DETERMINAÇÃO DA ASSOCIAÇÃO ENTRE AS VARIÁVEIS DO BANCO DE DADOS DA PRODUTIVIDADE DE ALGUNS PRODUTOS AGRÍCOLAS PRODUZIDOS NO BRASIL – UMA PREPARAÇÃO PARA A ANÁLISE FATORIAL

JOSÉ C. SOBRINHO[1], JADIR N. SILVA[2], DIOGO J. CARDOSO[3], RICARDO HOHER[4]

[1] Engenheiro Agrícola, Prof., Associado, UFSM – Campus Silveira Martins. Depto. Multidisciplinar, Silveira Martins – RS. Tel. 55 3224 4701, jcardosos@yahoo.com.br
[2] Matemático, Professor Titular – Departamento de Engenharia Agrícola – DEA/UFV – Viçosa – MG.
[3] Acadêmico do Curso de Ciência da Computação. Campus Central da UFSM – Santa Maria – RS
[4] Contador, Professor Assistente – UFSM – Campus Silveira Martins. Depto. Multidisciplinar, Silveira Martins – RS.

Apresentado no
XLIV Congresso Brasileiro de Engenharia Agrícola - CONBEA 2015
13 a 17 de setembro de 2015- São Pedro – SP, Brasil

RESUMO: Objetivou-se neste trabalho verificar se as p-variáveis dos dados de produtividade em toneladas por hectare de cevada, centeio, mamona, soja e trigo são correlacionados de algum modo. Foi avaliada a produtividade dos produtos agrícolas desde o ano de 1976 até 2014, perfazendo um total de 39 anos analisadas. Fez-se o teste de suposição da normalidade para a estatística univariada utilizando o teste de Shapiro-Wilk. Depois se utilizou o de teste de esfericidade de Bartlett para a matriz de correlação, onde foi confrontada a hipótese da matriz P_{pxp} ser igual à matriz identidade contra a hipótese alternativa da matriz P_{pxp}, ser diferente da matriz identidade, onde P_{pxp} é a matriz de correlação teórica das p-variáveis. Pela estatística univariada concluiu-se que todas as variáveis apresentaram distribuição normal, pelo teste de Shapiro-Wilk. Em relação ao teste de esfericidade de Bartlett para a matriz de correlação, observou-se que as variáveis não são mutuamente independentes, ou seja, são correlacionadas de algum modo, rejeitando-se a hipótese da matriz P_{pxp} ser igual à matriz identidade o que permite o ajuste do modelo de análise fatorial aos dados.
PALAVRAS–CHAVE: esfericidade, estatística, grãos.

DETERMINATION OF THE ASSOCIATION BETWEEN THE PRODUCTIVITY DATABASE VARIABLES OF SOME AGRICULTURAL PRODUCTS PRODUCED IN BRAZIL - A PREPARATION FOR THE FACTOR ANALYSIS.

ABSTRACT: the objective of this work was to verify if the p-variables of tons per hectare productivity data of barley, rye, castor beans, soybeans and wheat are correlated in some way. The productivity of agricultural products was evaluated from 1976 to 2014, a total of 39 years analyzed. It was made the supposition of normality test for the univariate statistical using the Shapiro-Wilk test. Then it was used the Bartlett's sphericity test for the correlation matrix, where it faces the hypothesis of the P_{pxp} matrix being equal to the identity matrix against the alternative hypothesis of the P_{pxp} matrix being different from the identity matrix, where P_{pxp} is the theoretical correlation matrix of the p-variables. The univariate statistical concluded that all variables were normally distributed, by the Shapiro-Wilk test. In relation to the Bartlett sphericity test for the correlation matrix, it was observed that the variables are not mutually independent, that is, they are correlated in some way, rejecting the hypothesis P_{pxp} matrix being equal to the identity matrix which allows the adjustment of the factor analysis model to the data.
KEYWORDS: sphericity, statistics, grains.

INTRODUÇÃO: A agricultura é um segmento de grande importância para o País desde o início da colonização no século XVI, a economia brasileira dependeu dela até metade do século XX, o que gerava grande dependência de alguns produtos como café, cana de açúcar, borracha, cacau e algodão (FURTADO, 2005; BAER, 2008). A produtividade sempre foi uma variável problemática na produção agrícola brasileira, embora não esteja relacionado à expansão da fronteira agrícola nacional, e sim com implantação de alta tecnologia na produção agrícola, seja em expansão no cerrado ou em qualquer região do país. Pesquisas de alto nível têm sido desenvolvidas na agricultura, sejam em nível de aplicativos, peças mecânicas e até mesmo em grandes empreendimentos em genética, ou outras atividades voltadas para o agronegócio brasileiro. Os parâmetros de cadeias produtivas são avaliados de diferentes formas, sejam no levantamento de dados ou análises de parâmetros de interesse ao mercado agropecuário. A estatística multivariada é importante na avaliação de cadeias produtivas, destacando-se o ajuste do modelo de análise fatorial aos dados de uma distribuição qualquer, em que se pressupõe que as variáveis respostas sejam correlacionadas de alguma forma entre si. Desse modo, quando as variáveis são provenientes de uma distribuição normal p-variada, é possível fazer o teste de hipótese para verificar se a matriz de correlação populacional é próxima ou não da matriz identidade, o teste que verifica esta proximidade é o teste de esfericidade de Bartlett. Ressalta-se que para a aplicação do teste de Bartlett há exigência que as variáveis envolvidas na análise tenham distribuição normal p-variada (MINGOTTI, 2013). Diante disto avaliou-se o banco de dados da produtividade de cevada, centeio, mamona, soja e trigo com vistas à detecção ou não da normalidade das distribuições dos dados.

MATERIAL E MÉTODOS: foi levantado no site da CONAB (2015) o parâmetro produtividade em quilogramas por hectare de cevada, centeio, mamona, soja e trigo. Avaliou-se a produtividade dos produtos agrícolas desde o ano de 1976 até 2014, perfazendo um total de 39 anos analisadas. Fez-se o teste de suposição da normalidade para a estatística univariada utilizando o teste de Shapiro-Wilk, cujas hipóteses foram: H_0: a amostra provém de uma população Normal e H_1: a amostra não provém de uma população Normal. Ocorrerá a rejeição de H_0 ao nível de significância 5%, se W_{calc} for menor que $W_{5\%}$ tabelado com 39 graus de liberdade. O teste de esfericidade de Bartlett para a matriz de correlação, cuja estatística T, foi determinada por:

$$T = -\left(n - \frac{1}{6}(2p + 11)\right)\left(\sum_{j=1}^{p} \ln(\hat{\lambda} i)\right) \quad \text{.....} \quad (1)$$

em que:
T = estatística do teste T;
n = número de elementos avaliados em cada variável no banco de dados;
p = quantidade de variáveis avaliadas;
$\hat{\lambda} i$ = autovalores da matriz de correlação amostral, i variando de 1 a 5; determinado pelo aplicativo Statistica;
As hipóteses testadas foram:
$H_o : P_{pxp} = I_{pxp}$;
$H_1 : P_{pxp} \neq I_{pxp}$
em que,
H_o : as variáveis são independentes;
H_1 : as variáveis não são independentes;
P_{pxp} é a matriz de correlação teórica das cinco variáveis e
I_{pxp} é a matriz identidade.
A estatística T tem distribuição aproximadamente qui-quadrado com $\frac{1}{2}p(p-1)$ graus de liberdade. Então, rejeita-se H_o se o valor observado de T for maior ou igual ao valor crítico da distribuição qui-quadrado para o nível de significância 5% e p=5 com n=39.

RESULTADOS E DISCUSSÃO: Na Tabela 1, apresentam-se os parâmetros relativos à produtividade em quilogramas por hectare para as variáveis cevada, centeio, mamona, soja e trigo, perfazendo um total de cinco variáveis, os dados estão apresentados da forma que foram coletados na página do banco de dados da CONAB (2015) desde o ano de 1976 até 2014. Ao utilizar-se o teste de

Shapiro Wilk, 1965, verificou-se que as distribuições das variáveis apresentaram distribuição normal, conforme Tabela 2. Na Tabela 3 tem-se a matriz de correlação entre as variáveis avaliadas, cujos valores da matriz P_{pxp} são apresentados, observa-se que as variáveis mais correlacionadas são trigo e cevada com coeficiente de correlação igual a 0,868828, em segundo lugar, têm-se as variáveis cevada e centeio com correlação amostral igual a 0,782434; em terceiro a correlação entre soja e cevada cujo valor foi igual a 0,768526. A correlação amostral de menor valor ocorreu entre mamona e centeio com valor igual a – 0,24086.

Tabela 1 - Parâmetros relativos à produtividade em quilogramas por hectare para as variáveis cevada, centeio, mamona, soja e trigo para os anos de 1976 até 2014.

Distribuição dos dados para os anos de 1976 até 1995					Distribuição dos dados para os anos de 1996 até 2014				
Cev	Cent	Man	So	Tr	Cev	Cent	Man	So	Tr
1018,1	912,0	806,3	1747,7	655,1	1939,2	783,5	642,7	2299	1604
1609,6	890,2	1140,9	1250,1	953,3	1923	802	141,8	2384	1593,1
1176,9	787,0	929,07	1251,3	734,0	2306,2	1269,8	334,8	2367	1919,4
1148,6	860,6	687,6	1700,2	878,84	2117,4	971,4	549,6	2395	1130
1460,8	899,0	592,9	1781,2	1048,7	2012,8	1194,4	495	2751	1868
628,5	64,9	429,3	1535,8	651,7	1524	1055	574	2567	1420
1114,2	937,5	594,9	1727,6	1134,2	2700	1308	673	2816	2123
956,2	738,0	541,2	1674,1	1008,0	2678	1346	646	2329	2375
1295,3	984,1	810,3	1807,7	1654,0	2762	1308	975	2245	2121
1671,7	843,1	616,6	1369,4	1441,0	2795	1535	703	2419	2063
1842,4	1300	386,8	1851,2	1786,4	2287	1372	602,0	2822,6	1176
1297,4	1034,4	677,7	1693,0	1675,0	2692	1343	758	2816	2170
2189,4	853,6	453,1	1952,9	1656,8	2989	1298	587	2629	2088
2000,9	1097,5	489,4	1740,1	1006,3	2599	1333	637	2927	2070
2126,4	1442,3	560,1	1580	1434	3230	1333	644	3115	2736
2147,9	1400	641,9	2027	1371	3451	1522	193	2651	2672
1935,4	1153,8	276,1	2150	1250	3510	1800	179,9	2938,3	2502
1926,3	1195,1	537,3	2179	1478	2606	1944	441,2	2854,2	2162
2137,4	1185,2	569,6	2221	1474	2606	1944	324,8	3033,3	2162
2526,4	1388,9	391,8	2175	1745					

Cev = cevada; Cent = centeio; Mam = mamona; So = soja; Tr = trigo.

Tabela 2 – Resultado da análise pelo teste de Shapiro-Wilk, 1965.

Estatística W	*CENTEIO*	*CEVADA*	*MAMONA*	*SOJA*	*TRIGO*
W Calculado	0,951898257	0,976541294	0,9705033	0,95255378	0,967667965
W tabelado (5%, 39)	0,939	0,939	0,939	0,939	0,939

Os parâmetros W calculados do teste de Shapiro Silk, 1965 apresentaram valores maiores que o valor crítico tabelado a 5% de probabilidade com 39 elementos avaliados, indicando a normalidade das distribuições das variáveis avaliadas.

Tabela 3 – Matriz de correlação populacional para as variáveis centeio, cevada, mamona, soja e trigo.

	CENTEIO	*CEVADA*	*MAMONA*	*SOJA*	*TRIGO*
CENTEIO	1	0,782434	-0,24086	0,666292	0,711347
CEVADA	0,782434	1	-0,2795	0,768526	0,868828
MAMONA	-0,24086	-0,2795	1	-0,36616	-0,27762
SOJA	0,666292	0,768526	-0,36616	1	0,750603
TRIGO	0,711347	0,868828	-0,27762	0,750603	1

Os autovalores da matriz R_{pxp} da Tabela 3 são iguais a 3,141762; 0,876296; 0,329222; 0,256555 e 0,120312 e o valor da estatística T é igual a:

T = – (39 – $\frac{1}{6}$* (2 * 5 + 11)) * (ln(3,141767) + ln(0,876296) + ln(0,329222) + ln(0,256555) + ln(0,120312)) = 126,96

O valor crítico da distribuição qui-quadrado com 5% de probabilidade e $\frac{1}{2} * 5 * (5-1)$ = 10 graus de liberdade é 18,307 e, portanto, menor que a estatística T calculada, cujo valor foi de 126,96. Rejeita-se H_o, uma vez que o valor calculado de T foi maior que o valor crítico da distribuição qui-quadrado para o nível de significância 5% e p=5 com n=39. Estatística

CONCLUSÕES: todas as variáveis apresentaram distribuição normal pela estatística W; o teste de esfericidade de Bartlett para a matriz de correlação indicou que as variáveis não são mutuamente independentes e estão correlacionadas de algum modo.

REFERÊNCIAS:

BAER, W. **The Brazilian economy.** Boulder - USA, Lynne Rienner Publishers, 2008, 6th edition 443p.

CONAB (COMPANHIA NACIONAL DE ABASTECIMENTO). **Acompanhamento da safra brasileira**, 8° Levantamento maio/2015. Disponível em: <http//www.conab.gov.br>. Online. Acesso em: 25 maio. 2015.

FURTADO, C. **Formação econômica do Brasil**. São Paulo: Companhia Editora Nacional, 2005. 32ª ed. 256p.

MINGOTTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**, Belo Horizonte. Editora UFMG, 2013, 2ª Ed. 303p.

SHAPIRO, S. S.; WILK, M.B. An Analysys of Variance Test for Normality (complete samples), **Biometrika**, London, v.52. n. 3-4, p. 591-611, Dec: 1965.