Anais X SBSR, Foz do Iguaçu, 21-26 abril 2001, INPE, p. 1641-1647, Sessão Poster

Análise multivariada de dados dendrométricos e radiométricos referentes à uma plantação de Pinus

CLAUDIA DE ALBUQUERQUE LINHARES [1]
FLÁVIO JORGE PONZONI [1]

[1]INPE - Instituto Nacional de Pesquisas Espaciais
Caixa Postal 515 - 12201-097 - São José dos Campos - SP, Brasil
linhares@dpi.inpe.br
flavio@ltid.inpe.br

Abstract In remote sensing studies, it is a common procedure to relate information acquired from orbital images to those obtained from laboratory or field works. To analyze how both of these two types of information are related, the statistical analysis is the most indicated technique, mainly considering scatterplots, regression models and correlation coefficients. Researches involving vegetation are the most frequently carried out and this type of approach is very usual to understand how spectral data are related to canopy parameter. For a given set of forest parameter, the multivariate statistical analysis is applied to show the potential of this tool for both providing a model and supporting an adequate data interpretation.

Keywords: biophysical parameter, multivariate analysis.

1 Introdução

Estudos envolvendo relações entre parâmetros biofísicos de dosséis florestais e parâmetros radiométricos existentes em imagens orbitais têm se tornado cada vez mais freqüentes (Sousa, 1997, Bernardes, 1998). Nesses estudos, é comum a aplicação de testes estatísticos com o objetivo de avaliar o grau de correlação entre estes parâmetros através de abordagens que incluem a análise individual da ação de um parâmetro específico sobre outro (variável independente versus variável dependente).

Freqüentemente, são ajustados modelos entre parâmetros biofísicos e parâmetros radiométricos, de forma que seja possível inferir sobre as características de um dossel florestal a partir de dados de imagens orbitais. Entretanto, quando da utilização de vários parâmetros simultaneamente, nem sempre os modelos são obtidos considerando-se todos os aspectos estatísticos necessários. Quando se dispõe de um grande número de variáveis, é importante entender como estas variáveis estão se comportando e como estão se relacionando. O estabelecimento de um modelo de regressão múltiplo requer uma análise detalhada das variáveis envolvidas.

É neste contexto que a estatística multivariada aparece como uma ferramenta fundamental e que na maioria das vezes não é utilizada. Assim, o objetivo deste trabalho foi explorar os dados dendrométricos e radiométricos disponíveis e provenientes de plantios de Pinus ssp., de modo a ilustrar a aplicabilidade e a importância da análise multivariada e tentar ajustar um modelo explicativo dos valores de radiância refletida pelo dossel na imagem, a partir de um conjunto de variáveis biofísicas.

2 Material e Métodos

A área de estudo consiste em plantios de *Pinus* spp. de uma Fazenda da Duratex Florestal S/A, das espécies *Pinus caribea* var. *bahamensis* (PCB), *Pinus caribea* var. *caribea* (PCC), *Pinus caribea* var. *hondurensis* (PCH), *Pinus elliottii* var. *elliottii* (PEE) e *Pinus oocarpa* var. *oocarpa* (POO), cujas idades variam de 9,5 a 13,2 anos, caracterizando talhões maduros.

A variável dependente foi a imagem proporção vegetação (Pinus) obtida a partir do modelo linear de mistura espectral aplicado a imagens TM/Landsat_5. As variáveis independentes consideradas inicialmente e, em princípio, explicativas da proporção vegetação, foram IAF (Índice de Área Foliar), DAF (Distribuição Angular de Folhas), DAP (Diâmetro à Altura do Peito), GAP (≈ clareira), Volume de Madeira, Idade, Altura e Área Basal.

Foram aplicadas técnicas de análise multivariada de dados, incluindo uma análise preliminar dos dados (estatística descritiva, correlações entre variáveis, necessidade de transformação dos dados, análise da matriz de correlação, etc.), a escolha das variáveis independentes mais adequadas, o ajuste de um modelo, análise dos resíduos, entre outras.

3 Resultados

3.1 Análise Preliminar e Estatística Descritiva

Primeiramente observou-se a ordem de grandeza das variáveis, visando detectar discrepâncias que pudessem causar problemas na análise. As menores grandezas foram referentes aos dados de GAP (entre 0 e 1) e as maiores, referentes aos dados de Volume (centenas). Isto não representou um problema uma vez que estas variáveis foram eliminadas na etapa de seleção das melhores variáveis para o modelo. Na **Tabela 1** estão listadas todas as variáveis iniciais.

Tabela 1

PINUS	VOLUME	IAF	DAF	GAP	IDADE	DAP	ALTURA	ÁREA BASAL
51,687	133,300	3,720	55	0,145	10,7	19,0	17,6	27,9
40,955	99,800	3,540	54	0,167	10,7	19,0	17,6	28,9
56,313	74,252	4,905	54	0,081	10,4	18,0	12,5	30,8
54,273	69,300	5,250	56	0,074	11,6	18,9	13,4	28,9
38,808	187,592	3,375	49	0,177	10,9	19,0	17,2	30,5
52,493	262,408	5,025	58	0,085	10,9	18,6	19,3	40,3
42,450	144,000	4,470	54	0,104	10,9	19,1	17,2	31,5
49,091	191,748	4,455	60	0,122	10,9	20,2	19,3	32,3
38,808	190,600	5,100	54	0,078	11,8	22,8	20,8	44,3
52,493	259,712	4,245	54	0,114	13,2	22,4	23,4	28,1
52,493	241,625	3,810	52	0,129	14,1	25,9	23,9	30,5
52,493	257,700	2,790	60	0,274	14,8	24,8	25,7	29,8
36,111	131,900	2,685	59	0,263	12,1	19,0	18,5	29,4
44,039	222,504	4,635	51	0,085	12,0	23,1	22,1	42,5
51,924	205,828	5,145	44	0,052	10,7	19,0	19,9	38,4
39,342	173,828	3,990	49	0,109	10,7	18,9	19,9	34,6
55,038	210,500	4,800	59	0,094	12,0	23,0	22,2	41,1
47,431	120,096	3,585	53	0,148	10,4	17,0	16,6	24,5
50,904	156,292	2,985	58	0,222	10,4	18,3	17,3	27,7
38,808	189,228	3,240	42	0,174	10,4	19,2	19,2	32,2
51,687	136,328	5,055	48	0,067	10,7	18,5	19,7	28,9

51,300	185,556	4,005	51	0,118	10,6	18,2	18,9	31,0
51,495	114,075	4,260	47	0,094	10,5	20,1	17,9	31,6
41,731	125,175	2,835	63	0,256	10,6	19,1	17,6	27,1
52,493	243,192	4,665	47	0,079	14,0	23,6	24,3	32,2
59,500	77,000	3,855	44	0,143	9,5	16,5	10,6	22,5
52,493	311,092	3,585	54	0,157	14,5	25,2	27,1	37,7
33,235	83,000	3,465	53	0,165	9,6	16,7	10,6	24,5
48,258	78,000	3,225	52	0,181	9,6	16,0	11,0	27,7
51,572	66,200	4,635	51	0,090	11,1	17,9	12,1	28,8
41,731	123,596	4,230	57	0,122	10,3	19,0	16,9	32,4
43,457	144,996	3,615	48	0,139	10,2	22,1	18,4	33,1
46,892	122,058	3,510	61	0,176	11,7	21,5	18,4	39,3
55,230	107,604	3,930	52	0,134	11,7	18,9	17,5	30,0
41,731	123,200	3,540	57	0,170	10,6	18,6	17,0	27,5
47,431	121,800	3,375	56	0,187	10,7	19,4	17,7	27,4

A **Tabela 2** mostra algumas estatísticas descritivas dos dados e é possível perceber que os dados de volume apresentam valores mínimos e máximos bem extremos, gerando alta variância, enquanto que as demais variáveis mostraram-se com menor variância. Os valores de simetria e curtose (achatamento/alongamento) da curva de distribuição indicam que os dados de idade são os que mais se distanciam do padrão normal.

Tabela 2

	N	**Média**	**Variância**	**Desvio Padrão**	**Mínimo**	**Máximo**	**Simetria**	**Curtose**
PINUS	36	47,672	42,064	6,48566	33,235	59,500	-0,434897	-0,78866
VOLUME	36	157,919	4072,353	63,81499	66,200	311,092	0,516324	-0,53083
IAF	36	3,987	0,534	0,73056	2,685	5,250	0,100817	-1,00419
DAF	36	53,222	25,549	5,05462	42,000	63,000	-0,268584	-0,34690
GAP	36	0,138	0,003	0,05580	0,052	0,274	0,791884	0,25051
IDADE	36	11,264	1,797	1,34039	9,500	14,800	1,362005	1,27495
DAP	36	19,903	6,215	2,49302	16,000	25,900	0,884589	0,04644
ALTURA	36	18,314	15,781	3,97249	10,600	27,100	-0,066042	0,20241
ABASAL	36	31,553	26,865	5,18313	22,500	44,300	0,869352	0,29082

Os gráficos de média e de desvio padrão das variáveis (**Figura 1**) confirmam a informação da **Tabela 2**, onde a maior variância ocorre com os dados de volume de madeira, contrastando com os demais. Quando esta variável é eliminada, as demais tornam-se mais homogêneas.

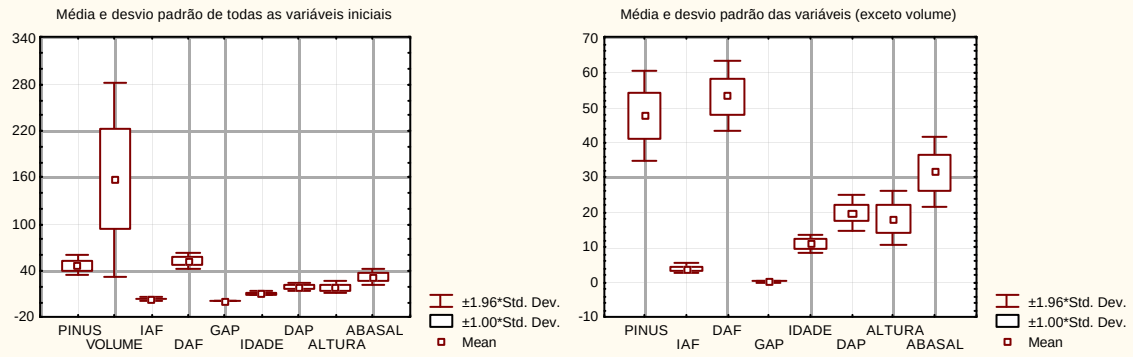

Figura 1

A normalidade dos dados foi verificada através da plotagem dos valores de cada variável sobre a reta de probabilidade acumulada, paralelamente aos testes de Shapiro-Wilk, que apresentaram altos valores de W (acima de 0,9), indicando que os dados mostraram-se normalmente distribuídos, não comprometendo a análise posterior.

Após observar todos estes aspectos preliminares dos dados e de posse de algum conhecimento do comportamento de cada variável, finalmente gerou-se a matriz de correlação (**Tabela 3**), para iniciar o processo de escolha das variáveis a serem consideradas em um primeiro momento, descartando-se as demais que não fossem adequadas. Uma primeira análise visual na **Tabela 3** permite observar que as correlações das variáveis independentes com a variável dependente PINUS são baixas (elipse vermelha), ao contrário do que é esperado como premissa para um bom ajuste de um modelo de regressão múltipla. A variável mais correlacionada com PINUS é o IAF (elipse verde). Observa-se ainda uma alta correlação entre IAF e GAP (elipse rosa) e entre volume de madeira e idade, DAP e altura (elipse azul), indicando que apenas duas destas seis variáveis são realmente necessárias para explicar PINUS.

Tabela 3
(Valores de r = coeficiente de correlação)

	PINUS	VOLUME	IAF	DAF	GAP	IDADE	DAP	ALTURA	ABASAL
PINUS	1	0,12	0,38	-0,07	-0,34	0,26	0,14	0,09	-0,04
VOLUME	0,12	1	0,09	0,01	-0,07	**0,71**	**0,74**	**0,89**	0,54
IAF	**0,38**	0,09	1	-0,23	-0,95	0,01	0,07	0,03	0,47
DAF	-0,07	0,01	-0,23	1	0,44	0,20	0,13	0,09	0,05
GAP	-0,34	-0,07	**-0,95**	0,44	1	0,06	-0,04	-0,01	-0,41
IDADE	0,26	0,71	0,01	0,20	0,06	1	0,86	0,80	0,30
DAP	0,14	0,74	0,07	0,13	-0,04	0,86	1	0,86	0,53
ALTURA	0,09	0,89	0,03	0,09	-0,01	0,80	0,86	1	0,52
ABASAL	-0,04	0,54	0,47	0,05	-0,41	0,30	0,53	0,52	1

A variável dependente PINUS possui baixa correlação com as variáveis independentes (em vermelho), ou seja, nenhuma variável independente explica os valores da imagem vegetação, dando uma prévia idéia de que esta relação radiométrica/dendrométrica não é passível de ajuste em um modelo. As variáveis independentes IAF/ GAP (em rosa) e Volume/ Idade/ DAP/ Altura (em azul) apresentam alta correlação entre si. Estes resultados fazem sentido, pois IAF e GAP foram obtidos através do mesmo instrumento (LAI-2000) e são inversamente proporcionais. As correlações entre Volume, Idade, DAP e Altura têm sido demonstradas em diversos trabalhos. E finalmente, a baixa correlação entre a imagem vegetação (PINUS) e as demais variáveis também é um fato comprovado por alguns trabalhos, ou seja, existem diversos aspectos relacionados à estrutura e fisionomia do dossel que dificultam o estabelecimento de uma relação direta entre imagem e parâmetros biofísicos em alguns casos. Aparentemente, tem-se que quatro variáveis deveriam ser excluídas da análise por não acrescentarem nenhuma ou pouca informação explicativa da variável dependente. Para fundamentar esta escolha, utilizou-se da técnica de Principais Componentes.

3.2 Principais Componentes

A análise por Principais Componentes é útil principalmente quando se tem dados com informações redundantes, pois o objetivo é gerar um novo conjunto de dados não

correlacionados. No caso da análise multivariada, esta técnica serve para auxiliar na escolha das variáveis independentes. Inicialmente, com todas as variáveis, obteve-se que a primeira componente possui 42,87% da variância total de todas as variáveis juntas e a segunda componente, com 26,84%. O gráfico da **Figura 2** comprova o que a matriz de correlação já havia mostrado: alta correlação entre Idade, Volume, Altura e DAP, indicando que deve-se escolher por apenas uma delas.

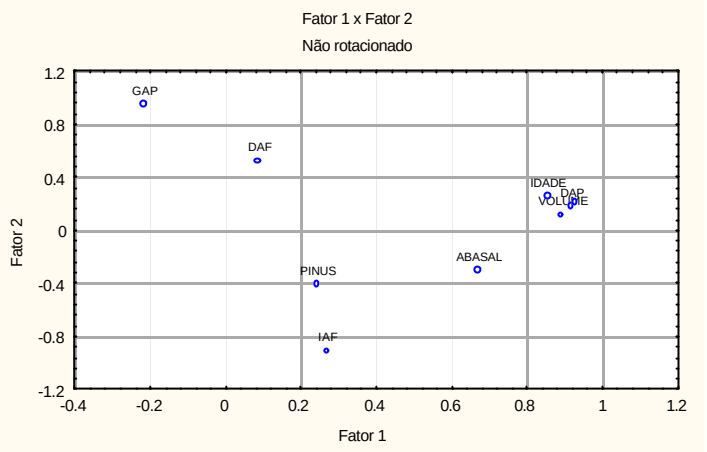

Figura 2

Segundo a análise por Principais Componentes, as variáveis a serem descartadas seriam GAP, optando-se pelo IAF devido à sua relação com a biomassa, além do IAF ser a variável que está mais correlacionada com PINUS e as variáveis Volume, Idade e Altura, optando-se por manter o DAP por ser um parâmetro freqüentemente estimado em campo e que fisicamente possui maior significado para explicar a estrutura do dossel.

3.3 Ajuste do primeiro modelo

Selecionadas as variáveis, seguiu-se para o ajuste de um modelo aos dados. Entrando-se com as variáveis no pacote, o modelo ajustado foi:

$$\hat{Y} = 24{,}02383 + 5{,}43353IAF + 0{,}06588DAF + 0{,}9569DAP - 0{,}65182\,ABASAL \qquad \mathbf{r = 0{,}54885798}$$

O Coeficiente de Correlação (r) foi baixo, indicando uma relação entre PINUS e as variáveis explicativas não muito forte. O valor de β_0 neste caso não possui significado real, uma vez que não há observações das variáveis independentes fora do intervalo estudado, não sendo seguro inferir sobre seu comportamento.

3.4 Teste F para relação de regressão

Este teste apenas verifica a significância dos parâmetros e indica a existência de uma relação de regressão e não a utilidade do modelo em se realizar predições. Testando H_0: $\beta_i = 0$ contra H_1: nem todos β_i igual a zero, tem-se que $F^* > F_t$, rejeitamos H_0 e conclui-se que nem todos os β_i são iguais a zero e portanto, há uma relação de regressão. Para saber qual β_i é diferente de zero, realiza-se um outro teste, a partir do cálculo da estatística t. Testando-se a hipótese H_0: $\beta_i = 0$ contra H_1: $\beta_i \neq 0$, tem-se, ao nível de confiança de 95%: 1) rejeita-se H_0 para β_{iaf} e β_{abasal}, concluindo que estes parâmetros são diferentes de zero e contribuem para o entendimento da

informação radiométrica proporção Pinus; 2) aceita-se H_0 para β_0, concluindo que o valor 24,02 encontrado não é significativo, ou seja, o modelo passa pela origem e 3) aceita-se H_0 para β_{daf} e β_{dap}, concluindo que estes parâmetros são iguais a zero, não possuindo qualquer relação com a variável dependente proporção Pinus.

Assim, DAF e DAP poderiam ser excluídas do modelo, o qual ficaria:

$\hat{Y} = 40{,}54926 + 4{,}53355IAF - 0{,}34713ABASAL$ **r = 0,45244358**

Ajustando-se um modelo com todas as variáveis, o valor de r é 0,58. Como já foi visto, utilizando apenas as quatro variáveis selecionadas após análise preliminar dos dados, temos r = 0,54. Este valor é igual se, daquelas quatro variáveis, retirando DAF, de onde conclui-se que esta variável em nada ajuda na compreensão da imagem PINUS (seu valor de correlação com PINUS é o mais baixo). Quando daquelas quatro variáveis retira-se o DAP, o valor de r é igual a 0,45, o mesmo de quando usa-se apenas IAF e ABASAL (segundo modelo, acima), mais uma vez mostrando que DAF não é relevante. Portanto tem-se suas opções: usar IAF, DAP e ABASAL com r = 0,54 ou usar apenas IAF e ABASAL com r = 0,45.

Outras tentativas de combinação de variáveis mostraram que quando o IAF não está presente, o valor de r é abaixo de 0,29 e ainda, que utilizando IAF/DAF/DAP ou IAF/DAP ou IAF/DAF, o coeficiente de correlação fica em torno de 0,38-0,39, o mesmo obtido quando se usa apenas o IAF. Isto confirma que a variável de maior importância nesta análise é o IAF e que DAF e DAP não contribuem para alterações significativas.

3.5 Ajuste do segundo modelo

Aplicando-se o método de *Stepwise*, obtém-se que as variáveis que devem incluídas no modelo são IAF, IDADE e ÁREA BASAL, excluindo-se DAF, DAP, GAP e Altura. A idade, que apresentava o segundo maior valor de correlação com PINUS, havia sido descartada pela alta correlação com DAP, entretanto, sua correlação com PINUS parece ser significativa para formulação do modelo. Ajustando-se novamente um modelo aos dados indicados pelo *Stepwise*:

$\hat{Y} = 23{,}28087 + 5{,}10197IAF + 1{,}84427IDADE - 0{,}53005ABASAL$ **r = 0,57732680**

Este resultado foi considerado muito bom, pois com apenas 3 variáveis tem-se o mesmo resultado que quando se usava todas as variáveis. Observando-se novamente agora o gráfico de Componentes Principais, nota-se que na primeira componente, aquela cujos dados estão descorrelacionados, IAF, IDADE e ABASAL apresentam-se bem diferenciados nos espaço. Portanto, este último modelo ajustado é o que será avaliado deste ponto em diante.

3.6 Análise de Resíduos

A análise dos resíduos é extremamente importante na identificação de problemas com os dados. A linearidade da função e a constância da variância do erro foram avaliadas através da plotagem dos gráficos ERROS x $\hat{Y}$ e X_i, observando-se que não houve nenhum comportamento tendencioso dos erros, ou seja, nada que indicasse que a função linear fosse inadequada. Concluiu-se também que não houve necessidade de transformação dos dados. A ausência de *outliers* foi verificada a partir dos gráficos do erro padronizado ($e/\sqrt{MSE}$) x X (Neter, 1974). A normalidade dos erros foi comprovada a partir do teste de Shapiro-Wilk (W=0,96958).

4 Conclusões

Dados reais e principalmente aqueles relativos aos fenômenos da natureza possuem alta probabilidade de não serem modeláveis. O comportamento por vezes caótico da natureza muitas vezes frustra as tentativas de se ajustar modelos de regressão múltipla, não pela ineficácia do método, mas pela complexidade do mundo real e pela dificuldade em modelá-lo.

A principal conclusão a que se chega é sobre a importância da análise multivariada não apenas na identificação e na implementação de um modelo de previsão, mas principalmente na compreensão de como seus dados estão se relacionando e que fenômenos podem estar ocorrendo. Interações entre variáveis consideradas, erros correlacionados e não normais, entre outros, são aspectos que influenciam no comportamento do fenômeno analisado e que pode comprometer qualquer tentativa de modelagem. Assim, a análise multivariada colabora para um maior embasamento da pesquisa que está sendo realizada.

Em especial, para os dados aqui avaliados, percebe-se a dificuldade enfrentada na utilização das técnicas de sensoriamento remoto para estimativa de certos parâmetros da natureza, em especial, da vegetação. A matriz de correlação obtida já durante a análise preliminar dos dados mostrava alta correlação entre várias variáveis independentes, as quais eram parâmetros de caracterização da vegetação, cujas correlações são confirmadas pela literatura. Entretanto, nenhuma delas mostrou-se minimamente relacionada com a variável dependente, referente a um dado radiométrico oriundo de uma imagem TM/Landsat. Diversos trabalhos utilizando técnicas de Sensoriamento Remoto têm estudado a relação entre parâmetros biofísicos da vegetação e a informação espectral fornecida por imagens orbitais. As causas do sucesso/insucesso no estabelecimento desta relação depende de vários fatores, como a fisionomia vegetal estudada, porte, arquitetura do dossel, parâmetros que se deseja estimar, entre outros.

Alguns outros fatores devem ser considerados, como amostragem e número de variáveis. Uma amostragem superior à utilizada (ex. 500 amostras) e um número de variáveis estimadas em campo maior (ex. 20 diferentes parâmetros) provavelmente forneceriam uma representação mais fiel da população e maior liberdade de escolha das variáveis independentes mais correlacionadas à variável dependente e menos correlacionadas entre si para entrada no modelo.

Bibliografia

Bernardes, S. Índices de vegetação e valores de proporção na caracterização de floresta tropical primária e estádios sucessionais na área de influência da Floresta de Tapajós-Estado do Pará. (Dissertação de Mestrado) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 1996. 94p.

Neter, J.; Wasserman,W. *Applied Linear Statistical Models.* 1974.

Sousa, C.L., 1997. Uso de imagens-índice e de imagens-proporção para avaliar a quantidade de madeira em povoamentos de Pinus spp. (São José dos Campos: Instituto Nacional de Pesquisas Espaciais).