PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA
BRAZILIAN JOURNAL OF MARKETING, OPINION, AND MEDIA RESEARCH

PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia
ISSN: 1983-9456 (Impressa)
ISSN: 2317-0123 (On-line)
Editor: Fauze Najib Mattar
Sistema de avaliação: Triple Blind Review
Idiomas: Português e Inglês
Publicação: ABEP – Associação Brasileira de Empresas de Pesquisa

Como Danificar Seriamente um Estudo de Segmentação: Use Análise Fatorial como Insumo para *Cluster Analysis*[1]

How to Seriously Damage a Segmentation Study: Use Factor Analysis as Input for Cluster Analysis

Submissão: 28/mar./2014 - Aprovação: 22/abr./2014

Luiz Sá Lucas
Mestre em Programação Matemática pela Universidade Federal do Rio de Janeiro - UFRJ-COPPE. Graduado em Engenharia Elétrica – Sistemas pela Pontifícia Universidade Católica do Rio de Janeiro PUC-RJ. Diretor Técnico no Ibope Inteligência. Membro do Conselho da European Society of Marketing Research - ESOMAR**.** Tem publicado e apresentado vários artigos sobre técnicas matemáticas em revistas especializadas e Congressos/Conferências na América Latina, Europa e Ásia.
E-mail: luiz.lucas@ibopeinteligencia.com; luizsalucas@gmail.com
Endereço profissional: Rua da Assembleia - nº 98 - 12º andar - 20011-000 - Rio de Janeiro/RJ – Brasil.

Wagner Esteves
Mestrando em Engenharia de Produção pela Universidade Federal Fluminense – UFF. Graduado em Estatística pela Escola Nacional de Ciências Estatísticas – ENCE/IBGE. Coordenador de Planejamento no Ibope Inteligência.
E-mail: wagner.esteves@ibopeinteligencia.com

Larissa Catalá
Pós-Graduada em Pesquisa de Mercado pela Escola de Comunicação e Artes da Universidade de São Paulo – ECA-USP. Graduada em Estatística pela Universidade de Campinas – UNICAMP. Tecnologista em Estatística no IBGE.
E-mail: larissa.catala@ibge.gov.br; larissa.catala@gmail.com

[1] Este foi um dos trabalhos apresentados no 6º Congresso Brasileiro de Pesquisa - Mercado, Opinião e Mídia da ABEP (realizado em 24 e 25 de março de 2014), transformado em artigo por seu(s) autor(es), submetido à PMKT e aprovado para publicação.

RESUMO

Os consumidores não são iguais, eles têm diferentes necessidades, comportamentos de compra, propensões, fidelidade à marca etc. Daí a segmentação se transformar numa das principais técnicas para o posicionamento de uma marca. No entanto, um mau e generalizado hábito se cristalizou na pesquisa de mercado: o uso de Análise Fatorial seguida da aplicação de *Cluster Analysis* aos fatores obtidos. Este artigo apresenta uma argumentação para evitar esse uso, em particular do método da Análise de Componentes Principais – ACP (ou Principal Component Analysis – PCA, em inglês), como o primeiro passo. Apresenta-se como exemplo uma extensa segmentação com as duas abordagens (ACP e as variáveis originais), analisando qual dos dois métodos apresentou melhor resultado dentre as 720 segmentações efetuadas.

PALAVRAS-CHAVE:

Segmentação, análise fatorial, análise de componentes principais, análise de grupamento, posicionamento de marca.

ABSTRACT

Consumers are not equal. They have different needs, buying behavior, propensities, brand loyalty etc. Hence Segmentation becomes one of the key techniques in the positioning of a brand. However, a bad and generalized habit has crystallized in Market Research: the use of factor analysis, followed in tandem by an application of cluster analysis on these factors. The article presents an argument to prevent such use, in particular Principal Component Analysis as this first step. Here we present an exercise: an extensive segmentation with both approaches (PCA and the original variables), analyzing which of the two methods showed the best result among 720 segmentations we performed.

KEYWORDS:

Segmentation, factor analysis, principal components analysis, cluster analysis, brand positioning.

1. INTRODUÇÃO

O título deste artigo traz certo exagero. O dano nem sempre é desastroso e, como será visto, existem alguns poucos casos em que a Análise de Componentes Principais - ACP se justifica. No entanto, nos tipos de problemas defrontados em marketing, a superioridade de *Cluster Analysis* – CA, nas variáveis originais, se tornará evidente nos exemplos apresentados.

2. SEGMENTAÇÃO AO LONGO DO TEMPO

A década de 1930 foi extremamente fecunda na criação de diferentes tipos de Análise Fatorial, entre elas a ACP. Os algoritmos de *Cluster Analysis* só começaram realmente a aparecer em meados dos anos de 1960 e, ainda nos anos de 1980, a dominância das técnicas de Análise Fatorial - AF era tal que, mesmo a *Cluster Analysis* era realizada por meio de AF (a chamada *Q Factor Analysis*). Uma excelente descrição desse aspecto pode ser encontrada em Myers (1996) e Stewart (1981). Vale a pena repetir um pequeno trecho do excelente livro de Myers (1996):

> Embora o conceito geral de segmentação de mercado tenha sido formalmente apresentado pela primeira vez por Wendell Smith (1956), mercados já haviam sido segmentados por muitas décadas e mesmo antes disso. Talvez as primeiras formas de segmentação tenham sido baseadas no marketing *mix* (produto, preço, promoção e distribuição). Mesmo antigamente, muito provavelmente os mercadores levavam para os mercados, produtos que diferiam em termos de tipo e qualidade almejados, níveis de preço aceitáveis e/ou métodos de distribuição desejados. (tradução dos autores).

Myers (1996) comenta também a transição da posição de Henry Ford em 1900 "o consumidor pode ter a cor de carro que quiser desde que seja preto" (ênfase na produção, já que a demanda era suficiente para adotar essa posição) para a de Alfred Sloan (GM) em 1920 "um carro para cada bolso e propósito" (ênfase no mercado). Assim, o conceito de segmentação é antigo e se baseia no fato de que os consumidores não são iguais. De fato, a tendência na modelagem em marketing hoje procura ir mais longe: modelos de escolha, como *conjoint analysis*, procuram não só segmentar, mas também estimar preferências individuais. Há vários textos sobre o tema, mas o mais completo é o de Wedel e Kamakura (2000). Foi nesse texto e num *workshop* da ESOMAR com Steve Cohen, que a atenção foi despertada para as dificuldades associadas ao uso do ACP e de *Cluster Analysis* em conjunto.

3. SEGMENTAÇÃO E MARKETING

Gray (2013) apresenta uma sucinta descrição de segmentação que parece bastante útil. Segundo seu resumo, a segmentação é uma das mais importantes metodologias em pesquisa de mercado. De várias formas facilita melhores decisões e aumenta a lucratividade, uma vez que ajuda a:

- Entender o que motiva o comportamento dos diferentes consumidores de uma categoria de produto ou serviço.
- Revela padrões de comportamento e motivações do consumidor e os associa às suas categorias (aspectos demográficos, por exemplo).
- Indica como as várias marcas se posicionam conforme as necessidades dos consumidores nos segmentos.
- Identifica necessidades não atendidas.
- Modifica ofertas existentes para atrair maior volume de consumidores.
- Ajuda no desenvolvimento de novos produtos.
- Melhora a relação com os consumidores.

Gray (2013) ainda indica várias formas de segmentar o mercado, mas aqui se concentrará na maneira mais usual: aplicar métodos de *Cluster Analysis* em medições de necessidades, preferência por marcas, estilos de vida, dados demográficos, informações de bases de dados etc.[2]

4. SEGMENTAÇÃO E ESTRATÉGIA EM MARKETING

Conforme as necessidades estratégicas, a segmentação pode assumir diferentes formas:

- Segmentação por necessidades.
- Segmentação por estilo de vida.
- Segmentação demográfica.
- Segmentação comportamental.

Cada uma tem um objetivo específico, porém uma análise detalhada está fora do escopo deste artigo, inclusive por uma questão de espaço. Para maior aprofundamento sobre o assunto sugere-se a leitura de Kaden, Linda e Prince (2013). Outras referências incluem: King e Wang (2007), McDonald e Dunbar (2004), Dibb e Simkin (2010; 1996) e Wedel e Kamakura (2000).

5. FRAGILIDADES DA ACP *VERSUS* VARIÁVEIS ORIGINAIS

Inicialmente cabem aqui alguns comentários sobre a ACP. Uma famosa abordagem sobre as fragilidades da Análise Fatorial e da ACP, em particular, é apresentada no seminal livro de Stephen Jay Gould, *The mismeasure of men*, publicado em 1981. Com base nos argumentos ali apresentados é preciso muito cuidado ao interpretar os eixos de qualquer Análise Fatorial. Citando outro trabalho, atualmente em desenvolvimento, mas disponível na *Web* (SHALIZI, 2014):

> ACP é uma ferramenta bastante boa quando se precisa ou tenta uma redução na dimensão dos dados quando não se tem certeza do que exatamente usar. Tem algumas propriedades matemáticas interessantes... as dimensões encontradas por ACP... podem ser características reais dos dados ou apenas razoáveis e convenientes ficções e resumos. Que elas sejam reais é uma hipótese que esses métodos podem sugerir, mas para a qual eles só podem sugerir uma evidência muito fraca. Isso importa porque no final das contas nós fazemos *data mining* para descobrir conhecimento sobre o qual possamos *agir*. Uma coisa é fazer com que nossa ação seja apenas uma previsão que nos ajuda a ajustar nossos modelos à prática, mas outra é tentar agir sobre o mundo baseado em como as partes interagem uma com a outra. Para fazer isso direito, precisamos saber o que essas partes realmente são. (tradução dos autores).

Um fator tem uma característica essencialmente geométrica e não uma correspondência com um fenômeno real. Adotar essa correspondência é aquilo que Gould (1981) chamava de "reificar, imaginar que algo é real apenas porque podemos construí-lo de forma abstrata".

5.1 ACP *VERSUS* VARIÁVEIS ORIGINAIS

O que se segue é apoiado fortemente no texto de Yeung e Rizzo (2001). Segundo abordagem desses autores, diferentes algoritmos de *clustering* fornecem diferentes soluções. Isso leva a uma pergunta: qual é a certa? Acredita-se que a *clusterização* é um processo *ad-hoc* que, ao separar elementos em grupos, fornece uma descrição do universo que é útil aos propósitos de marketing (principalmente a descoberta de nichos/*targets*/mercados-alvo de interesse), porém não existe uma resposta certa sobre o assunto (SÁ LUCAS, 2007).

[2] Ver maiores informações sobre a ligação entre pesquisa e bases de dados em Sá Lucas (2007).

ACP é uma técnica de redução de dimensionalidade de um conjunto de dados que transforma as variáveis originais em novas variáveis (as Componentes Principais - CPs) que resumem as características dos dados. Essas componentes são descorrelatadas (mas não necessariamente independentes) e podem ser ordenadas de forma tal que a k-ésima componente seja aquela que tem a k-ésima maior variância no conjunto de componentes principais (CPs). A abordagem tradicional é utilizar as poucas primeiras CPs, pois elas capturam a maior parte da variação do conjunto de dados original.

Em contraste, as últimas CPs são consideradas como as que capturam o ruído residual nos dados. Cabe aqui uma observação: em teoria de sinais aleatórios é muito comum o uso do conceito de ruído branco, que nada mais é que uma variável aleatória com distribuição normal, média zero e uma variância dada. Ora, ruído branco não tem informação alguma e tem variância. Tomar variância como quantidade de informação é, no mínimo, temerário. Ao se tomar as primeiras PCs (as de maior variância) espera-se que elas extraiam a estrutura de *clustering* dos dados. Existem regras práticas para o número de fatores a serem extraídos, mas essas regras são informais e *ad-hoc*.

Por outro, segundo Yeung e Ruzzo (2001), existem considerações teóricas que indicam que as primeiras e poucas PCs podem não conter *cluster information*. Assumindo que os dados consistem na mistura de duas distribuições normais multivariadas com médias diferentes, mas com a mesma matriz de variância-covariância intra*cluster*, Chang (1983) mostrou que as primeiras CPs podem conter menos *cluster information* do que outras com menor variância. Ele, inclusive, gerou um exemplo artificial em dois grupos, em que a melhor separação entre eles se deu no subespaço gerado pela primeira e a última CP.

5.2 EXEMPLO

Para o exemplo será utilizada uma abordagem semelhante à de Yeung e Ruzzo (2001). Foram geradas, a partir de um *package* em R (*Cluster Generation*), bases de dados com 3, 4, 5, 6, 7 e 8 grupos, com três distintos graus de separação entre eles (Qiu e Joe, 2006a; 2006b). Tomaram-se sempre 20 variáveis em cada caso. Foram calculadas, em todos os 720 casos, as CPs. Para a seleção das variáveis mais importantes, utilizou-se uma técnica *ad-hoc*, mas bastante poderosa: tomou-se um agrupamento em cinco *clusters*, gerou-se um preditor por meio de Random Forest (Breiman, 2001).

Em cada caso (de 3 a 8 grupos) foram produzidas 20 segmentações de forma que, a cada vez, fosse incluída uma nova PC e uma nova variável, conforme as ordens apontadas (variância para CP e Random Forest para as variáveis). O algoritmo de clusterização foi o WDM (Sá Lucas, 2007). Como os *clusters* corretos eram conhecidos e fornecidos pelo *cluster generation*, calculou-se o grau de acerto do algoritmo pelo índice de Rand ajustado (aRI) (RAND, 1971). Quando o algoritmo reproduzia o *cluster* perfeitamente, o índice era igual a 1. No pior caso, o índice seria igual à zero. Os resultados são apresentados nas Figuras 1 a 9.

Na Figura 1 os *clusters* são pouco discriminados e, de forma geral, a ACP tem melhor desempenho do que as variáveis originais, embora ambos tenham um aRI bem abaixo de 1 (a clusterização não consegue reproduzir bem a segmentação correta). Na Figura 2, as variáveis originais começam a se destacar, embora com aRI ainda bem abaixo de 1. A situação da Figura 3 é semelhante à da Figura 2, com aRI ainda bem abaixo de 1.

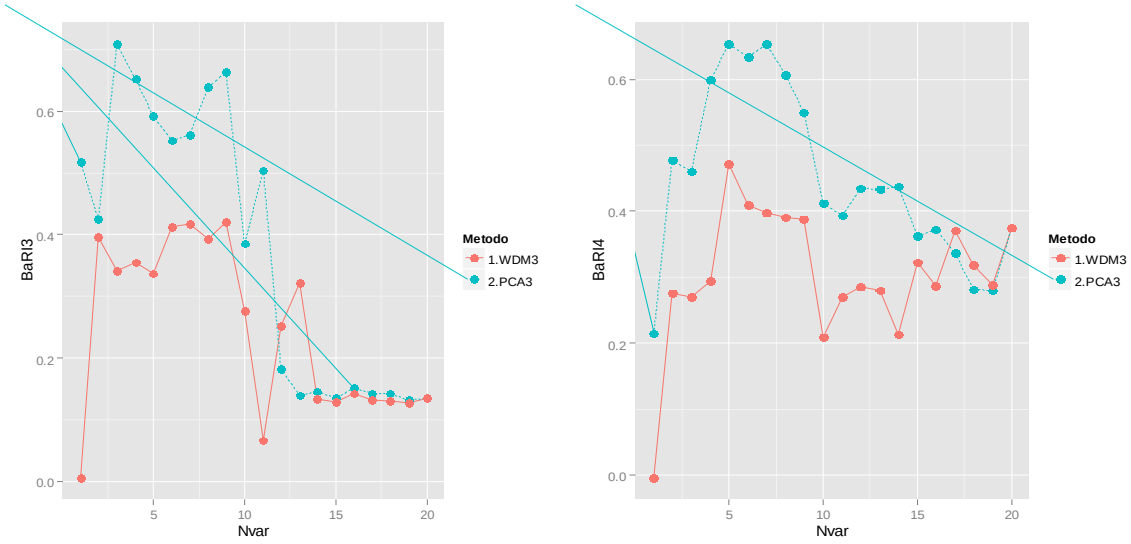

FIGURA 1
Comparação entre os Métodos – *Clusters* pouco discriminados (3 e 4 grupos).

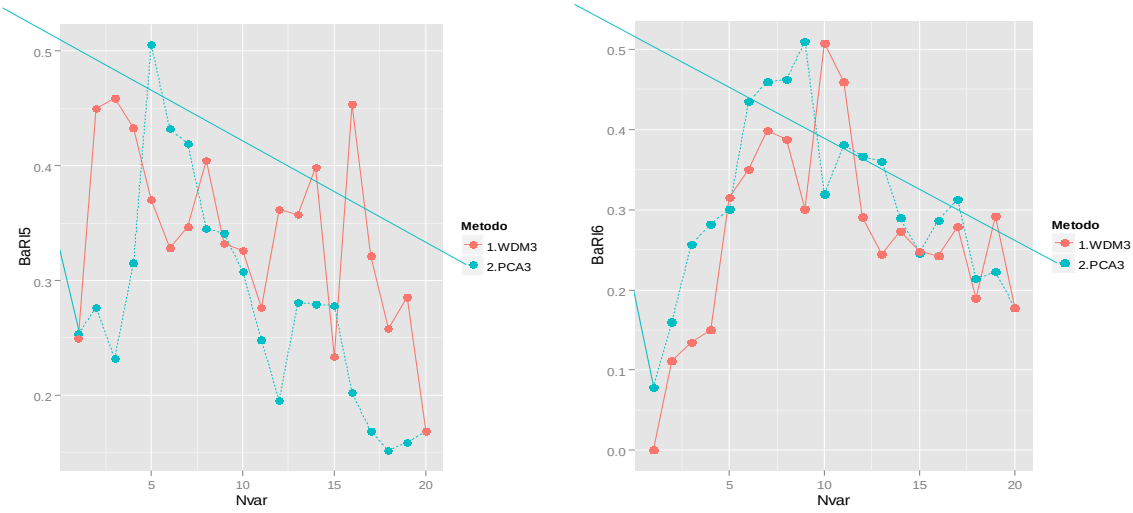

FIGURA 2
Comparação entre os Métodos – *Clusters* pouco discriminados (5 e 6 grupos).

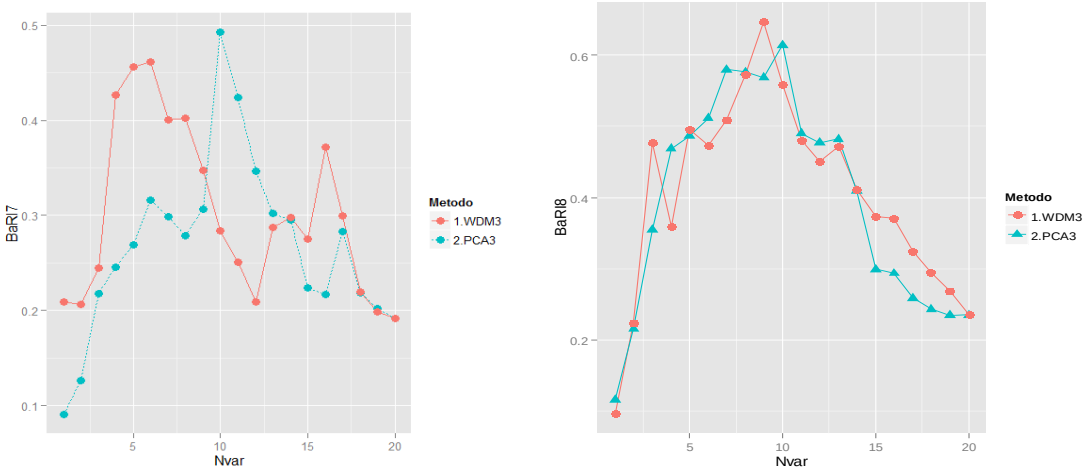

FIGURA 3
Comparação entre os Métodos – *Clusters* pouco discriminados (7 e 8 grupos).

Quando aumenta a discriminação, as variáveis originais começam a se destacar, como na Figura 4, mas com aRI mais próximo de 1. No caso da Figura 5, quando aumenta a discriminação, as variáveis originais começam a dominar, com aRI mais próximo de 1.

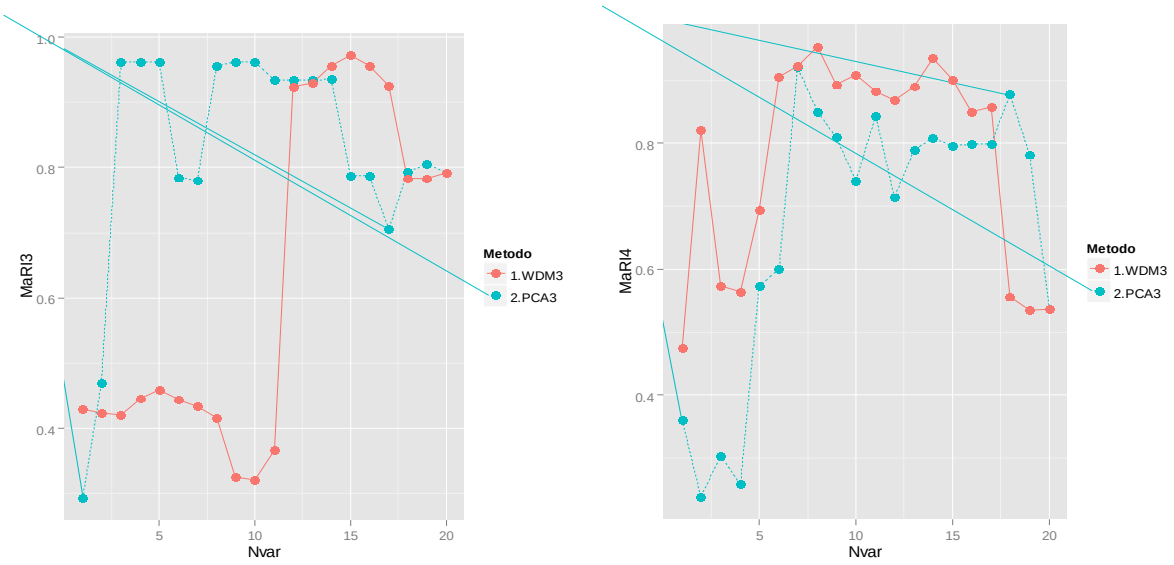

FIGURE 4
Comparação entre os Métodos – *Clusters* medianamente discriminados (3 e 4 grupos).

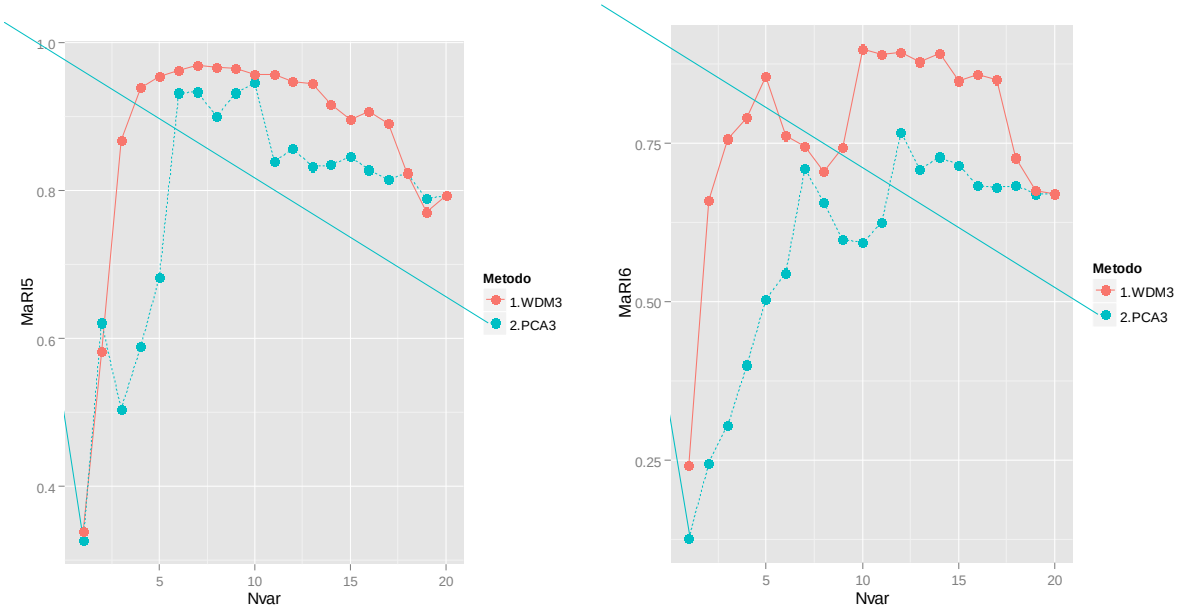

FIGURA 5
Comparação entre os Métodos – *Clusters* medianamente discriminados (5 e 6 grupos).

Novamente, na Figura 6 o caso se repete, quando aumenta a discriminação, as variáveis originais dominam, com aRI mais próximo de 1.

Quando aumenta a discriminação, conforme mostra a Figura 7, as variáveis originais dominam, com aRI chegando a ser igual a 1.

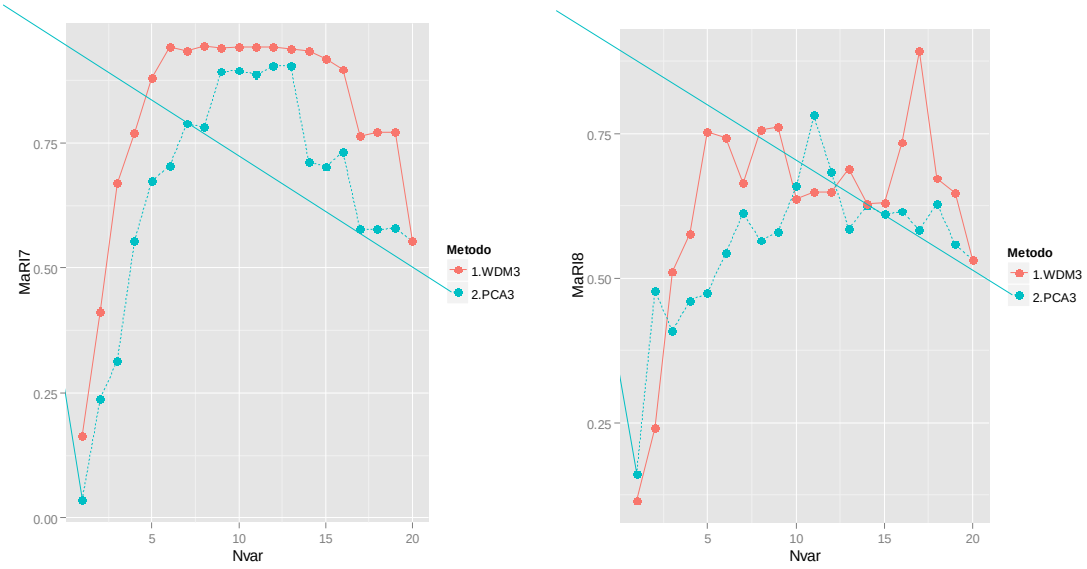

FIGURA 6
Comparação entre os Métodos – *Clusters* medianamente discriminados (7 e 8 grupos).

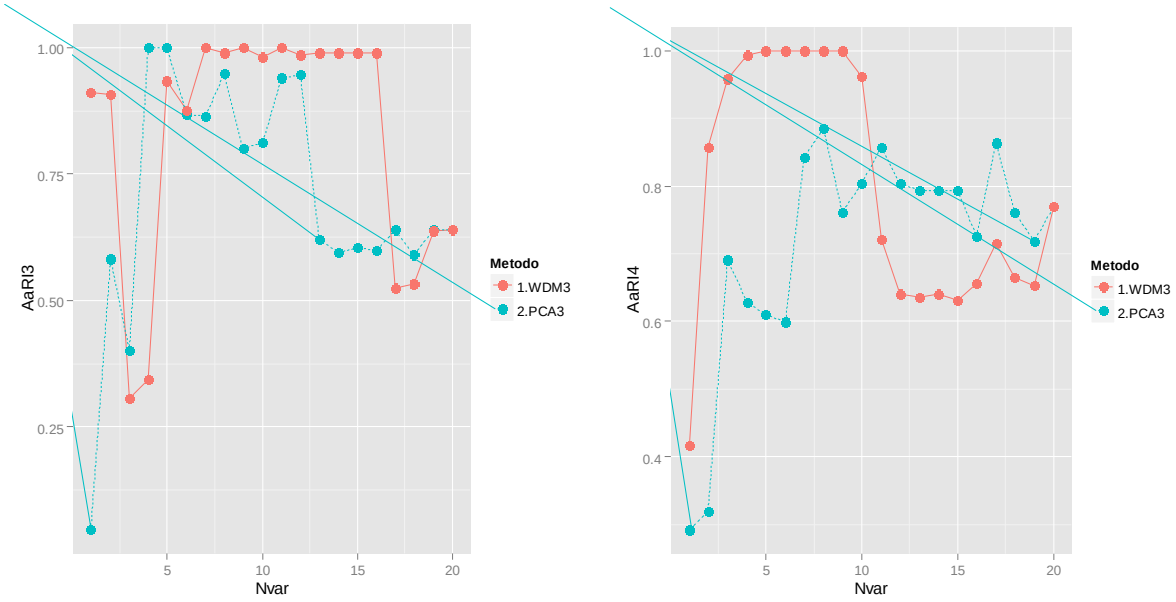

FIGURA 7
Comparação entre os Métodos – *Clusters* altamente discriminados (3 e 4 grupos).

Tudo se repete na Figura 8: quando aumenta a discriminação, as variáveis originais dominam, com aRI chegando a ser igual a 1.

Ainda nesse caso (Figura 9), quando aumenta a discriminação, as variáveis originais dominam, com aRI chegando a ser igual a 1.

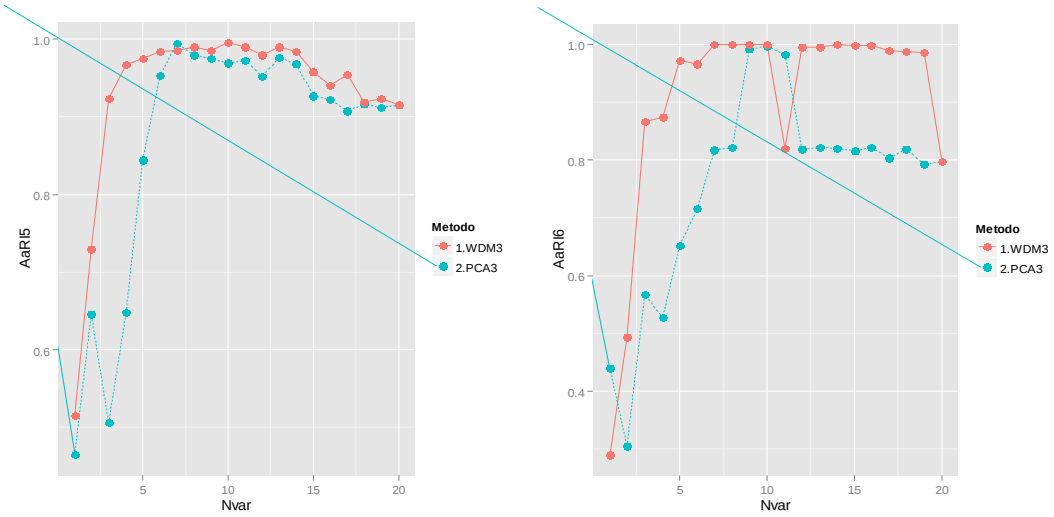

FIGURA 8
Comparação entre os Métodos – *Clusters* altamente discriminados (5 e 6 grupos).

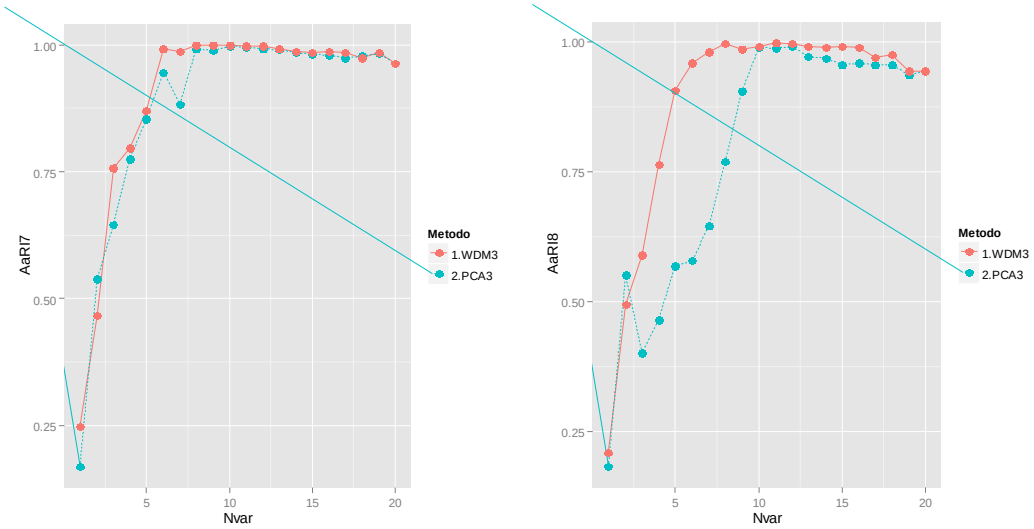

FIGURA 9
Comparação entre os Métodos – *Clusters* altamente discriminados (7 e 8 grupos).

6. CONCLUSÃO

À medida que aumenta a discriminação entre os grupos (melhora a segmentação), a dominância com o uso das variáveis originais aumenta. Se não há discriminação entre os grupos, a segmentação está mal feita. Ainda assim, a segmentação via ACP pode minimizar a distorção.

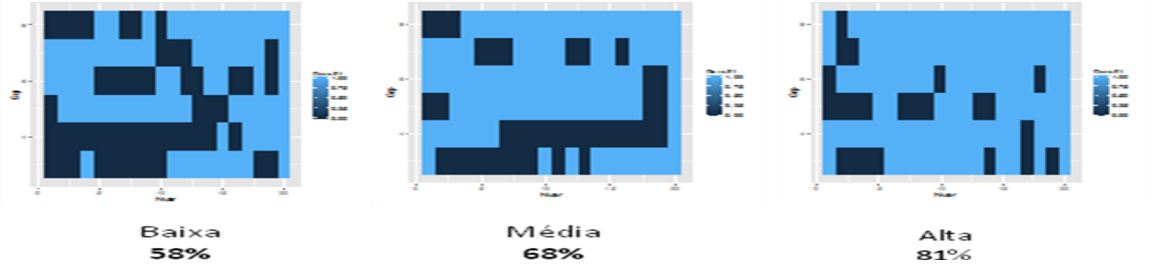

FIGURA 10
Heatmaps conforme a discriminação entre os grupos.

Os *Heatmaps* apresentados na Figura 10 resumem o desempenho dos métodos, sendo indicados os percentuais em que o uso das variáveis originais foi melhor ou igual ao das CPs (ACP), conforme a discriminação entre os grupos.

À medida que aumenta a discriminação entre os grupos, o grau de discriminação aumenta. Além disso, o percentual de acerto do uso das variáveis se torna mais flagrante: azul claro representa os pontos em que as variáveis originais têm igual, ou melhor, desempenho do que as componentes principais. Além disso, o uso de poucas ou de todas as variáveis prejudica o desempenho (ver Figuras 3 a 8). Assim poder-se-ia utilizar como critério geral, a adoção das variáveis originais num nível médio (cerca de 25% a 75%). O uso das componentes principais seria justificável se só se conseguisse atingir uma segmentação pouco clara.

7. REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, L. *Random Forests*, Machine Learning, 45, (1) pp. 5-32, 2001.

CHANG, W. *On using principal components before separating a mixture of two multivariate normal distributions*, Applied Statistics, 32, pp. 267-275, 1983.

DHILLON, I.; MODHA, D.; SPANGLER, W. Class Visualization of High-Dimensional Data with Applications, *Computational Statistics and Data Analysis*, 41, pp. 59-90, 2002.

DIBB, S.; SIMKIN, L. *Target segment strategy*. In: BAKER, M.; SAREN, M. (Org.). Marketing Theory, 2010.

DIBB, S.; SIMKIN, L. *The Market Segmentation Workbook*. London: Routledge, 1996.

GOULD, S. J. *The mismeasure of man*. New York: W.W. Norton & Company, 1981.

GRAY, K. *Think you Know Segmentation? Think Again! A Close Look at 4 Core Analysis*. Quirk´s marketing research media e-newsletter, 2013. Disponível em: <www.quirks.com/articles/2013/20131225-2.aspx>. Acessado em: 5 fev. 2014.

HUBERT, L.; ARABIE, P. Comparing Partitions. *Journal of Classification*, 1985, pp. 193-218.

KADEN, R.; LINDA, G.; PRINCE, M. *Leading edge marketing research.* Los Angeles: Sage Publications. 2013.

KING, D.; WANG, F. Time to Re-think Segmentation, 2007. Disponível em: <http://www.dmnews.com/time-to-re-think-segmentation/article/98990/>. Acessado em: 6 jan. 2014.

MCDONALD, M.; DUNBAR, I. *Market Segmentation* – How to do it, How to profit from it. Oxford: Elsevier, 2004.

MYERS, J. H. *Segmentation and positioning for strategic marketing decisions*, Chicago: American Marketing Association, 1996.

QIU, W.; JOE, H. Generation of Random Clusters with Specified Degree of Separation. *Journal of Classification*, 23 (2), 2006a, pp. 315-334.

QIU, W.; JOE, H. Separation Index and Membership Partial for Clustering. *Computation Statistics and Data Analysis*, 50, 2006b, pp. 585-603.

RAND, W. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66**,** 1971, pp. 846-850.

SÁ LUCAS, L., Joint segmenting consumers using both behavioral and attitudinal data. *Proceedings of the Sawtooth Software Conference*, 2007, pp. 199-218.

SHALIZI, C. *The Truth about Principal Components and Factor Analysis*. Advanced Data Analysis from an Elementary Point of View, 2014. Disponível em: <http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf>. Acessado em: 4 fev. 2014.

SMITH, W. Product Differentiation and Market Segmentaion as Alternative Marketing Strategies. *Journal of Marketing,* 21 (July), 1956, pp. 3-8.

STEWART, D. The Application and Misapplication of Factor Analysis in Marketing Research. Journal of Marketing Research, 18 (February), 1981, pp.51-62.

WEDEL, M.; KAMAKURA, W. Market Segmentation – Conceptual and Methodological Foundations, *International Series in Quantitative Marketing,* Boston: Kluwer Academic Publishers, 2000.

YEUNG, K.; RUZZO, W. An Empirical study of Principal Component Analysis for Clustering Gene Expression Data, *Bioinformatics*, 2001.