USO DA ANÁLISE DE COMPONENTES PRINCIPAIS PARA A CRIAÇÃO DE CLUSTERS COMO MECANISMO DE DIVERSIFICAÇÃO DE CARTEIRA DE ATIVOS DO SETOR AGROINDUSTRIAL

LUIZ FERNANDO OHARA KAMOGAWA; RICARDO MENDONÇA FONSECA; JOSÉ CÉSAR CRUZ JÚNIOR; VITOR AUGUSTO OZAKI;

ESALQ/USP

PIRACICABA - SP - BRASIL

lfokamog@esalq.usp.br

APRESENTAÇÃO SEM PRESENÇA DE DEBATEDOR

COMERCIALIZAÇÃO, MERCADOS E PREÇOS AGRÍCOLAS

Uso da análise de componentes principais para a criação de *clusters* como mecanismo de diversificação de carteira de ativos do setor agroindustrial

Grupo de Pesquisa: 1 – Comercialização, Mercados e Preços Agrícolas

RESUMO

O presente trabalho teve como objetivo a criação de *clusters* de comportamento de ativos relacionados à agroindústria (derivativos agropecuários e ações BOVESPA) baseados na análise de componentes principais; utilizando este resultado como ferramental de tomada de decisão na construção de carteiras de investimento. Isso, baseado no princípio que a redução da volatilidade da carteira via diversificação depende da baixa covariância entre eles, ou seja, que os ativos pertençam a diferentes *clusters*.

Palavras-chave: Análise de componentes principais, análise de *cluster*, diversificação de carteira, derivativos agropecuários e BOVESPA.

ABTRACT

The present paper had as a main objective the creation of some behavior clusters of assets related to the agribusiness (agriculture derivatives and BOVESPA stocks) based on the principal components analysis; using this result as a tool of asset portfolio composition.

This decision is based upon the principle that the decrease of portfolio volatility depends on the low covariance between the assets, or in other words, they must belong to different clusters.

Key-words: Principal components analysis, cluster analysis, portfolio diversification, agriculture derivatives and BOVESPA.

1. INTRODUÇÃO

A diversificação de carteira é um eficiente instrumento de redução de volatilidade e risco (Ross et al, 2002). Um exemplo da eficiência da diversificação é a comparação dos desvios-padrão do retorno diário de alguns ativos individuais com o indicador IBRX e FGV 100 (ambos indicadores baseados em uma carteira de ações da BOVESPA) (Tabela 1).

Tabela 1. Desvios-padrão do retorno diário para os índices IBRX e FGV 100 e para alguns ativos selecionados membros destes índices, entre 5/8/2004 a 24/3/2006.

Ativo	Desvio-padrão
IBRX	0,0131
FGV 100	0,0124
Unibanco PNT (UBBR11)	0,0499
Votorantim Celulose e Papel PN (VCPA4)	0,0437
Petrobrás PN (PETR4)	0,0415
Vale do Rio Doce ON (VALE3)	0,0385
Klabin PN (KLBN4)	0,0252
Perdigão PN (PRGA4)	0,0203

Fonte: Elaborado pelos autores com dados de PLADIN (2006)

Esta diversificação, no entanto, só é eficiente para situações onde os ativos tenham uma correlação, pelo ou menos, inferiores a 1 (Ross et al, 2002). Quanto maior for a correlação ou covariância entre os ativos que compõe esta carteira, menor é a eficiência desta diversificação.

Dados estes princípios, o uso dos componentes principais, uma vez que eles são baseados na decomposição da informação da matriz de covariância/variância, e a análise de clusters é adequada como ferramental para a tomada de decisão para a redução da volatilidade de uma carteira de investimentos via diversificação.

Topaloglou et al (2002) utilizou a mesma metodologia adotada no estudo (análise de componentes principais associadas à analise de clusters) e obteve bons resultados. Comparativamente a uma carteira tendo como critério apenas as médias ponderadas, foi obtido um resultado *ex-post* com maior retorno e menor variância.

Aguillera et al (1999) utilizou os componentes principais para analisar o comportamento de ações da Bolsa de Madri. Neste caso, entretanto, o interesse não era

diversificar a carteira, mas encontrar as características que levavam a um determinado comportamento futuro para prever preços.

2. OBJETIVOS

O presente trabalho tem como objetivos:
1) Obter os componentes principais da matriz de variância/covariância das variações de ativos relacionados à agroindústria;
2) Criação de grupos (clusters) de comportamento destes ativos a partir dos componentes principais;
3) Extrair destes resultados conclusões úteis para a composição de uma carteira de investimentos com ativos relacionados à agroindústria; e,
4) Verificar a possibilidade do uso deste ferramental como utensílio *ex-ante* no processo de tomada de decisões para compor uma carteira de investimento de menor variância.

3. METODOLOGIA

3.1 Componentes principais

Os componentes principais foram concebidos para explicar a matriz de variância/covariância com o mínimo de combinações lineares possíveis, um procedimento que permite a redução de dimensões de um sistema de equações (Johnson & Wichern, 2002). Uma técnica que cria novas variáveis que são uma combinação linear das originais (Sharma, 1996), que permite a redução da representatividade de uma matriz de variância/covariância de uma dimensão *p* original para uma, duas ou no máximo três dimensões (Johnson & Wichern, 2002; Sharma, 1996). Um critério muito utilizado no processo de tomada de decisão em investigações complexas, mas não seu meio final (Johnson & Wichern, 2001).

Algebricamente, como a distância total multivariada normalizada (distância de Mahalanobis ou Euclidiana) de um dado vetor $\boldsymbol{X}' = \left[X_1; X_2; \ldots; X_p \right]$ (onde: $X'_p = \left[x_{pt}; \ldots; x_{pT} \right]$) é dada pela equação (1) (Johnson & Wichern, 2002):

$$\left(X - \mu \right)' \Sigma^{-1} \left(X - \mu \right) \leq \chi^2_p \left(\alpha \right) \qquad (1)$$

Em que:

Σ é a matriz de variância/covariância do vetor ***X***;

μ é o vetor de médias de X_p; e,

p é o número de variáveis/dimensões.

É possível decompor esta distância total em um sistema de autovalores $\left(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \right)$ e seus autovetores correspondentes $\left(e_i; \ldots; e_p \right)$ da matriz Σ, dado que esta distância também pode ser representada pela forma (2) (Johnson & Wichen, 2002):

$$\left(X - \mu \right)' \Sigma^{-1} \left(X - \mu \right) = \sum_{i=1}^{p} \left(\frac{1}{\lambda_i} \right) \left(X' e_i \right)^2 \geq 0 \qquad (2)$$

Considerando a combinação linear (3):

$$\begin{aligned} Y_1 &= X'e_1 \\ Y_2 &= X'e_2 \\ &\vdots \\ Y_p &= X'e_p \end{aligned} \qquad (3)$$

Obtém-se que: $\text{var}(Y_i) = e_i'\Sigma e_i$ e $\text{cov}(Y_i, Y_k) = e_i'\Sigma e_k \quad (i,k = 1,2,\ldots,p)$. Assim, os componentes principais (quantidade de dimensões) serão aqueles capazes de aproximar ao máximo o valor de $\text{var}(Y_i)$ e $\text{cov}(Y_i, Y_k)$ (Johnson & Wichern, 2002). Dado que: $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \text{var}(Y_i)$, a quantidade de componentes principais pode ser dada tendo como critério o número de autovalores necessários que maximize a proporção (4) também conhecida como inércia parcial (Johnson & Wichern, 2002):

$$proporção = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, \; k \leq p. \qquad (4)$$

Se o primeiro, os dois primeiros, ou os três primeiros autovalores representarem grande parte desta proporção, estes componentes podem substituir as *p* dimensões originais sem grandes perdas de informação (Johnson & Wichern, 2002).

3.2 Processo de criação dos *clusters*

Os componentes principais podem ser utilizados para análises posteriores. Um dos possíveis usos é a criação de clusters a partir dos vetores criados, tendo como vantagem que os vetores criados são ortogonais entre eles (Sharma, 1996).

O processo de criação dos *clusters* é um processo de agrupamento baseado em critérios de similaridade ou distância (dissimilaridade). Estes critérios envolvem a necessidade de uma medida para esta similaridade, podendo ela ser simples (como o uso de critérios como a análise visual ou intuitiva) ou mais complexas (com o uso de critérios algébricos de medidas de distância como a distância estatística Euclidiana, a medida de Canberra e o coeficiente de Czekanowski) (Johnson & Wichern, 2002).

Foram utilizados dois processos no presente trabalho. Do conjunto inicial de autovetores foram identificados alguns *clusters* de comportamento pela simples análise gráfica e intuitiva. Destes *clusters* pré-identificados foram retirados aqueles que não faziam parte do grupo principal. Deste grupo principal, foi aplicada a metodologia não hierárquica de formação de *clusters* conhecida como *k-means*.

MacQueen (1967) citado por Johnson & Wichern (2002) sugeriu o termo *k-means* para descrever o algoritmo que classifica cada item como membro do cluster com o centróide mais próximo. De forma que, este centróide a partir da retirada de um membro de um grupo e realocação em outro (estimando-se novos centróides) em um processo contínuo até que ocorra a minimização da variância total dos membros em relação a estes centróides (Johnson & Wichern, 2002).

O número ótimo de *clusters* são aqueles em que seus membros possuem a maior similaridade média (todos os grupos) possível (conhecido como método de Wald), e ao

mesmo tempo tenham maior dissimilaridade média em relação a outros grupos (critério de *average linkage*) (Johnson & Wichern, 2002; Sharma, 1996). A escolha do número *k* de *clusters* a serem estimados foi feita utilizando ambos os critérios adotando como medida de distância a distância estatística Euclidiana.

3.3. Banco de dados

Foi 12 o número de ações negociadas na BOVESPA cujo principal ramo da empresa tenha relação com o setor agroindustrial (Tabela 2). O critério de escolha destas empresas foi fundamentado em: 1) ser membro do índice IBRX BOVESPA e 2) possuir uma série histórica suficiente para análise. Em função do critério 2 foi excluída da análise as ações do grupo Cosan (CSAN3), que apesar de ser um dos ativos mais líquidos e pertencente ao IBRX não foi considerada no estudo.

Foi 6 o número de derivativos agropecuários: 1 negociada na BM&F e 5 negociadas em bolsas de mercadorias norte-americanas (Tabela 3). Os critérios de escolha foram os mesmos das ações acima.

Adicionalmente, foram incluídas as 9 ações mais líquidas negociadas na BOVESPA (consequentemente membros do IBRX) e mais uma do varejo de alimentos (supermercados) (Tabela 4). E cinco indicadores adicionais (dólar comercial, IBRX, índice FGV 100 BOVESPA, risco Brasil mood's e CDI).

Tabela 2. Relação de empresas agroindustriais cujas ações são negociadas na BOVESPA utilizadas no estudo e informações adicionais.

Empresa (CÓDIGO)	(%) movimentação diária do IBRX	Principal ramo de atividade
Ambev PN (AMBV4)	3,734	Cervejas e refrigerantes
Aracruz PNB (ARCZ6)	0,894	Papel e celulose
VCP PN (VCPA4)	0,536	Papel e celulose
Sadia S/A PN (SDIA4)	0,527	Carnes e derivados
Natura ON (NATU3)	0,493	Produtos de uso pessoal
Souza Cruz ON (CRUZ3)	0,472	Cigarros e fumos
Klabin S/A (KLBN4)	0,426	Papel e celulose
Perdição PN (PRGA4)	0,344	Carnes e derivados
Suzano Papel (SUZB5)	0,253	Papel e celulose
Duratex PN (DURA4)	0,170	Madeira
Fosfértil PN (FFTL4)	0,135	Fertilizantes e defensivos
Ripasa PN (RPSA4)	0,115	Papel e celulose

Fonte: BOVESPA (2006)

Tabela 3. Relação de derivativos utilizados no estudo.

Derivativo	Bolsa negociada
Boi gordo	BM&F (Brasil)
Soja	CBOT (Chicago)
Trigo	CBOT (Chicago)
Açúcar	NYBOT (Nova Iorque)
Café	NYBOT (Nova Iorque)
Milho	CBOT (Chicago)

Fonte: PLADIN (2006)

Tabela 4. Relação das empresas de alta liquidez negociadas na BOVESPA e uma do varejo, utilizados no estudo e informações adicionais.

Nome da empresa (CÓDIGO)	(%) movimentação diária do IBRX	Principal ramo de atividade
Petrobras PN (PETR4)	12,303	Exploração e refino petróleo
Petrobras ON (PETR3)	9,358	Exploração e refino petróleo
Vale do Rio Doce PNA (VALE5)	7,481	Minerais metálicos
Bradesco PN (BBDC4)	6,718	Banco
Itaúbanco PN (ITAU4)	6,119	Banco
Vale do Rio Doce ON (VALE3)	6,025	Minerais metálicos
Unibanco UNT (UBBR11)	3,329	Banco
Itausa PN (ITSA4)	2,596	Banco
Bradesco ON (BBDC3)	2,491	Banco
Pão de Açúcar-CBD PN (PCAR4)	0,853	Alimentos

Fonte: BOVESPA (2006)

A série de cotações e valores foi diária no período compreendido entre 5/8/2004 e 24/3/2006. Das séries originais foi extraída a variação diária (comportamento) e desta variação diária foi feita a análise de componentes principais e de *cluster*. A fonte de dados foi o PLADIN (2006).

4. RESULTADOS

Pelo critério da inércia parcial, o uso de apenas dois componentes principais não foi suficiente para representar as informações contidas na matriz de variância/covariância. Com isso, para uma maior segurança foi adicionado um terceiro componente principal, de forma que os três principais representam 67% da inércia total de Σ.

Pela análise visual, de imediato, foram identificados 5 *clusters* de comportamento (Figura 1) (Anexo). Dos *clusters* formados, foram retirados os quatro não pertencentes ao grupo principal (1-VALE3 e VALE5; 2-UBBR11 e ITAU4; 3-PETR3 e PETR4; e, 4-BBDC3 e BBDC4) e foi feita nova análise.

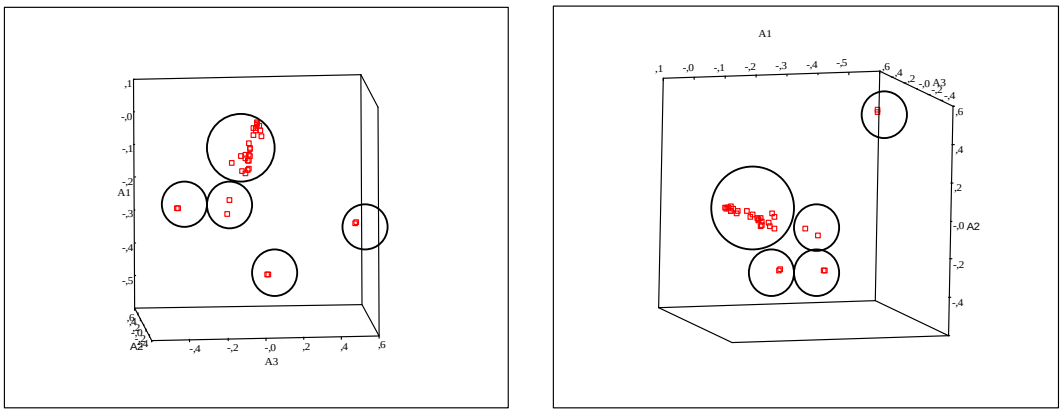

Figura 1. Resultado gráfico dos componentes principais para a série de valores de interesse do estudo e os *clusters* de comportamento para diferentes perpectivas.

Fonte: Elaborado pelos autores.

Do grupo principal, graficamente é possível notar que existe um grupo de ações ou derivativos mais associados ao risco Brasil, câmbio e juros; e outro mais associado aos índices IBRX e FGV 100 (Figura 2).

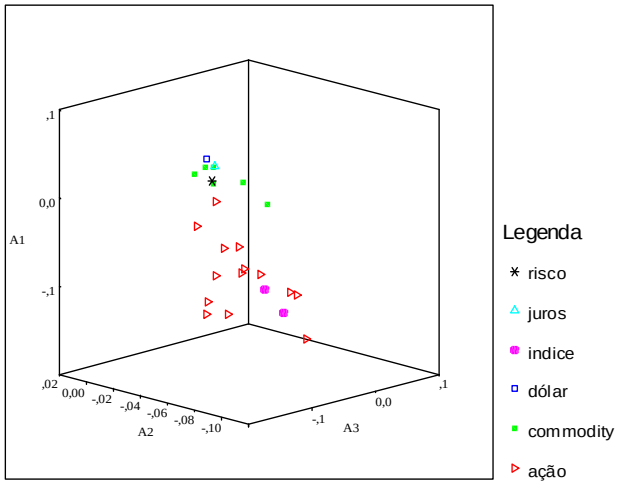

Figura 2. Resultado gráfico reduzido dos componentes principais para a série de valores de interesse do estudo.
Fonte: Elaborado pelos autores.

Foram adotados três cenários para o número de *clusters* (*k*=2, 3 e 4). Pelo critério *k-means* os *clusters* foram formados segundo a (ANEXO 1). E seus respectivos centróides são dados pela (Tabela 5):

Tabela 5. Centróides por número de *clusters* e por *cluster*.

Número de *clusters* (*k* correspondente)	λ_1	λ_2	λ_3
2 (*k*=1)	-0,01204	-0,00198	0,00180
(*k*=2)	-0,009978	-0,05494	-0,05776
3 (*k*=1)	-0,13150	-0,01872	-0,04832
(*k*=2)	0,00947	0,00834	0,00855
(*k*=3)	-0,10671	-0,07320	-0,13198
4 (*k*=1)	0,00947	0,00834	0,00855
(*k*=2)	-0,13150	-0,01872	-0,04832
(*k*=3)	-0,10671	-0,0732	-0,13198
(*k*=4)	-0,13641	-0,09951	-0,06479

Fonte: Elaborado pelos autores.

Pelo critério de Wald, o número ótimo de *clusters* são 3 (Tabela 5). Já pelo critério de *average link*, 4 seria o número ótimo de *clusters* (Tabela 6). Entretanto, é observado pela formação dos componentes principais (Tabela 5) que os centróides dos *clusters* entre 3 e 4 componentes principais são os mesmos, a exceção do *cluster* criado (Tabela 5). Dado que o quarto *cluster* criado entre 3 e 4 componentes inclui apenas um membro (ANEXO 1), foi optado por 3 *clusters* (Figura 3).

Tabela 6. Distâncias Euclidianas dentro (critério de Wald) e entre (*average linkage*) *clusters* por número de *clusters*.

Número de *clusters*	Dentro (*within*)	Entre (*between*)

2	0,1450	0,0076
3	0,1373	0,0165
4	0,1623	0,0332

Fonte: Elaborado pelos autores.

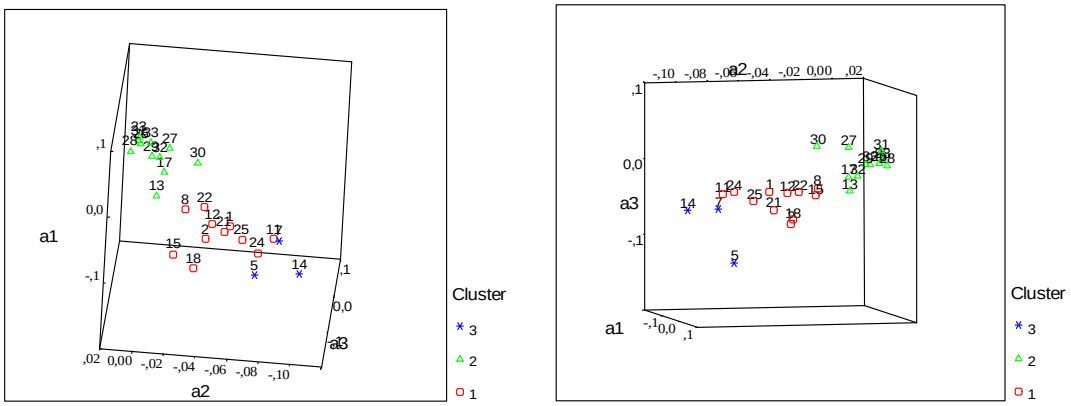

Figura 3. Resultado gráfico reduzido dos componentes principais para a série de valores de interesse do estudo e os *clusters* de comportamento (*k*=3) sob diferentes perspectivas.

Fonte: Elaborado pelos autores.

A formação destes *clusters* indica uma separação de comportamento entre as ações e as commodities (ANEXO 1). De um lado, as ações (na sua maioria) estão, naturalmente, mais relacionadas aos indicadores IBRX e FGV 100. Do outro, as commodities estão melhor relacionadas com ao risco, ao câmbio e aos juros. A única ação pertencente a este grupo são as ações da Fosfértil (FFTL4), que hipoteticamente, por ser uma empresa química com direcionamento ao setor agrícola, está diretamente relacionadas ao preço das commodities e ao câmbio (por duas vias: pela interferência no preço das commodities e pela interferência no preço do petróleo – um dos principais insumos de produção da empresa); e da Natura (NATU3) que aparentemente está associada aos indicadores de risco interno (câmbio, risco mood´s e juros).

Um grupo de ações está alheio ao comportamento dos grupos acima: as ações da Votorantin Celulose e Papel (VCPA4), da Sadia (SDIA4) e da Klabin (KLBN4); que formam um *cluster* próprio (ANEXO 1).

O resultado que surpreende é o distanciamento do comportamento de empresas tradicionalmente exportadoras do comportamento do dólar, principalmente as ações do *cluster* 3 que contém duas das maiores exportadoras nacionais de celulose e papel (ANEXO 1) e (Figura 3).

Outro resultado interessante é a dissimilaridade da Perdigão (PRGA4) e da Sadia (SDIA4) do câmbio, uma vez que, são grandes exportadoras de alimentos. Um possível efeito é que estas empresas dependem de outros insumos indexados ao dólar (como as enzimas de derivados de petróleo e as matrizes). Dessa forma, segundo a visão dos agentes de mercado, é mais forte o efeito do câmbio sobre as commodities do que o câmbio sobre o valor e volume das vendas. Do outro lado, conforme esperado, estas empresas têm um comportamento mais distante das commodities soja e milho (importantes insumos de produção destas empresas) (ANEXO 1) e (Figura 3).

Adicionalmente, conforme o esperado, empresas do mesmo ramo tenderam a ficar restritas ao mesmo *cluster*. Um exemplo disso são os clusters formados pelos bancos, empresas de petróleo e de mineração. Possibilitando excluir o comportamento destas

empresas do comportamento dos indicadores do setor agroindustrial (tanto ações quanto commodities).

6. CONCLUSÕES

O uso dos componentes principais trouxe a tona resultados interessantes e úteis para o processo de tomada de decisão de investimento em ativos relacionados à agroindústria. O investimento apenas em commodities ou apenas em ações de empresas ligadas à agroindústria não parece ser uma boa medida de diversificação de carteira. Uma boa medida é uma carteira que possua tanto commodities quanto estas ações. Para ser mais exato, uma carteira que possua os membros dos *clusters* 1, 2 e 3, mas preferencialmente a combinação entre o 1 e 3.

Na realidade, temos que todos estes ativos agroindustriais estão relativamente relacionados comparados aos ativos mais líquidos do IBRX e que não fazem parte deste grupo (ANEXO 1). Assim, uma diversificação mais abragente deve incluir tanto ativos agroindustriais quanto os ativos fora do cluster principal (no caso os bancos, a Petrobrás e a Vale do Rio Doce) (ANEXO 1).

Pelos resultados obtidos e sua coerência, pode-se considerar que os componentes principais podem ser um bom ferramental mais abragente de tomada de decisão de ativos, não ficando restrito ao setor agroindustrial. Assim, é possível dizer que o trabalho cumpriu satisfatoriamente os objetivos traçados.

Como próximo passo, para corroborar os resultados, poderia ser feito uma análise *ex-post* compondo uma carteira hipotética utilizando, em conjunto com outros critérios mais tradicionais, os critérios obtidos no estudo, verificando se a variância desta é reduzida comparativamente a uma outra carteira (ex: IBRX).

Uma outra possibilidade é a extensão dos resultados obtidos pela incorporação de outros ativos (da própria BOVESPA) para verificar se existem outros *clusters* de comportamento, principalmente setoriais (ex: aço, petróleo e bancos).

REFERÊNCIAS BIBLIOGRÁFICAS

AGUILLERA, A.M.; OCEANA, F.A.; VALDERRAMA, M.J. Stochastic modeling for evolution of stock prices by means of functional principal components analysis, **Applied Stochastic Models in Business and Industry**, v.15, n.4, p.227-234, 1999.

BOVESPA. **Informações para investidores**, www.bovespa.com.br (24/3/2006).

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**, New Jersey: Prentice Hall, 5th edition, 2002. 767p.

PLADIN. **Dados históricos**, www.pladin.com.br (24/03/2006).

ROSS, S.A.; WESTERFIELD, R.W.; JAFFE, J.F. **Corporate finance**, New York: McGraw-Hill, 2nd edition, 2002. 776p.

SHARMA, S. **Applied multivariate techniques**, New York: John Wiley and Sons, 1st edition, 1996, 493p.

TOPALOGLOU, N.; VLADIMIROU, H.; ZENIUS, S.A. CVar model with selective hedging for international asset allocation, **Journal of Banking & Finance**, v.26, n.7, p.1535-1561, 2002.

ANEXO 1

Tabela A1. Variáveis, o número da variável e a composição dos *clusters* por número de *clusters*.

Variável	*n*	*k*=2 (número do *cluster*)	*k*=3 (número do *cluster*)	*k*=4 (número do *cluster*)
CRUZ3	1	2	1	2
SUZB5	2	2	1	2
VALE3	3	-	-	-
VALE5	4	-	-	-
VCPA4	5	2	3	3
UBBR	6	-	-	-
SDIA4	7	2	3	4
RPSA4	8	2	1	2
PETR3	9	-	-	-
PETR4	10	-	-	-
PRGA4	11	2	1	4
PCAR4	12	2	1	2
NATU3	13	1	2	2
KLBN4	14	2	3	4
ITSA4	15	2	1	2
ITAU4	16	-	-	-
FFTL4	17	1	2	1
DURA4	18	2	1	2
BBDC3	19	-	-	-
BBDC4	20	-	-	-
ARCZ6	21	2	1	2

AMBV4	22	2	1	2
Dólar comercial	23	1	2	1
IBRX	24	2	1	4
FGV 100	25	2	1	4
Boi BM&F	26	1	2	1
Soja Chicago	27	1	2	1
Trigo Chicago	28	1	2	1
Açúcar NY	29	1	2	1
Café NY	30	1	2	1
Milho Chicago	31	1	2	1
Risco Brasil	32	1	2	1
CDI	33	1	2	1

Fonte: Elaborados pelos autores.