Capítulo 15

Análises Estatísticas Utilizadas em Dados de Alfafa

Alfredo Ribeiro de Freitas
Reinaldo de Paula Ferreira

Introdução

Nas pesquisas com alfafa as características mais comumente avaliadas e analisadas, dada a importância econômica, são as produções de matéria verde e de matéria seca, a altura da planta, a relação caule/folha, a digestibilidade da folhagem, a proteína bruta, a aceitação fenotípica, a incidência de pragas, as doenças e a persistência. As cinco primeiras características são contínuas, isto é, podem assumir qualquer valor dentro de determinado limite finito. No entanto, a aceitação fenotípica (AF), que dá idéia do comportamento geral da cultivar, e a incidência de pragas e doenças (IPD) são avaliadas subjetivamente por meio de notas e são consideradas variáveis aleatórias discretas, isto é, os valores são finitos, inteiros e enumeráveis. Geralmente, a AF é avaliada por meio de cinco notas (1-excelente; 2-ótimo; 3-bom; 4-regular; 5-ruim) e a IPD, por meio de quatro (0-susceptível; 1-baixa resistência; 2-moderadamente resistente e 3-altamente resistente). Finalmente, a persistência (P), é uma variável do tipo contínua, mas não é avaliada diretamente na planta e sim em função da cobertura vegetal no início (CI) e no final (CF) da área útil da parcela, de acordo com a seguinte fórmula:

$$P = 100 - \left(\frac{CI - CF}{CI} \right) x\,100$$

Naturalmente, em outros tipos de pesquisas com a cultura da alfafa, como por exemplo, envolvendo vacas leiteiras e pastejo rotacionado, vários outros tipos de variáveis são também avaliadas, dentre elas: proteína bruta, nutrientes digestíveis totais, fibra detergente neutro, produção de leite, de gordura e de proteína, entre outras. Entretanto, elas podem ser classificadas em contínuas, discretas ou na forma de percentagens, porém, o procedimento de análise estatística descrito neste capítulo continua válido.

O delineamento experimental

De maneira geral, os experimentos com alfafa são conduzidos em delineamentos experimentais em blocos casualizados com parcelas divididas (*split-plot*). No experimento aqui descrito, para fins de abordagem da aplicação de técnicas biométricas, nas parcelas principais são alocados aleatoriamente os acessos ou as variedades a serem comparadas, denominadas "tratamentos A"; nas subparcelas, são considerados os cortes das plantas realizados no tempo, geralmente mensais, os quais são denominados "tratamentos B". Como já descrito anteriormente, várias variáveis podem ser avaliadas nas pesquisas com a cultura da

alfafa, porém, neste capítulo, serão considerados apenas os dados de produção de matéria seca (PMS), kg/ha.

A descrição do experimento

Para fins de exemplificação, serão considerados os dados experimentais obtidos de ensaio que faz parte de um estudo conduzido pelo programa de melhoramento da cultura da alfafa, em execução na Embrapa Pecuária Sudeste, São Carlos, SP, desde junho de 2004. Foi utilizado o delineamento em blocos casualizados, com duas repetições, sendo que nas parcelas principais foram distribuídos aleatoriamente 92 acessos ou cultivares de alfafa denominados "tratamentos A"; destas, 91 foram provenientes do Instituto Nacional de Tecnologia Agropecuária, Argentina, e uma delas, considerada como testemunha, corresponde à cultivar Crioula, tradicionalmente cultivada no País. Como subparcela, foram considerados 20 cortes mensais sucessivos no período das águas e da seca, denominados "tratamentos B", realizados quando aproximadamente 10 % das plantas estavam em florescimento.

É importante mencionar que, na análise de dados em blocos casualisados com parcelas subdivididas, tanto os tratamentos A quanto os tratamentos B são distribuídos aleatoriamente nas parcelas e subparcelas, respectivamente. A razão da aleatorização é evitar a correlação entre os erros dos dados dos tratamentos alocados às parcelas e entre os dados dos tratamentos alocados às subparcelas.

Entretanto, como os cortes das forrageiras são realizados no tempo, não é possível, portanto, serem distribuídos aleatoriamente às subparcelas. Tal fato faz com que o procedimento estatístico apropriado de análise é o de medidas repetidas-MR (FREITAS et al., 2001; LITTELL et al., 1996; LITTELL et al., 1998; REIEZIGEL, 1999). Na análise de MR, a ordem das observações realizadas no indivíduo é fundamental e, geralmente, existe uma estrutura de correlação entre elas, sendo as mais próximas mais correlacionadas.

Dentre alguns exemplos de análises de MR na agricultura pode-se citar o crescimento corporal dos animais, a produção de leite, a contagem de ecto e endoparasitos, a produção em cortes das forrageiras, o tempo de armazenamento e a qualidade de produtos. Na análise de MR, a unidade experimental onde são realizadas as avaliações no tempo geralmente é denominada "indivíduo".

Na Tabela 1, é apresentada a estrutura dos dados do experimento realizado em blocos casualizados, com duas repetições; a parcela é representada pela combinação variedade-bloco, constituindo 184 indivíduos, sendo que em cada um deles são realizadas as 20 avaliações. Esta estrutura facilita ao leitor a compreensão

dos princípios teóricos necessários para o emprego da análise de medidas repetidas. Para padronizar a descrição da análise, as variedades ou acessos serão denominados "Tratamento" e as avaliações no tempo, "Corte".

Tabela 1. Esquema de obtenção dos dados de produção de matéria seca.

Tratamento (Variedade, Bloco)	Indivíduo	Corte (Avaliações no indivíduo)				
1, 1	1	$y_{1,1,1}$	$y_{1,1,2}$	...	$y_{1,1,19}$	$y_{1,1,20}$
1, 1	2	$y_{1,1,1}$	$y_{1,1,2}$	...	$y_{1,1,19}$	$y_{1,1,20}$
...	...			...		
92, 1	92	$y_{92,1,1}$	$y_{92,1,2}$	...	$y_{92,1,19}$	$y_{92,1,20}$
1, 2	93	$y_{1,2,1}$	$y_{1,2,2}$	...	$y_{1,2,19}$	$y_{1,2,20}$
2, 2	94	$y_{2,2,1}$	$y_{2,2,2}$	...	$y_{2,2,19}$	$y_{2,2,20}$
...	...			...		
92, 2	184	$y_{92,2,1}$	$y_{92,2,2}$	...	$y_{92,2,19}$	$y_{92,2,20}$

Fonte: Freitas, arquivo pessoal.

O modelo estatístico

Na Tabela 2, é apresentado o quadro de análise de variância, que contém as fontes de variação com os respectivos graus de liberdade. A compreensão desta tabela é facilitada a partir da Tabela 1. Com as informações das Tabelas 1 e 2, o próximo passo é a elaboração do modelo matemático para a análise de medidas repetidas, que é dado por:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + t_k + (\alpha t)_{ik} + \varepsilon_{ijk}, \quad (1)$$

em que:

y_{ijk} é o valor observado da PMS no corte k, no indivíduo j e no tratamento i;

μ é o efeito médio global;

α_i é o efeito fixo do tratamento i (i=1,2,...t);

β_j é o efeito fixo de bloco (j=1,2,...b);

δ_{ij} é o efeito aleatório da unidade experimental j no tratamento i;

t_k é o efeito fixo do corte k (k=1,2,...c);

$(\alpha t)_{ik}$ é o efeito da interação de tratamento e corte e;

ε_{ijk} é o erro aleatório associado à PMS no corte k, na unidade experimental j e no tratamento i.

Tabela 2. Quadro de análise de variância.

Fontes de variação	GL	SQ
Blocos – B	b-1 = 1	SQB
Tratamentos –T	t-1 = 91	SQT
Resíduo a	(b-1) (t-1) = 91	SQEa
(Parcelas)	bt-1 = 183	SQP
Cortes – C	c-1 = 19	SQC
Interação T x C	(c-1) (t-1) = 1729	SQTxC
Resíduo b	t(b-1) (c-1) = 1748	SQEb
Total	(btc-1) = 3679	

Fonte: Freitas, arquivo pessoal.

Considerando-se o modelo estatístico (1), para quaisquer duas respostas avaliadas no indivíduo, por exemplo, nos tempos k e l ($k \neq l$), verifica-se que os erros são correlacionados. Designando-se covariâncias e variâncias por Cov e Var, respectivamente, tem-se $Cov(y_{ijk}, y_{ijl}) = Cov(\delta_{ij} + e_{ijk}, \delta_{ij} + e_{ijl}) = Var(\delta_{ij}) + Cov(e_{ijk}, e_{ijl})$; entretanto, quando duas respostas são avaliadas em indivíduos diferentes, elas não são correlacionadas, isto é, $Cov(y_{ijk}, y_{i'j'k'}) = 0$, não importando se $i = i'$ ou $i \neq i'$; $k = k'$ ou $k \neq k'$. A principal conseqüência dessa estrutura de correlação é que os erros associados às avaliações dentro de indivíduos são representados por uma matriz, isto é, $V(\varepsilon_{ijk}) = R$.

Se a aleatorização entre as avaliações no tempo (corte) fosse possível, os erros e_{ijk} seriam independentes, identicamente distribuídos e descritos por $V(\varepsilon_{ijk}) = R = \sigma^2 I$, em que σ^2 é a estimativa da variância residual e I a matriz de identidade de ordem n (n=btc). Nesse caso, a análise seria a de blocos casualizados com parcela subdividida realizada de modo convencional.

Antes de ilustrar os procedimentos de realizar análise de medidas repetidas, é fundamental conhecer a qualidade dos dados, o que é feito por meio de análise exploratória.

Análise exploratória

A análise exploratória geralmente consiste no uso do diagrama de caixa (*Box–plot*), cálculo de coeficientes de simetria, de curtose e teste de normalidade dos erros. Diagramas de caixa são representações gráficas de um conjunto de dados, que possibilitam comparar distribuições de dados contínuos e revelar características importantes, como a dispersão dos dados em torno da média, o grau e a direção da

simetria, a existência de heterogeneidade de variâncias, a presença de *outliers*, entre outras (CLEVELAND, 1994; DIGBY e KEMPTON, 1996; GOWER e HAND, 1996).

Uma alternativa fácil e prática para utilizar o diagrama de caixa é por meio do módulo INSIGHT do software *Statistical Analysis System* - SAS. Entretanto, é importante descrever a construção básica do diagrama de caixa. Inicialmente, ordenam os dados X_1, ..., X_n, da amostra em que o diagrama será utilizado; em seguida, determinam os elementos de posição 25 %, 50 % e 75 %, denominados, respectivamente, primeiro (Q_1), segundo (Q_2) e terceiro quartil (Q_3); a diferença ($Q_3 - Q_1$) é a diferença interquartílica.

Um exemplo do uso do diagrama de caixa para os dados de produção de matéria seca (PMS) de alfafa dos 20 cortes é apresentado na Fig. 1. A parte inferior e superior da caixa corresponde, respectivamente, ao quartil Q_1 e Q_3, a linha horizontal cheia no meio indica a mediana ou segundo quartil (Q_2); as caixas estreitas (*"whiskers"),* acima e abaixo da caixa central, possuem distância não superior a 1,5 vezes a distância interquartílica. As marcas individuais nos extremos das caixas estreitas são consideradas dados discrepantes e candidatas a serem *outliers*.

Possíveis *outliers* são determinados a partir de quatro valores calculados em função de Q_1, Q_2 e Q_3, ou seja, $L_1 = Q_1 - 1{,}5(Q_3 - Q_1)$; $L_2 = Q_1 - 3{,}0(Q_3 - Q_1)$; $U_1 = Q_3 + 1{,}5(Q_3 - Q_1)$ e $U_2 = Q_3 + 3{,}0(Q_3 - Q_1)$. Se as marcas individuais estão dentro do intervalo L_1 e L_2 e U_1 e U_2, os dados são apenas discrepantes; se os dados são menores que L_2 e maiores que U_2, eles são *outliers*. O gráfico superior da Fig. 1 representa os dados originais, isto é, com a presença de *outliers*; no gráfico inferior, os *outliers* já foram eliminados da amostra.

Observando-se o gráfico superior (Fig. 1), verifica-se que os maiores valores de PMS, pela ordem, foram obtidos nos cortes 2, 1 e 3, enquanto os menores foram obtidos nos cortes 9, 17, 18 e 19. Observa-se ainda que a mediana de PMS, ao longo dos cortes, não tem ocorrência casuística, pois segue uma tendência devida ao efeito de sazonalidade na produção, explicada pela existência de períodos de seca (abril a setembro) e de águas (outubro a março), que ocorreram durante a realização dos 20 cortes.

Utilizando-se os valores de delimitações internas (L_1 e U_1) e externas (L_2 e U_2) do diagrama de caixa, foram observadas a ocorrência de 11 *outliers* nos cortes 2, 4, 5, 8, 13, 14, 18 e 20 no gráfico superior (Fig. 1). Os valores de delimitações internas (L_1 e U_1) e externas (L_2 e U_2) do diagrama de caixa são apresentados na Tabela 3.

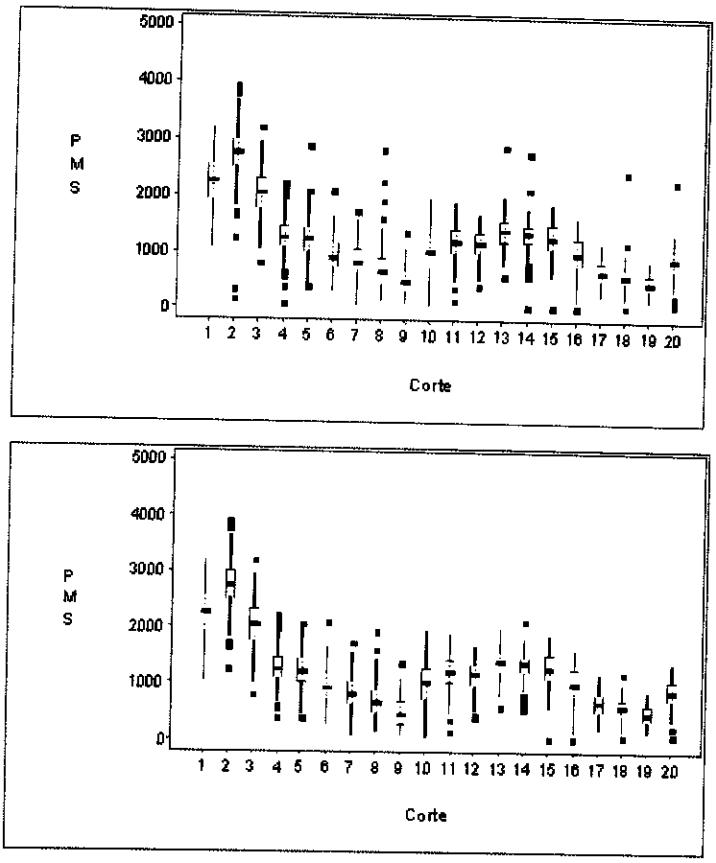

Fig. 1. Diagrama de caixa - com *outliers* (acima) e sem (abaixo).
Fonte: Freitas, arquivo pessoal.

Tabela 3. Valores de delimitações internas (L_1 e U_1) e externas (L_2 e U_2).

Corte	L_1	L_2	U_1	U_2
1	975,06	31,41	3491,46	4435,11
2	1789,75	1073,24	3700,43	4416,93
3	955,04	145,26	3114,44	3924,21
4	610,01	103,77	1959,97	2466,20
5	423,41	-179,08	2030,05	2632,54
6	113,50	-510,79	1778,26	2402,54
7	70,53	-507,21	1611,17	2188,91
8	-147,24	-770,97	1516,04	2139,77
9	-267,40	-828,98	1230,17	1791,75
10	-56,94	-852,33	2064,10	2859,29
11	438,77	-145,57	1997,01	2581,35
12	455,04	-81,78	1886,56	2423,38
13	728,66	237,21	2039,22	2530,68
14	830,20	446,51	1938,56	2238,91
15	522,99	-65,69	2092,79	2681,46
16	126,44	-553,11	1938,56	2618,10
17	99,62	-332,34	1251,50	1683,45
18	96,98	-273,78	1085,66	1456,41
19	45,46	-292,70	947,22	1285,38
20	273,69	-198,54	1532,97	2005,2

Fonte: Freitas, arquivo pessoal.

Uma vez detectado que um dado aberrante é um *outlier*, ele deve ser eliminado para se prosseguir com a análise. A presença de um *outlier* aumenta a variância dos dados, interfere na sua distribuição, provoca heterogeneidade de variâncias e aumento acentuado no coeficiente de assimetria e de curtose. .

Na Tabela 4, pode-se visualizar a influência de *outliers* nos dados de PMS quanto aos valores dos coeficientes de simetria, curtose, coeficiente de variação e no teste de normalidade. O coeficiente de simetria é uma medida da forma da distribuição dos dados com relação à distribuição da curva normal, isto é, se a distribuição tem mais viés em uma direção do que em outra, e assume três valores: zero - a distribuição dos dados tem a forma de sino e tem-se que a média, a mediana e a moda apresentam valores iguais; positivo - a cauda da curva de dados tem viés para a direita e tem-se a média > a mediana > a moda; negativo - a cauda da curva tem viés para a esquerda e a média < a mediana < a moda. Observa-se que a simetria foi negativa para os dados de PMS dos cortes 1 a 4, 10 a 12, 14 a 16, indicando que a cauda desta curva é viesada à esquerda.

Quanto à curtose, ela indica uma medida do grau de achatamento da distribuição normal, podendo assumir três valores: mesocúrtica - os dados têm distribuição normal; nesse caso, o valor da curtose é zero; leptocúrtica – a distribuição é mais pontiaguda do que a normal, e a curtose é positiva; platicúrtica – a distribuição é mais achatada do que a normal, e a curtose é negativa. Observa-se (Tabela 4) que os coeficientes de curtose foram negativos para os dados de PMS dos cortes 1, 9, 10, 12 e 16, mostrando uma distribuição mais achatada (platicúrtica) do que a normal. A curtose é uma medida do grau de achatamento de uma distribuição em relação à curva normal. Segundo Cochran e Cox (1978), a simetria, a curtose e a não-normalidade dos dados produzem estimativas viciadas dos efeitos fixos, interferem no uso dos testes t e F e na heterogeneidade da variância do erro, sendo mais problemáticas em análises multivariadas. Quanto ao fato de a presença de *outlier* aumentar a variância dos dados, verifica-se sensível redução do CV nos dados de PMS em todos os cortes em que estes foram eliminados.

Para verificar se a distribuição normal se ajusta aos dados, geralmente utiliza-se o teste de *Shapiro-Wilks* para amostras menor que 2000. Por meio desse teste usam-se a estatística W ($0< W \leq 1$) e a sua probabilidade ($0< Prob \leq 1$); valores próximos de zero, para ambas, W e Prob, indicam que a distribuição dos dados se afasta da curva normal. Observa-se que, nos cortes em que houve *outliers*, a distribuição normal não se ajusta aos dados; porém, após a retirada dos *outliers*, essa distribuição se ajusta razoavelmente aos dados. O leitor interessado em maiores

detalhes sobre o diagrama de caixa e as estatísticas apresentadas acima pode consultar o módulo *Insight* do SAS (SAS INSTITUTE, 2002–2003).

Análise de Variância – Anova

Uma vez elaborado o modelo estatístico e efetuada a análise exploratória, a etapa seguinte é análise de variância propriamente dita. Vários *softwares* podem ser utilizados para esta tarefa – SPSS, STAT, etc. No entanto, um dos mais versáteis e que será demonstrado aqui é o SAS. Dois procedimentos do SAS que podem ser utilizados em análises de medidas repetidas é o GLM e o MIXED.

Tabela 4. Valores dos coeficientes de simetria, de curtose, testes de normalidade de *Shapiro-Wilk:* S-W e coeficientes de variação.

Corte	Coeficiente de simetria		Coeficiente de curtose		Valor de p S-W		Coeficiente de variação,%	
	Com *outlier*	Sem *outlier*	Com *outlier*	Sem *outlier*	Com *outlier*	Sem *outlier*	Com *outlier*	Sem *outlier*
1	-0,13		-0,20		0,3683		19,19	
2 (2)[1]	-1,55	–0,27	7,09	1,20	<0,0001	0,0350	17,73	14,83
3	-0,15		0,16		0,9148		19,86	
4 (1)	-0,16	0,14	1,82	1,04	0,0032	0,0477	23,34	22,24
5 (1)	0,60	-0,13	3,47	0,23	0,0001	0,8633	25,99	24,23
6	0,38		0,48		0,0466		30,60	
7	0,00		0,10		0,8159		35,02	
8 (2)	1,77	0,60	7,18	0,73	<0,0001	0,0012	51,3 3	44,80
9 (3)	0,58	0,58	-0,22	–0,22	<0,0001	0,0000	51,72	51,72
10	-0,08		-0,55		0,3595		37,45	
11	-0,41		0,47		0,0921		23,76	
12	-0,43		-0,20		0,0166		22,85	
13 (1)	0,72	–0,30	5,02	0,04	<0,0001	0,3061	19,35	17,71
14 (2)	-0,07	–0,51	6,58	1,28	<0,0001	0,0008	20,51	17,55
15	-0,41		0,90		0,0147		22,51	
16	-0,30		-0,23		0,0666		31,31	
17	0,03		-0,52		0,4059		28,95	
18(1)	2,69	-0,07	21,31	0,06	<0,0001	0,6324	38,55	31,68
19	0,21		-0,47		0,1753		30,91	
20(1)	0,61	-0,34	4,54	-0,47	<0,0001	0,1753	28,39	30,91

[1] Número de *outliers* por corte.
Fonte: Freitas, arquivo pessoal.

Uso do procedimento GLM

Para se realizar a análise de variância por meio do GLM, como a descrita na tabela 2, e se ter certeza de que o teste F obtido é exato para os testes de hipóteses, é necessário que os erros e_{ijk} atendam às suposições de independência, normalidade e homogeneidade de variâncias. Para à última suposição, a matriz de variâncias e covariâncias $V(\varepsilon_{ijk}) = R$ precisa atender à condição de esfericidade.

Mas, o que vem a ser esfericidade? Uma exigência inicial para a matriz R ser esférica é a circularidade. Uma matriz de variâncias e co-variâncias é circular quando a diferença entre os valores entre quaisquer duas medidas repetidas tem variâncias constantes. Para o leitor usuário do SAS e interessado em conhecer mais profundamente as restrições do GLM para realizar análises de medidas repetidas, recomenda-se a leitura dos trabalhos de Scheiner e Gurevitch (1993), Littell et al. (1996) e Littell et al. (1998). Quando os dados de medidas repetidas atendem às condições acima, o procedimento GLM pode ser executado conforme o programa 1.

No programa 1, a opção *Title* indica o título da análise; a opção *Proc Glm* indica que o GLM está sendo usado para a análise; a opção *Class* indica que os efeitos de Bloco, Tratamento e Corte são considerados na análise (ver Tabela 2); a opção *Model* mostra o modelo estatístico a ser utilizado (ver Tabela 2). Na opção *Test, H* indica o efeito a ser testado e *E* indica o erro correto para se testar este efeito; por exemplo, o efeito de tratamento é testado com o efeito bloco (tratamento). A opção *Lsmeans* fornece as médias obtidas por quadrado mínimo; a opção *Adjust* produz comparações múltiplas que são ajustadas para os valores de probabilidade p e para os intervalos de confiança para as diferenças de médias consideradas duas a duas; a opção *CI* produz intervalos de confiança das médias; a opção *Pdiff* produz todas as diferenças de médias pareadas. Para utilizar este programa, o arquivo de dados, conforme descrito na Tabela 1, deve conter quatro colunas (Bloco, Tratamento, Corte, PMS). Uma saída parcial do programa 1 é apresentada na Tabela 5, com a probabilidade do teste F (prob > F), indicando que os efeitos de tratamento, corte e interação tratamento x corte são altamente significativos.

```
PROGRAMA 1
TITLE - ANALISE DE VARIANCIA PELO GLM;
PROC GLM;
CLASS BLOCO TRATAMENTO CORTE ;
MODEL PMS= TRATAMENTO BLOCO(TRATAMENTO) CORTE TRATAMENTO
*CORTE/SS3 ;
TEST H= TRATAMENTO E = BLOCO(TRATAMENTO);
LSMEANS TRATAMENTO CORTE /STDERR PDIFF = ALL ADJUST=TUKEY
CL;RUN;
```

Tabela 5. Resultados parciais do programa 1.

Efeito	GL Numerador	Prob > F
Tratamento	91	0,0028
Corte	19	<0,0001
Tratamento*Corte	1728	<0,0005

Fonte: Freitas, arquivo pessoal.

Para propósitos de melhoramento genético, o programa 2, com a opção *Random*, produz uma tabela com a esperança dos quadrados médios - E(Q.M.) para os vários efeitos do modelo (Tabela 6). Observando-se a coluna E(Q.M.), verifica-se que o denominador apropriado para testar tratamento é bloco(tratamento), enquanto o efeito de corte e interação tratamento x corte são testados com o resíduo. As estimativas dos componentes de variância residual (σ^2) e de bloco dentro de tratamentos ($\sigma^2_{B(T)}$) foram, respectivamente, 39945 e 31418, indicando que da variabilidade total existente devida aos fatores aleatórios ($\sigma^2 + \sigma^2_{B(T)}$), 56,0 % é devida ao erro aleatório e 44,0 % ao erro de bloco dentro de tratamentos.

A opção *Test* após *Random* produz testes de hipótese dos efeitos usando os erros apropriados determinados na E(Q.M.), ou seja, os mesmos resultados proporcionados pelo comando *Test H* = Tratamento *E* = Bloco (Tratamento). Observando-se a coluna E(Q.M.) da Tabela 7 verifica-se que os efeitos de corte e da interação tratamento x corte, além dos componentes fixos, contêm uma fração devida à variância residual; o efeito de tratamento, por outro lado, alem dos componentes fixos associados a este efeito, contém uma fração devida aos dois fatores aleatórios (σ^2, $\sigma^2_{B(T)}$).

```
PROGRAMA 2
TITLE PROGRAMA 2- ANALISE UNIVARIADA PELO GLM;
PROC GLM; CLASS BLOCO TRATAMENTO CORTE ;
MODEL PMS=TRATAMENTO BLOCO(TRATAMENTO) CORTE
TRATAMENTO*CORTE /;
RANDOM BLOCO(TRATAMENTO)/TEST; RUN;
```

Tabela 6. Resultados parciais do programa 2.

Causa da Variação	GL	Q.M.	E(Q.M.)	F	P-valor
Tratamento: T	91	1191471	$\sigma^2 + 19{,}848\,\sigma^2_{B(T)} +$ Q(T,T*C)	1,80	0,0028
Bloco(Tratamento): B(T)	92	663529	$\sigma^2 + 19{,}848\,\sigma^2_{B(T)}$	16,61	<0,0001
Corte: C	19	62917117	σ^2 + Q(C,T*Corte)	1575,10	<0,0001
Interação T x C	1728	46761	σ^2 + Q(T*C)	1,17	0,0005
Resíduo	1734	39945	σ^2		

Fonte: Freitas, arquivo pessoal.

Uso do procedimento MIXED

Como já visto, o GLM somente pode ser utilizado em *Anova* de medidas repetidas (MR) em situações restritas. O MIXED, por outro lado, é mais poderoso e possibilita explorar todos os recursos dessa análise.

O fator crucial para se utilizar adequadamente o MIXED é escolher a matriz correta $V(\varepsilon_{ijk}) = R$. Quando a distribuição dos dados é pelo menos aproximadamente normal, o *MIXED* disponibiliza ao usuário cerca de 40 tipos de matriz *R* e ainda fornece três critérios para se selecionar a mais adequada delas (BOZDOGAN, 1987; WOLFINGER, 1993): *AIC* (*Akaike's Information Criterion*), *BIC* (*Bayesian Information Criterion*") e -2*Res Log Likelihood* (-2L).

Quando várias estruturas de *R* são calculadas, aquelas com valores menores para *AIC* são as mais apropriadas. No entanto, para se escolher qual delas apresenta o melhor ajuste, é necessário construir um teste denominado "teste da razão de verossimilhança restrito" para se comparar as matrizes duas a duas. O resultado é um teste de qui-quadrado (χ^2), com graus de liberdade dado pela diferença do número de parâmetros das duas matrizes, o que eqüivale testar a hipótese Ho: As duas matrizes são iguais versus Ha: As duas matrizes são diferentes.

Uma das maneiras de se calcular a matriz R e utilizar os recursos básicos do MIXED para análise de dados é por meio do programa 3 descrito a seguir. Entretanto, uma das dúvidas freqüentes do usuário é como escolher as matrizes R que serão avaliadas e, em seguida, como selecionar a mais adequada. Como já comentado, o *MIXED* disponibiliza ao usuário cerca de 40 tipos de matriz R; destas, três obrigatoriamente devem ser avaliadas - Simetria composta (*CS*), *Huynh-Feldt* (H-F) e Não-estruturada (*UN*). Porém, de acordo com trabalhos com forrageiras e principalmente com alfafa (FREITAS et al., 2007a; FREITAS et al., 2007b), duas outras estruturas também devem ser avaliadas – Auto-regressiva de primeira ordem - *AR(1)*, a Auto-regressiva de primeira ordem de média Móvel - *ARMA(1,1)*. Estas

estruturas podem ser visualizadas no procedimento *MIXED* (SAS INSTITUTE, 2002–2003).

No programa 3, a opção *Repeated Corte* indica que *Corte* é considerado medidas repetidas; a opção sub=bloco (tratamento) indica que a unidade experimental ou indivíduo é representada por bloco(tratamento); na opção *type* é especificada a estrutura da matriz de variância e covariância a ser testada. No exemplo, está sendo avaliada a matriz ARMA(1,1); para calcular cada uma das matrizes UN, H-F, CS, AR(1), etc, em cada execução do programa, substitui-se o nome da matriz na opção *Type*. A opção *Lsmeans* fornece as médias obtidas por quadrado mínimo; a opção *Stderr* e *Pdiff* produz, respectivamente, os erros-padrão da média para tratamento e corte e as probabilidades das diferenças de médias pareadas; a opção *Adjust* produz comparações múltiplas que são ajustadas para os valores de p, para os intervalos de confiança e para as diferenças de médias consideradas duas a duas; a opção *CI* produz intervalos de confiança das médias.

```
PROGRAMA 3
TITLE PROGRAMA 3 – CALCULO DA MATRIZ R = ARMA(1,1);
PROC MIXED;
CLASS BLOCO TRATAMENTO CORTE ;
MODEL PMS= TRATAMENTO CORTE TRATAMENTO*CORTE;
REPEATED CORTE /SUB=BLOCO(TRATAMENTO) TYPE=ARMA(1,1) ;
LSMEANS TRATAM CORTE /STDERR PDIFF = ALL ADJUST=TUKEY CL;RUN;
```

Utilizando-se o programa 3 para calcular as matrizes *AR(1), ARMA(1,1)* e *CS,* os resultados com relação *aos* graus de liberdade, valores de *-2 Res Log Likelihood* e de *AIC* para estas estruturas são apresentados na Tabela 7. Verifica-se que *ARMA(1,1)*, *AR(1) e CS,* pela ordem, são as que representam mais adequadamente os erros entre avaliações dentro de indivíduos, pois são as que possuem os menores valores de *AIC.* Para comparar *AR(1)* e *CS*, uma vez que ambas tem dois parâmetros, não é necessário utilizar o teste de χ^2; a mais adequada é a *AR(1)*, pois tem o menor valor de *AIC*. Para comparar *AR(1)* e *ARMA(1,1)*, tem-se χ^2 = |13957,5 - 14123,8| = 166,3 com grau de liberdade = 2-1=1. Consultando-se χ_1^2 = 166,3 em uma tabela de χ^2, verifica-se que o valor de p é menor que 0,001, sugerindo que a matriz *ARMA(1,1)* é a que melhor ajusta a correlação entre as medidas repetidas. Esta estrutura será a utilizada nos exemplos a seguir. A Fig. 2 apresenta a estrutura de cada uma dessas três matrizes considerando uma dimensão 4 x 4.

Tabela 7. Graus de liberdade, valores de *-2 Res Log Likelihood* e de *AIC* para as estruturas de co-variância avaliadas.

Estrutura	GL	*-2 Res Log Likelihood*	*AIC*
AR(1)	1	14123,8	14127,8
ARMA(1,1)	2	13957,5	13963,5
CS	1	14128,1	14132,1

Fonte: Freitas, arquivo pessoal.

$$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$$

Simetria Composta - cs

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Auto-regressiva de Primeira Ordem - AR(1)

$$\sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$$

Auto-regressiva de Primeira Ordem de Média Móvel - ARMA(1,1)

Fig. 2. Estruturas de variâncias e co-variâncias: CS, AR(1) e ARMA(1,1).

Fonte: Freitas, arquivo pessoal.

A escolha das matrizes *AR(1), ARMA(1,1) e CS,* para serem avaliadas, foi feita com base na literatura (FREITAS et al., 2007a; FREITAS et al., 2007b) e na análise da Fig. 1. Observando-se essa figura, verifica-se que as médias de PMS ao longo dos cortes seguem uma tendência que pode ser explicada, principalmente, pelo efeito de sazonalidade na produção, devido à existência de períodos de seca (abril a setembro) e de águas (outubro a março), que ocorreram durante a realização dos 20 cortes. Nesta situação, matrizes como a Auto-regressiva de Primeira Ordem - *AR(1)*, que contém um parâmetro auto-regressivo (ρ) e a variância residual σ^2 e, a Auto-regressiva de Primeira Ordem de Média Móvel - *ARMA(1,1)*, que, além de ρ e σ^2, inclui o parâmetro que modela o componente de média móvel (γ), são fortes candidatas.

Quanto à Simetria Composta (CS), em que as variâncias são iguais e as co-variâncias também são iguais, observa-se que é uma estrutura que deve ser sempre escolhida para ser testada pelo fato de que atende à condição de circularidade (LITTELL et al., 1998). Se essa matriz for a escolhida, o usuário pode utilizar também o procedimento GLM para analisar os dados.

Procedimento *GLM* versus *MIXED*

A divergência fundamental entre os procedimentos GLM e MIXED está nos erros ε_{ijk} associados a cada indivíduo, conforme descrito no item Análise de variância (*Anova*). Esta diferença reflete no cálculo dos erros-padrão (EP) associados às médias e demais estatísticas derivadas destes, tais como testes de hipóteses, intervalos de confiança, entre outras. No GLM e MIXED, o EP é obtido, respectivamente, da raiz quadrada de $\sigma^2L(X'X)^-L'$ e $L(X'V^{-1}X)^-L'$, em que X é a matriz de especificação, L é a matriz de hipótese, σ^2 é o quadrado médio residual, V a matriz de variâncias e covariâncias, isto é, o EP no GLM é função principal do σ^2, enquanto no MIXED é função de V.

A diferença entre os procedimentos GLM e MIXED quanto aos erros-padrão e estatísticas derivadas destes está ilustrada, respectivamente, nos resultados dos programas 1 e 3. Entretanto, para finalidade didática da comparação dos resultados entre os dois procedimentos, para o efeito de tratamento, foi considerada a análise de dados de PMS de apenas cinco cultivares de alfafa: Barbara, Crioula, P 30, P 5715 e LE N 4.

Nas Tabelas 8, 9 e 10 são apresentadas, respectivamente, as médias e erros-padrão para tratamento e corte dos procedimentos GLM e MIXED. Na organização destas tabelas, foram utilizados os resultados do *Lsmeans, Adjust e Pdiff,* enquanto a opção *CI* foi utilizada na organização da Fig. 3.

Na Tabela 8 são apresentados os valores de Pr > F para os efeitos de tratamentos, cortes e interação tratamentos x cortes; houve significância apenas para o efeito de cortes. Como os dados eram praticamente balanceados, os dois métodos são concordantes quanto às estimativas de efeitos fixos. No entanto, observa-se que os erros-padrão obtidos por máxima verossimilhança (MIXED) são maiores do que os obtidos por quadrados mínimos (GLM). Esta superioridade pode ser também observada nos intervalos de confiança, com 95 % de probabilidade para as médias de tratamentos e de cortes (Fig. 2), principalmente para o efeito de tratamentos.

Tabela 8. Valores do teste F para GLM e MIXED.

Efeito	Prob > F	
	GLM	MIXED
Tratamento	0,8524	0,8932
Corte	<0,0001	<0,0001
Tratamento*Corte	0,9662	0,9563

Fonte: Freitas, arquivo pessoal.

Tabela 9. Médias e erros-padrão para as cinco principais variedades de alfafa obtidas por meio dos procedimentos GLM e MIXED.

Tratamento	GLM	MIXED
LE N 4	1653,1 ± 31,2 a	1653,1 ± 155,33 a
Bárbara	1519,5 ± 31,2 bc	1519,5 ± 155,33 a
Crioula	1532,8 ± 32,0 bc	1534,3 ± 155,47 a
P 30	1565,0 ± 32,0 ab	1565,5 ± 155,47 a
P 5715	1434,2 ± 31,2 c	1434,2 ± 155,33 a

* Médias com letras diferentes na coluna diferem (P < 0,05).

Fonte: Freitas, arquivo pessoal.

Tabela 10. Médias e erros-padrão.

Corte	GLM	MIXED
1	2563,4 + 62,5 b	2563,5 ± 92,1 b
2	3182,5 + 62,5 a	3182,5 ± 92,1 a
3	2616,7 + 62,5b	2616,7 ± 92,1 b
4	1748,9 + 62,5c	1748,9 ± 92,1 c
5	1584,3 + 68,75cde	1586,5 ± 95,7 cdef
6	1342,7 + 62,5df	1342,7 ± 92,1 def
7	1296,1 + 62,5bef	1296,1 ± 92,1 efg
8	1269,2 + 68,7ef	1273,5 ± 95,7 fg
9	928,6 + 62,5gh	928,5 ± 92,1 hi
10	1644,5 + 62,5cd	1644,5 ± 92,1 cd
11	1557,9 + 62,5cde	1557,9 ± 92,1 cdef
12	1406,7 + 62,5def	1406,7 ± 92,1 defg
13	1578,8 + 62,5cde	1578,8 ± 92,1 cdef
14	1565,9 + 62,5cde	1565,9 ± 92,1 cdef
15	1613,2 + 62,5cd	1613,2 ± 92,1 cde
16	1395,5 + 62,5def	1395,5 ± 92,1 defg
17	866,4 + 68,7gh	894,5 ± 92,1 hi
18	819,1 + 62,5h	819,1 ± 92,1 i
19	652,2 + 62,5h	652,2 ± 92,1 i
20	1159,5 + 62,5fg	1159,5 ± 92,1 gh

Fonte: Freitas, arquivo pessoal.

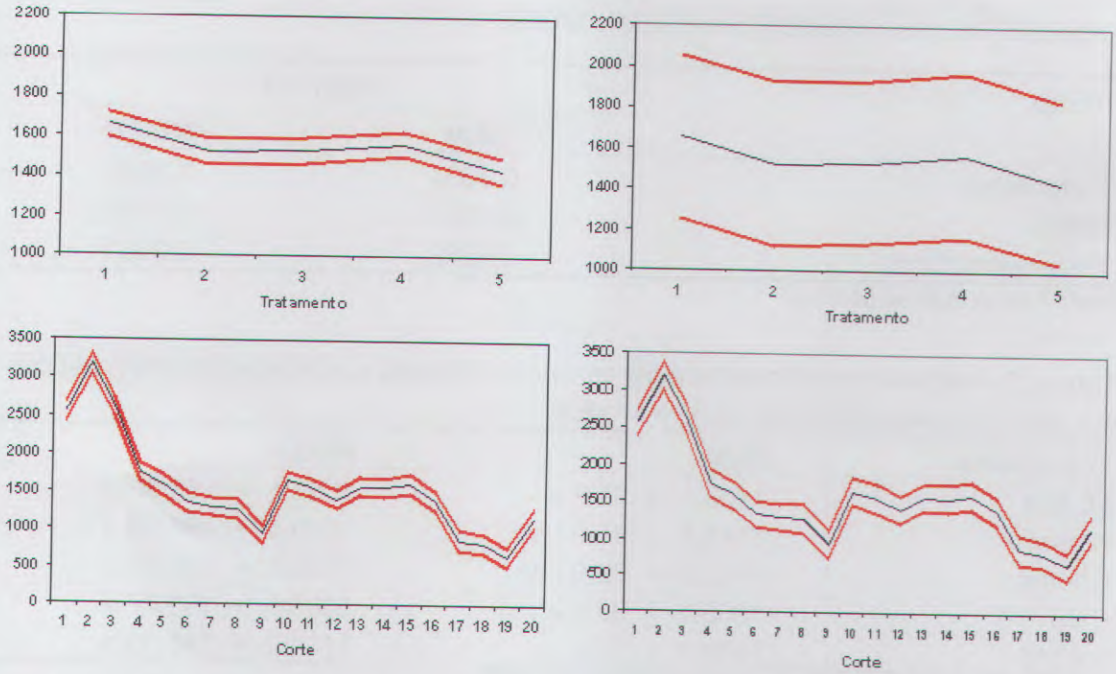

Fig. 3. Intervalos de confiança com 95 % de probabilidade para produção de matéria seca (PMS), kg/ha: as três linhas em ordem decrescente de valor de PMS indicam o limite superior, a média e o limite inferior.

Fonte: Freitas, arquivo pessoal.

Considerações gerais

No presente estudo, foram abordadas análises considerando-se dados de PMS de alfafa. No entanto, as demais características da cultura, como produção e composição da forragem, bem como aquelas associadas ao desempenho animal, como produção e composição do leite, pesos e ganhos de peso, escore corporal, entre outras, podem ser analisadas considerando-se o enfoque de medidas repetidas. Para esta situação, as ferramentas de análises mais apropriadas são os procedimentos GLM e MIXED do SAS. Como foi demonstrado, o GLM é apropriado para ajustar modelos lineares gerais pelo método dos quadrados e permite executar várias análises univariadas, multivariadas por meio da opção *Manova* e de medidas repetidas por meio da opção *Repeated*. Para análises de medidas repetidas, que é o assunto central tratado aqui, o GLM proporciona resultados incorretos e limitados na maioria dos casos, pois produz estimativas inadequadas dos erros-padrão para a maioria das estimativas. O GLM somente produz resultados corretos em análises de medidas repetidas quando a condição de circularidade e esfericidade é atendida, ou equivalentemente, a matriz de variância-covariância dentro de indivíduos é do tipo Simetria Composta (CS). Este fato, no entanto, pode induzir o usuário a resultados equivocados quando usa o procedimento GLM, pois muitas vezes ele encontra que a

matriz CS apresenta um ajuste adequado dos dados dentro de indivíduos, simplesmente porque pelo GLM ele não tem possibilidade de testar outras estruturas. Quando usa o procedimento MIXED, certamente o usuário vai encontrar uma matriz mais apropriada do que a CS, pois este procedimento disponibiliza ao usuário cerca de 40 tipos de estruturas de matriz de covariâncias, quando os dados têm distribuição normal (LITTLE et al. 1996).

Dentre algumas vantagens do MIXED em relação ao GLM em análises considerando o enfoque de medidas repetidas, pode-se citar: a) o GLM não permite a modelagem da estrutura de covariância dos dados, pois seus resultados são incorretos na maioria das vezes, além de ser bastante limitado, principalmente em análise de medidas repetidas. O MIXED seleciona a estrutura de covariância mais adequada aos dados; com isso, elimina-se a necessidade do teste de esfericidade, para saber que análise utilizar: univariada ou multivariada; b) o MIXED usa o método da máxima verossimilhança; o GLM usa quadrado mínimo; c) com o MIXED, os dados de medidas repetidas são analisados na sua forma original; o GLM requer transformação das variáveis originais em ortogonais; d) os dados podem ter estrutura completa ou incompleta; o GLM elimina da análise os indivíduos com dados perdidos; e) geralmente as avaliações adjacentes são mais fortemente correlacionadas que as demais e a variabilidade da resposta dos indivíduos em função do tempo é crescente; o GLM não permite modelar estruturas que atendam este tipo de comportamento dos dados.

Indiscutivelmente, o procedimento MIXED fornece os recursos necessários para os diversos tipos de análises de dados obtidos da cultura da alfafa e também do desempenho animal.

Referências

BOZDOGAN, H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. **Psychometrika**, Baltimore, v. 52, n. 3, p. 345-370, 1987.

CLEVELAND, W.S. **The elements of graphing data**. New Jersey: AT&T Bell Laboratories; Murray Hill, 1994. 297 p.

COCHRAN, W.G.; COX, D.F. **Deseno experimentales**. Mexico:Trillas, 1978. 661 p.

DIGBY, P.G.N.; KEMPTON, R.A. **Multivariate analysis of ecological communities**. London: Chapman & Hall, 1996. 206 p.

FREITAS, A. R. ; DESTEFANI, C. R.; FERREIRA, R. P.; MOREIRA, A. Distribuição de dados de 20 cortes de rendimentos de matéria seca de 92 acessos de alfafa, aplicando análise de medidas repetidas. In: REUNION DE LA ASOCIACION LATINOAMERICANA DE PRODUCCION ANIMAL, 20.; REUNION DE LA ASOCIACION PERUANA DE PRODUCCION ANIMAL, 30.; CONGRESO INTERNACIONAL DE GANADERIA DOBLE PROPOSITO, 5., 2007, Cusco. **Anais...** Cusco: Associacion Peruana de Produccion Animal, 2007. 1 CD-ROM.

FREITAS, A. R. de; BARIONI JUNIOR, W.; FERREIRA, R. de P.; DESTEFANI, C.; SANTOS, A. R. dos; MOREIRA, A. Estudo de medidas repetidas: uma aplicação a dados de forrageiras. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 52.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 12., 2007, Santa Maria, RS. **Resumos...** Santa Maria: UFSM: RBRAS, 2007. 1 CD-ROM.

FREITAS, A. R. de; PRIMAVESI, O.; CORREA, L. A.; PRIMAVESI, O.; POTT, E. B.; MASCIOLLI, A. S. Repeated measurement analyses of forages in cropping systems. In: INTERNATIONAL GRASSLAND CONGRESS, 19., 2001, Piracicaba. **Proceedings...** Piracicaba: Fealq, 2001. p. 1046-1047.

GOWER, J.C.; HAND, D.J. **Biplots**. London: Chapman & Hall, 1996. 277 p.

LITTELL, R. C.; HENRY, P. R.; AMMERMAN, C. B. Statistical analysis of repeated measures data using SAS procedures. **Journal of Animal Science**, Champaign, v. 76, p. 1216-1231, 1998.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D. **SAS System for Mixed Models**. Cary: Statistical Analysis System Institute, 1996. 633 p.

REIEZIGEL, J. Analysis of experimental data with repeated measurement. **Biometrics**, Washington, v. 55, p. 1059-1063, 1999.

SAS INSTITUTE. **User's Guide**. versão 9.1.3, versão para Windows. Cary: 2002-2003.

SCHEINER, S. M.; GUREVITCH, J. The design and analysis of ecological experiments. New York: Chapman & Hall, 1993. 445 p.

WOLFINGER, R. Covariance structure selection in general mixed models. **Communications in Statistics - Simulation and Computation**, New York, v. 22, n. 4, p.1079-1106, 1993.