início do enchimento dos grãos, denominado período reprodutivo da cultura (R1-R5).

4. O quarto período compreende do enchimento dos grãos até a maturação fisiológica, denominado período de maturação de grãos da cultura (R5-R7).

Como o déficit hídrico causa efeitos diferenciados em função do estádio de desenvolvimento da planta, Jensen (1968) incorporou este aspecto no seu modelo. O modelo de predição do Jensen considera a sensibilidade da planta ao déficit hídrico para cada fase do seu ciclo de desenvolvimento vegetativo da seguinte forma:

$$\frac{Y}{Ym} = \prod_{i=1}^{n} \left(\frac{ETR}{ETm}\right)_i^{\lambda_i}, \tag{2}$$

Em que:

$\frac{Y}{Ym}$: o rendimento relativo de grãos em um ano agrícola;
ETR: a evapotranspiração real;
ETm: a evapotranspiração máxima;
$\left(\frac{ETR}{ETm}\right)_i$: o consumo relativo de água da cultura no estadio de desenvolvimento i;
λ_i: a sensibilidade relativa da planta ao déficit hídrico em um estádio de desenvolvimento i;

O modelo de Jensen (2) é o modelo de predição no qual foi baseado este trabalho.

4 Modelo proposto

Incorporando as variáveis indicadoras, resultantes da análise de agrupamentos, ao modelo (2), temos o novo seguinte modelo:

$$\begin{aligned} log\left(\frac{Y}{Ym}\right) &= \beta + \lambda_1 log\left(ETR/ETm\right)_1 + \lambda_2 log\left(ETR/ETm\right)_2 + \lambda_3 log\left(ETR/ETm\right)_3 + \\ &+ \lambda_4 log\left(ETR/ETm\right)_4 + A_1 + A_2 + A_3 + A_4 + \epsilon, \end{aligned} \tag{3}$$

sendo:

$\frac{Y}{Ym}$: o rendimento relativo de grãos em um ano agrícola;
ETR: a evapotranspiração real;
ETm: a evapotranspiração máxima;
$\left(\frac{ETR}{ETm}\right)_i$: o consumo relativo de água da cultura no estadio de desenvolvimento i;
λ_i: a sensibilidade relativa da planta ao déficit hídrico em um estádio de desenvolvimento i;
$A_{[i]}$: variável indicadora do agrupamento, no i-ésimo período fenológico.

5 Resultados

O números de grupos formados método das k-médias para cada ano e estádio fenológico são dados na Tabela 1:

Tabela 1: Número de grupos formados pelo método das k-médias.

	Fase 1	Fase 2	Fase 3	Fase 4	Ano
2003	3	6	4	6	4
2004	12	10	5	8	4
2005	10	3	10	5	7
2006	6	3	3	4	3
2007	3	4	7	3	3

Utilizando a variação da informação (VI), podemos observar se os agrupamentos formados se diferem quanto ao estado fenológico e ano observado. A variação da informação varia de 0 a 100%, onde quanto mais próximo de 100%, mais parecidos são os agrupamentos formados, seguem abaixo as Tabelas 2 e 3 com os valores de VI.

Tabela 2: Variação da informação segundo os estádios fenológicos por ano.

Ano 2003	V2-R1	R1-R5	R5-R7	***Ano 2006***	V2-R1	R1-R5	R5-R7
S-V2	0,28	0,21	0,25	S-V2	0,23	0,30	0,23
V2-R1		0,23	0,31	V2-R1		0,23	0,21
R1-R5			0,25	R1-R5			0,23
Ano 2004	V2-R1	R1-R5	R5-R7	***Ano 2007***	V2-R1	R1-R5	R5-R7
S-V2	0,49	0,37	0,44	S-V2	0,26	0,33	0,12
V2-R1		0,47	0,47	V2-R1		0,36	0,30
R1-R5			0,38	R1-R5			0,36
Ano 2005	V2-R1	R1-R5	R5-R7	***Ano 2008***	V2-R1	R1-R5	R5-R7
S-V2	0,42	0,37	0,38	S-V2	0,49	0,23	0,38
V2-R1		0,47	0,18	V2-R1		0,46	0,46
R1-R5			0,45	R1-R5			0,18

O modelos para os anos de 2005 e 2006 foram os que melhor e pior previniram, respectivamente, o rendimento de soja, porém os modelos ajustados quando comparados ao modelo sugerido por Jensem, apresentaram uma melhora na previsão do rendimento de soja em todos os anos. Comparando as estimativas do rendimento de soja do modelo de Jensen com o proposto por nós, percebe-se uma notável melhora quando incorporado o agrupamento. Análisando a medida do erro quadrático médio temos as porcentagens melhoradas indicadas pela Tabela 4:

- Modelo final ajustado para o ano de 2005

$$log\left(\frac{Y}{Ym}\right) = \beta + \lambda_3 log\left(ETR/ETm\right)_3 + \lambda_4 log\left(ETR/ETm\right)_4 + A_1 + A_2 + A_3 + A_4 + \epsilon, \qquad (4)$$

Tabela 3: Variação da informação segundo os anos por estádios fenológicos.

S-V2	2004	2005	2006	2007	2008	***R1-R5***	2004	2005	2006	2007	2008
2003	0,45	0,40	0,26	0,20	0,31	2003	0,30	0,48	0,17	0,34	0,18
2004		0,36	0,53	0,47	0,52	2004		0,37	0,32	0,33	0,27
2005			0,44	0,42	0,53	2005			0,52	0,36	0,48
2006				0,23	0,40	2006				0,36	0,18
2007					0,35	2007					0,38
V2-R1	2004	2005	2006	2007	2008	***R5-R7***	2004	2005	2006	2007	2008
2003	0,45	0,26	0,27	0,31	0,46	2003	0,39	0,28	0,30	0,28	0,45
2004		0,28	0,42	0,50	0,58	2004		0,38	0,38	0,40	0,39
2005			0,25	0,38	0,47	2005			0,21	0,20	0,20
2006				0,28	0,48	2006				0,21	0,15
2007					0,58	2007					0,18

Tabela 4: Porcentagem melhorada quando adotado o modelo 3.

Ano	2003	2004	2005	2006	2007
Porcentagem	66.41	60.26	99.77	49.52	65.83

- Modelo final ajustado para o ano de 2006

$$log\left(\frac{Y}{Ym}\right) = \beta + \lambda_3 log\left(ETR/ETm\right)_3 + A_2 + A_4 + \epsilon, \tag{5}$$

6 Conclusão e Discussão

A idéia de agruparmos municípios parecidos quanto a medidas meteorológicas e incorporá-las aos modelos de previsão de rendimento de soja mostrou-se interessante, melhorando, na pior das situações, 49,52% das estimativas do rendimento de soja, sendo que para 2005 esta melhora foi de quase 100%. Lembrando que neste trabalho, foram utilizados 27 municípios, devido as dificuldades que tivemos para colher as informações necessárias, ou seja, com um pouco mais de recursos pode-se melhorar ainda mais a previsão.

Outro fator importante são os meses em que a cultura permanece em campo. Diferentes regiões possuem diferentes épocas de semeadura, estas normalmente compreendidas entre outubro e dezembro. Um maior banco de dados e uma melhor discriminação das epócas de semeadura pode contribuir muito para a melhora dos modelos.

7 Agradecimentos

Este trabalho foi parcialmente apoiado financeiramente pela CAPES e CNPq.

Referências

Doorembos, J., Kassan, A. H., 1979. Yiel response to water, fao irrigation and drainage paper 33 Edition. FAO, Rome.

Dourado-Neto, D., Garcia, A. G., Fancelli, A. L., et al, 1999. Balance hídrico cíclico y secuencial: estimación de almacenamiento de agua en el solo. Scientia. Agrícola, Piracicaba, v.46,jul.

Jensen, M. E., 1968. Water consumption by agricultural plants. In: KOZLOWSKI, T. T.(ed) Water deficits and plant growth, vol 2, cap.1 Edition. New York: Academic Press, pp. 1–22.

Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data - An Introduction to Cluster Analysis. Wiley, NY.

Matzenauer, R., Barni, N., Machado, F., et al, 1998. Análise agroclimática das disponibilidades hídricas para a cultura de soja na região do planalto médio do rio grande do sul. Relatório técnico, Ver. Bras. De Agromet., Santa Maria, v.t, n.2, p. 263-275.

Meilă, M., 2002. Comparing clusterings. `UW` statistics technical 418, University of Washington.

Pereira, A. R., Villa, N. A., Sediyama, G. C., 1997. Evapotranspiração. FEALQ, Piracicaba.

Souza, E. F., 2007. Comparação e escolha de agrupamentos: uma proposta utilizando a entropia. Dissertação de mestrado em estatística, Instituto de Matemática e Estatística - Universidade de São Paulo.