Análise multivariada e transformada wavelet aplicadas na modelagem da assinatura digital multiespectral da radiação foliar refletida em mudas de eucaliptos com bacteriose

José Raimundo de Souza Passos[1]; Eniuce Menezes de Souza[2]; Edson Luiz Furtado[3]; João Ricardo Favan[4]; André Stefanini Jim[5]; Márcia Lorena Alves dos Santos[6].

Introdução

A bacteriose foliar do eucalipto se caracteriza "inicialmente por lesões encharcadas do tipo anasarca, internervurais, angulares e anfígenas, concentradas ao longo da nervura principal, nas margens da folha ou distribuídas aleatoriamente sobre o limbo". Sua ocorrência abrange os estados de Amapá, Bahia, Minas Gerais, São Paulo, Pará, Mato Grosso do Sul e Rio Grande do Sul. Além da redução da produtividade, o controle e manejo desta doença reduz os custos de produção como também reduz a emissão bactericidas no ambiente. Os agentes causais da bacteriose foram identificados como *Xanthomonas axonopodis* e *Pseudomonas cichori*. A reflectância foliar é sensível ao estresse das plantas à mudança na pigmentação, reação de hipersensibilidade e degradação celular. Podemos assim, associar a reflectância foliar de plantas como sendo uma *assinatura digital ou espectral*, um *padrão* de resposta, podendo variar, para um binômio patógeno–hospedeiro, segundo dois componentes: o temporal – associado a evolução da doença, e para um dado tempo do processo infeccioso, aos comprimentos de onda – resultado de interação do espectro eletromagnético com a estrutura foliar. Os objetivos deste trabalho são:

a) aplicar a técnica multivariada de componentes principais para a redução da dimensionalidade das variáveis aleatorias de reflectância foliar;
b) aplicar a técnica multivariada da função linear e quadrática discriminante de Fisher para classificação e validação dos tratamentos propostos;
c) aplicar transformadas wavelets na modelagem das assinaturas digitais associadas as reflectâncias foliares de eucaliptos segundo os tratamentos propostos;
d) ajustar modelos de regressão logística para a severidade tendo como fatores os coeficentes wavelets;
e) propor modelos de regressão múltipla que associem a severidade aos comprimentos de onda.

[1] Departamento de Bioestatística. Instituto de Biociências. Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP. SP, Brasil.

[2] Departamento de Estatística da Universidade Estadual de Maringá-PR.

[3] Departamento de Produção Vegetal – FCA/UNESP, Botucatu-SP.

[4] FATEC/Pompéia Shunji Nishimura, Pompéia-SP.

[5] Doutor PPG Ciência Florestal – FCA/UNESP, Botucatu-SP.

[6] M.Sc. Bioestatística - Universidade Estadual de Maringá - UEM

2. Metodologia

A técnica de redução da dimensionalidade tem por objetivo reduzir a dimensão dos dados referentes às variáveis reposta de reflectância da dimensão 128 (comprimentos de onda na faixa 966,31 nm a 1685,09 nm). Dentre as técnicas existentes literatura, foi utilizada a técnica da Análise Multivariada denominada de componentes principais – que tem como fundamento a construção de combinações lineares das variáveis aleatórias, no caso, comprimento de onda em nm. Essas combinações lineares possuem propriedades ótimas em termos de variância, buscando-se novas variáveis (coordenadas) que maximizem a variância e não sejam correlacionadas entre si. Nesta redução da dimensionalidade foram considerados os dois primeiros componentes principais. O programa estatístico utilizado: SAS – Free Statistical Statistical Software, SAS University Edition.

Após a redução da dimensionalidade pela técnica componentes principais, será utilizada a técnica da função discriminante linear canônica, que tem como base a classificação de k grupos através de funções lineares no espaço multidimensional utilizando-se de métricas como distância euclidianas e distância de Mahalanobis – que leva em consideração a matriz de variância e covariância das observações. Assim, poderemos obter a validação cruzada – probabilidade de má classificação, que nos informa sobre o percentual de acerto e de erro do modelo obtido via componentes principais. O programa estatístico utilizado: SAS – Free Statistical Statistical Software, SAS University Edition.

A transformada *wavelet* discreta não decimada (TWDND) foi aplicada em cada covariável, decompondo-as em 4 níveis de resolução segundo a wavelet de Haar. Formalmente, uma analise de multirresolução é uma sequência crescente, $\{V_j, j \in \mathbb{Z}\}$, de subespaços fechados de $L^2(\mathbb{R})$, representando os sucessivos níveis de decomposição, tais que eles satisfaçam às seguintes condições:

$MR1 \quad \ldots V_{-1} \subset V_0 \subset V_1 \subset \ldots$

$MR2 \quad L^2(\mathbb{R}) = \overline{\cup_j V_j},$

$MR3 \quad \cap_j V_j = \lim_{j \to -\infty} V_j = 0$

$MR4 \quad f(t) \in V_j \Leftrightarrow f(2t) \in V_{j+1}, \forall\, j,$

$MR5 \quad V_{j+1} = V_j \oplus W_j, W_j \perp V_j$

$MR6$ Existe $\phi \in L^2(\mathbb{R})$, denominada função escala, tal que $\{\phi(x-k); k \in \mathbb{Z}\}$ constitui uma base ortogonal para V_0.

Foi feita a aplicação da TWDND para a variável preditora, cujo codificação é R966, sendo que o mesmo procedimento foi utilizado para os demais 127 comprimentos de onda. Foram ajustados modelos lineares generalizados conderando a variável resposta como a planta sadia e doente (inoculada) e função de ligação logito, tendo como covariáveis os comprimentos de onda. Assim, o modelo logístico irá relacionar a probabilidade de inoculação π_{ij}, associado a i-ésima observação com a j-ésima variável preditora comportamento de onda como se segue,

$$log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_u + \cdots + \beta_j x_j$$

em que $i = 1, \ldots 48$ e $j = 1, \ldots 128$. Os modelos foram construídos considerando uma convariável da cada vez, isto é, para cada conjunto de dados com 48 observações totais (24 inoculados e 24 controle) foram construídos 128 modelos de regressão logística considerando cada faixa de comprimento de onda indivudualmente. Para comparar a capacidade preditiva dos modelos de regressão logística em cada nível, utilizou-se tres diferentes pseudos coeficientes de determinação característics dos modelos lineares generalizados, sendo estes o pseudo R^2 de Mc Fadden's, o pseudo R^2 de Cox & Snell e o pseudo R^2 de Nagelkerke. Além disso, a área sobre a curva ROC (AUC) foi calculada com o intuito de avalizar os valores fornecidos pelos pseudos R^2, considerando a vasta discussão presente na literatura com pontos positivos e negativos sobre a aplicabilidade destes coeficientes no contexto de modelos lineares generalizados.

3. Resultados e Discussão

O menor percentual de erro de classificação geral com a taxa de erro igual a 25%, refere-se ao 3º. dia após a infecção (Tabela 1). Este resultado mostra que, após o inicio do processo infeccioso, no 3º dia houve uma mudança na assinatura digital da reflectância foliar das mudas infectadas com a bacteria. Conforme os resultados obtidos (Figura 1) para os dados em questão, concluímos que os modelos híbridos de regressão logística aplicados aos coeficientes escalas do último nível suave (s4) da decomposição *wavelet* otimizam a predição do diagnóstico da bacteriose foliar das mudas de clones híbridos de *E. grandis* x *E. urophylla* ao longo dos comprimentos de ondas considerados neste estudo para a tomada das medidas de reflectância. Enfatizando que, a partir da faixa de comprimento de onda de 1126 nm, os modelos atingem pseudos $R^2 = 1$, tanto Mc Fadden como Nagelkerke, apresentando eficiência máxima em explicar a variabilidade da variável dicotômica definida como: muda inoculada ou não (controle). Nessa perspectiva, sugere-se uma restrição das faixas de comprimentos de ondas para a medição da reflectância, sendo suficiente medir para comprimentos superiores a 1126 nm para o diagnóstico da planta.

Tabela 1 – Número de observações, percentual de classificação e taxa de erro da validação cruzada pela modelagem da função linear discriminante de Fisher a partir dos dois primeiros componentes principais das combinações lineares das reflectâncias foliares (tempo=3 dias).

		Tratamento classificado		
		Controle	Infectado	total
Tratamento original	controle	19	5	24
		79,17%	20,83%	100%
	infectado	7	17	24
		29,17%	70,83%	100%
	total	26	22	48
		54,17%	45,83%	100%
Taxa do erro		20,83%	29,17%	25,00%

Obs: a) as porcentagens foram calculadas considerando os totais das linhas, por exemplo, porcentagem de acertos no controle: (19/24)100=79,17%. As taxas de erro são complementares as taxas de acertos, por exemplo, no caso anterior, para o controle: taxa de erro= (5/24)100=20,83%.

b) significância do Teste do concordância *kappa*: valor-p = 0,0005.

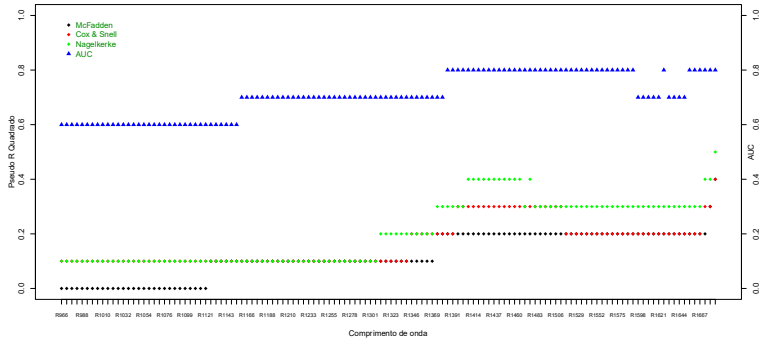

Figura 1 – Pseudos R^2 e AUC para os 128 modelos logísticos usuais do diagnóstico de bacteriose ao longo dos comprimentos de onda, após aplicação da transformada *wavelet* discreta não decimada, com 4 níveis de resolução segundo a wavelet de Haar.

O modelo ajustado de regressão linear múltipla considerando a variável dependente o logaritmo da severidade adicionado da unidade $(\log(\text{severiade} + 1))$ e as variáveis independentes como as 128 reflectâncias foliares na faixa de 966,31 nm a 1685,09 nm. Os valores–p das estimativas dos parâmetros foram todos inferiores a 0,0001 e o R^2_{adj} (ajustado) foi de 64,42%.

$$\hat{y} = -1{,}65X_1 + 1{,}74X_2 - 1{,}41X_3 + 1{,}36X_4 + 0{,}03X_5 + X_6 - 1{,}07X_7$$

em que,

$\hat{y} = log(severiade + 1)$; $X_1 = R1194\_016$; $X_2 = R1110\_853$; $X_3 = R1255\_856$;

$X_4 = R1278\_414$; $X_5 = R1397\_462$; $X_6 = R1638\_673$; $X_7 = R1644\_46$

4. Referências Bibliográficas

FURTADO, E. L.; DIAS, D. C.; OHTO, C. T.; ROSA, D. D. **Doenças do eucalipto no Brasil**. 1. ed. Botucatu: 74 p., 2009.

HUANG, Jing Feng; BLACKBURN, George Alan. Optimizing predictive models for leaf chlorophyll concentration based on continuous wavelet analysis of hyperspectral data. **International Journal of Remote Sensing**, 32:24, 9375-9396, 2011.

MAHLEIN, Anne-Katrin; OERKE, E. ; STEINER, U.; DEHNE, H.. Recent advances in sensing plant diseases for precision crop protection. **European Journal of Plant Pathology**, 133:197–209, 2012.

MORETIN, P. A. **Ondas e Ondaletas – Da Análise de Fourier à Análise de Ondaletas**. Edusp. São Paulo, 1999.

NELDER, J. A; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society Series A (Journal of the Royal Statistical Society. Series A (General)**, Vol. 135, No. 3) 135 (3): 370–384. doi:10.2307/2344614, 1972.