UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

MARIA VERÔNICA GONÇALVES ROCHA

ANÁLISE FATORIAL COM O MÉTODO DE EXTRAÇÃO ANÁLISE DE COMPONENTES PRINCIPAIS APLICADOS AO CONSUMO DE COMBUSTÍVEL

Campina Grande
Julho, 2014

MARIA VERÔNICA GONÇALVES ROCHA

ANÁLISE FATORIAL COM O MÉTODO DE EXTRAÇÃO ANÁLISE DE COMPONENTES PRINCIPAIS APLICADOS AO CONSUMO DE COMBUSTÍVEL

Trabalho de Conclusão de Curso apresentado ao curso de curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, em cumprimento às exigências legais para obtenção do título de Bacharel em Estatística.

Orientador:

Prof. Dr. Edwirde Luiz Silva

Campina Grande

Julho 2014

R672a Rocha, Maria Verônica Gonçalves.
Análise fatorial com método de extração análise de componentes principais aplicados ao consumo de combustível [manuscrito] / Maria Verônica Gonçalves Rocha. - 2014.
34 p. : il.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Prof. Dr. Edwirde Luiz da Silva, Estatística".

1. Estatística multivariada. 2. Análise fatorial. 3. Consumo de combustível. I. Título.

21. ed. CDD 519.53

MARIA VERÔNICA GONÇALVES ROCHA

ANÁLISE FATORIAL COM O MÉTODO DE EXTRAÇÃO ANÁLISE DE COMPONENTES PRINCIPAIS APLICADOS AO CONSUMO DE COMBUSTÍVEL

Trabalho de Conclusão de Curso apresentado ao curso de curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, em cumprimento às exigências legais para obtenção do título de Bacharel em Estatística.

Aprovada em: 30 / 07 /2014

Prof. Dr. Edwirde Luiz Silva – DE/CCT/UEPB
Orientador

Prof. Dr. Mácio Augusto de Albuquerque – DE/CCT/UEPB
Examinador

Prof. Ms. Juarez Fernandes de Oliveira – DE/CCT/UEPB
Examinador

Campina Grande – PB

DEDICATÓRIA

Dedico este trabalho a Deus por ter me dado forças e me ajudado todo esse tempo, não só na universidade, mais em todos os dias da minha vida, cuidando de mim com o seu carinho e amor especial.

A meu pai Aderson (en memorian), a minha mãe Marta a guerreira que sempre me incentivou para prosseguir e não desistir assim como todos os meus irmãos José Rogério e Klebérson e irmãs Valdete e Vanusa, cunhados e cunhadas.

Ao meu esposo Afonso e meus filhos André e Anderson, por terem paciência nos momentos que tive de me ausentar para estudar.

Aos meus amigos, em especial Lenilma e Lourenilson que direto ou indiretamente me ajudaram a chegar até aqui, com carinho e atenção.

Aos meus professores pela dedicação em passar seus ensinamentos e experiências.

AGRADECIMENTOS

Primeiramente agradeço a Deus, o centro e o fundamento de tudo em minha vida, por renovar a cada momento a minha força e disposição e pelo discernimento concedido ao longo dessa jornada.

Ao meu professor orientador, Prof. Dr. Edwirde, que teve paciência e que me ajudou bastante a concluir este trabalho, e era um desejo antigo meu que o mesmo me orientasse no meu trabalho final desde que estudei com o ele no meu primeiro ano de universidade.

Ao Prof. Ms. Juarez que desde o dia que o mesmo fez a minha matricula sempre me deu incentivo para acreditar no curso que tinha escolhido, sempre falando para não desistir e por aceitar e fazer parte da minha banca.

Ao Prof. Dr.. Mácio por aceitar fazer parte da minha banca.

Aos meus professores que durante muito tempo me ensinaram e que me mostraram o quanto estudar é bom, mesmo com as minhas dificuldades.

Ao Prof Onildo e Sá de Cálculo I pela paciência e incentivo, que tiveram e sempre me chamando de guerreira.

Ao meu pai Aderson (em memorian), minha mãe Marta guerreira que sempre me ajudou e incentivou a continuar, aos meus irmãos Klebérson, Rogério e Ricardo (em memorian), a minhas irmães Valdete e Vanusa, ao meus sobrinhos Vitor e Vinicius as minhas cunhadas Ana e Silvanira e aos meus cunhados Batista e Luciano que sempre me incentivaram.

Ao meu esposo Afonso pelo carinho atenção e paciência, e aos meus filhos André e Anderson que direto ou indiretamente sempre me deram incentivos para chegar até aqui.

Aos meus amigos e em especial a Lenilma e Lourenilson pelo carinho e atenção que sempre tiveram comigo.

A Andrea da lanchonete que muito me serviu nos momentos de dificuldades a confiança depositada foi essencial, da mesma forma a Nilson e Paula da Xerox pela sua atenção e confiança.

Então tomou Samuel uma pedra, e a pôs
entre Mizpá e Sem, e chamou-lhe Ebenézer; e
disse: **Até aqui nos ajudou o SENHOR.**

RESUMO

Este trabalho acadêmico, de punho monográfico, cujo conteúdo foi elaborado para apresentar a estrutura contextual e metodológica da estatística multivariada que é uma ferramenta poderosa na análise de dados, em um conjunto de métodos estatísticos nos permitindo confrontar diversas variáveis, tendo como propósito de simplificar e nos facilitar a interpretação do fenômeno estudado, possibilitando em particular aplicar as técnicas de análise fatorial e análise de componentes principais, com isso verificando a quantificação dessas componentes bem como a medida de contribuição de cada uma explicando o comportamento dos indicadores iniciais constituem os resultados mais importantes da aplicação de métodos de análise fatorial das componentes principais. Será efetuada uma análise fatorial para reduzir o conjunto de dados relativo ao consumo de combustível na cidade e na estrada de variáveis medidas a um conjunto menor de fatores interpretados.

Palavra-chave: Fatorial, componentes principais, combustível

ABSTRACT

This academic work, monographic fist, whose content was designed to present the contextual and methodological framework of the multivariate statistic that is a powerful tool to analyze data in a set of statistical methods allowing us to confront several variables, with the purpose of simplifying and facilitate the interpretation of the phenomenon studied, allowing in particular to apply the techniques of factor analysis and principal component analysis, thereby verifying the quantification of these components and the extent of contribution of each one explaining the behavior of the initial indicators are the most important results applying methods of factor analysis of the main components. A factor analysis will be performed to reduce the data set to fuel consumption in the city and on the highway of measured variables to a smaller set of factors interpreted.

Keyword: Factorial, principal components, fuel

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

1 INTRODUÇÃO

A análise multivariada é considerada uma das áreas da estatística mais importante pelo fato de estudar os fenômenos observando e analisando suas diversas respostas de forma simultânea.

A análise multivariada é o conjunto de métodos estatísticos que visa simultaneamente analisar conjuntos de dados multivariados, no sentido de que existem várias variáveis medidas para cada objeto ou indivíduo estudado, ou seja, obtém um melhor entendimento do fenômeno objeto de estudo que os métodos estatísticos univariados e bivariados são incapazes de conseguirem.

As técnicas de análise multivariada são úteis para descobrir regularidades no comportamento de duas ou mais variáveis e para testar modelos alternativos de associação entre tais variáveis, incluindo a determinação de quando e como dois ou mais grupos diferem em seu perfil multivariado. Quando analisamos dados associados nós esperamos explicar variações de acordo com um ou mais dos seguintes pontos de vista:

a) Determinação da natureza e do grau de associação entre um conjunto de variáveis dependentes e um conjunto de variáveis independentes.

b) Achar uma função ou fórmula pela qual nós podemos estimar valores das variáveis dependentes a partir das variáveis independentes, o chamado problema da regressão.

c) Estabelecer a significância estatística associada aos itens anteriores.

A estatística multivariada possibilita aplicar técnicas e métodos multivariados a problemas específicos de cada área do conhecimento. Pode ser dividida em três métodos: dependência, interdependência e estruturais. Aqui será detalhada dentre outras existentes, apenas duas das técnicas aplicadas a dados numéricos do método de interdependência: análise fatorial e análise de componente principal.

É usada para analisar as inter-relações entre um grande número de variáveis métricas que explicam estas relações em termos de um número menor de variáveis chamadas fatores (não observável) ou componentes (se observável).

O objetivo do trabalho é fazer uma ilustração da técnica multivariada análise fatorial. Para isso utilizou-se diversos modelos de carros considerando as seguintes variáveis: Gás (combustível), peso, estrada, cilindro, deslocamento e Nox.

1.1 Marco Histórico

A Estatística Multivariada teve o seu início, como corpo teórico diferenciado, no início do século, a partir dos contributos de Pearson(1901), Fisher (1936), Hottelling (1931), Wilks (1932) e Bartlett (1947).

Outras obras mais recentes, mas já considerados clássicas, como as de Kendall (1957, 1975), Anderson (1958, 1984. Morrison (1967, 1976) e Mardia, Kent e Bibby (1982) são as escolhas acertadas para um principiante, embora Morrison não foque métodos como os de análise de clusters e ordenação multidimensional, Mardia e seja demasiado matemático e Kendall utilize uma notação nem sempre de fácil compreensão. Mas recentes e talvez mais acessíveis ao leitor com menor preparação matemática e mais preocupado com a aplicação dos métodos e a interpretação dos resultados são os livros de Chatfield e Collins (1980), Dillon e Goldstein (1984) Hair, (1992) Anderson, Tatham e Black (1987) e Everit e Dunn (1991). Manly (1986) constitui uma introdução para não matemáticos, mas que apenas poderá servir de ponto de partida nada mais.

1.2 Finalidade da análise multivariada

- Fornecer métodos cujo objetivo é o estudo conjunto de dados multivariados que a análise estatística de várias variáveis seja simplificada;

- Auxilia o pesquisador a tomar melhores decisões no contexto em que se encontra considerando as informações do conjunto de dados analisados.

2 MATERIAIS E MÉTODOS

Para fins de ilustração, suponha que tenha sido realizada pesquisa sobre 20 tipos de carros. Para cada tipo de carro foram consideradas as seguintes características: consumo de combustível na cidade e na estrada, peso do automóvel, quantidade de cilindros, deslocamento, gás e óxidos de nitrogênios (Nox).

Tabela 1: Modelos de carros e algumas de suas variáveis

Carro	**Cidade**	**Estrada**	**Peso**	**Cilindro**	**Deslocamento**	**Gás**	**No_x**
Chev, Camaro	19	30	3545	6	3,8	12	34,4
Chev, Cavalier	23	31	2795	4	2,2	10	25,1
Dodge Neon	23	32	2600	4	2	10	25,1
Ford Taurus	19	27	3515	6	3	12	25,1
Honda Accord	23	30	3245	4	2,3	11	25,1
Lincoln Cont,	17	24	3930	8	4,6	14	25,1
Mercury Mystique	20	29	3115	6	2,5	12	34,4
Mitsubishi Eclipse	22	33	3235	4	2	10	25,1
Olds, Aurora	17	26	3995	8	4	13	34,4
Pontiac Grand Am	22	30	3115	4	2,4	11	25,1
Toyota Camry	23	32	3240	4	2,2	10	25,1
Cadillac DeVille	17	26	4020	8	4,6	13	34,4
Chev, Corvette	18	28	3220	8	5,7	12	34,4
Chrysler Sebring	19	27	3175	6	2,5	12	25,1
Ford Mustang	20	29	3450	6	3,8	12	34,4
BMW 3-Series	19	27	3225	6	2,8	12	34,4
Ford Crown Victoria	17	24	3985	8	4,6	14	25,1
Honda Civic	32	37	2440	4	1,6	8	25,1
Mazda Protege	29	34	2500	4	1,6	9	25,1
Hyundai Accent	28	37	2290	4	1,5	9	34,4

2.1 Análise fatorial

A análise fatorial inclui um conjunto de técnicas estatísticas cujo objetivo é representar ou descrever um número de variáveis hipotéticas. Por outras palavras, a análise fatorial permite identificar novas variáveis, em número menor que o conjunto inicial, mas sem perda significativa de informação contida neste conjunto.

A análise fatorial é uma técnica estatística destinada a representar um processo aleatório multivariado por meio da criação de novas variáveis, derivadas das variáveis originais e, geralmente, em menor número, que representa as comunalidades do processo restando às variáveis espúrias serem não descritas pelo modelo fatorial. Ela permite concluir se é possível explicar este padrão de correlações através de um menor número de variáveis. Será uma análise exploratória utilizada com o objetivo reduzir a dimensão dos dados. Mas a análise fatorial pode também ser confirmatória se for utilizada para testar uma hipótese inicial de que os dados poderão ser reduzidos a uma determinada dimensão e de qual a distribuição das variáveis segundo essa dimensão. No entanto, nem sempre esta divisão é clara quando se aplica a análise fatorial.

A técnica da análise fatorial foi originalmente desenvolvida por um psicólogo, Spearman, como instrumento de análise das faculdades humanas. Partindo da hipótese da existência de um só fator inteligência e da impossibilidade de medir diretamente, Spearman desenvolveu a análise fatorial para poder estudar o fator inteligência indiretamente a partir das correlações entre diferentes testes. Segundo Thurstone (1947) retornou a ideia inicial de Spearman, mas desenvolveu-a por acreditar existir mais que um fator inteligência. Daqui resultou o desenvolvimento de uma análise fatorial que permita detectar mais do que um fator subjacente aos dados iniciais

Foi introduzida em 1904, muitos dos exemplos iniciais aparecem na área da psicologia e ciências sociais, na tentativa de identificar-se os fatores relacionados com a inteligência humana e liga-los, de algum modo, à etnia. Devido a certa subjetividade e falta de unidade de suas soluções, a análise fatorial tem sido alvo de críticas ao longo dos anos (HILLS, 1977; KACHIGAN, 1991). Porém, é uma das técnicas multivariadas mais conhecidas e tem sido muito utilizada em áreas como química, educação, geologia, marketing, entre outras. Alguns exemplos comuns são: a identificação do perfil dos consumidores ou os fatores que os levam a comprar, ou preferir certos produtos e marcas (Werner et al., 2003); a análise do posicionamento de produtos e serviços perante os concorrentes de mercado, a descoberta dos fatores relacionados à satisfação do individuo com sua vida pessoal, atividade profissional, ou produtos de um modo geral (Jobson, 1996; Jackson, 2003), elaboração de índices diferenciados de qualidade (MINGOTI, 2001) etc. vários exemplos interessantes na área de pesquisa de opinião e mercado podem ser encontrados em Dillon; Goldstein (1984).

A análise fatorial, em sua versão clássica de determinar os fatores ortogonais que descrevem aproximadamente e sucessivamente os vetores-resposta de n indivíduos a um conjunto constituído por m testes psicológicos, relaciona-se com os trabalhos de Karl Pearson (1901) e Charles Spearman (1904). Este último trata, pela primeira vez, do que hoje se conhece como as variáveis latentes mencionadas anteriormente. Assim é a inteligência e são desta mesma natureza, muitas outras variáveis psicológicas, sociais e econômicas (SOUZA, 1988).

Nos diferentes casos concretos de estudo de populações deparamo-nos muitas vezes com dezenas de indicadores, tornando-se absolutamente necessário reduzir a dimensão da análise para que a situação se torne compreensível. Se estudarmos **n** indivíduos para os quais existe informação sob a forma de **p** variáveis, para descrever toda a informação são necessários ½p (p+3) parâmetros, isto é:

a) Média de cada variável: $\mu_i = \frac{1}{N}\sum_u^N x_{ui}$

b) Variância de cada variável: $\sigma_i^2 = \frac{1}{N} \sum_{u=1}^N (x_{ui} - \mu_i)^2$

c) Covariâncias para cada par de variáveis: $\sigma_{ij} = \frac{1}{N} \sum_{u=1}^N (x_{ui} - \mu_i)(x_{uj} - \mu_j)$.

d) Correlações para cada par de variáveis: $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$

sendo x_{ui} o termo geral da matriz de dados originais correspondente ao valor da variável i para o indivíduo u (u = 1, 2, ..., n; i = 1, 2, ..., p).

Por exemplo, para uma população caracterizada por dez atributos, teríamos que calcular e analisar 65 parâmetros.

Do mesmo modo que a análise de componentes principais, a análise fatorial tem como objetivo principal descrever a variabilidade original do vetor aleatório X, em termos de um número menor m de variáveis aleatórias, chamadas de fatores comuns e que estão relacionadas com o vetor original X através de um modelo linear. Neste modelo, parte da variabilidade de X é atribuído aos fatores comuns, sendo o restante da variabilidade de X atribuído às variáveis que não foram incluídas no modelo, ou seja, ao erro aleatório.

Em linhas gerais, o que se espera é que as variáveis originais $X_i, \ i = 1,2,\dots,p$ estejam agrupados em subconjuntos de novas variáveis mutuamente não correlacionadas, sendo que a análise fatorial teria como objetivo o encontro destes fatores de agrupamento. Deste modo, em casos nos quais se tem um número grande de variáveis medidas e correlacionadas entre si, seria possível, a partir da análise fatorial, identificar-se um número menor de novas variáveis alternativas, não correlacionadas e que de algum modo sumarizassem as informações principais das variáveis originais. Estas novas variáveis alternativas são chamadas de fatores ou variáveis latentes.

A análise fatorial adapta como principio que cada variável pode ser decomposta em duas partes: uma parte comum e uma parte única. A primeira é a parte da sua variação partilhada com outras variáveis, enquanto que a segunda é específica da sua própria variação. Assim, uma distinção importante entre os dois métodos provém do montante de variância analisada: enquanto que na análise de componentes principais se considera a variação total presente no conjunto das variáveis originais, na análise fatorial só é retida a variação comum partilhada por todas as variáveis. As componentes principais são expressas como combinações lineares das variáveis originais. Por exemplo, para m componentes e p variáveis (m $\leq$ p):

$$CP_1 = a_{11}X_1 + a_{21}X_2 + \dots + a_{p1}X_p$$

$$CP_2 = a_{12}X_1 + a_{22}X_2 + \dots + a_{p2}X_p$$

...

$$CP_m = a_{1m}X_1 + a_{2m}X_2 + ... + a_{pm}X_p$$

Na análise fatorial, cada variável observada é descrita como uma função dos fatores comuns (FC) e de um fator único. Para (m < p) fatores comuns:

$$X_1 = a_{11}FC_1 + \alpha_{21}FC_2 + ... + \alpha_{m1}FC_m + e_1$$
$$X_2 = \alpha_{12}FC_1 + \alpha_{22}FC_2 + ... + \alpha_{m2}FC_m + e_2$$
$$...$$
$$X_p = \alpha_{1p}FC_1 + \alpha_{2p}FC_2 + ... + \alpha_{mp}FC_m + e_p$$

Os dois conjuntos de equações mostram uma diferença adicional entre os dois tipos de análise: a presença ou não de uma parcela de erro. Na análise de componentes principais o erro está ausente, as variáveis observadas são medidas sem erro e as variáveis latentes são combinações lineares perfeitas dessas variáveis. Na análise fatorial o erro explica uma parceira da variância de respectiva variável, não explicada pelos fatores comuns.

Segundo (GUTS et al, 2013) a determinação dos fatores está orientada a encontrar na medida máxima possível as correlações entre os valores $x_1, ..., x_p$. Os fatores se selecionam em uma quantidade tal que se pode explicar da melhor maneira possível a matriz de correlação para $x_1, ..., x_p$. Como se observa, a análise fatorial está orientada ao estudo das correlações. Os fatores $f_1, ..., f_m$ encontram-se através da equação:

$$x_i = \sum_{k=1}^{m} l_{im} f_m + e_i \qquad (1)$$

Sendo $i = 1, ..., p$; x_i denominam variáveis; l_{im} é a carga fatorial do *m*-ésimo fator para a *i*-ésima variável (ou de *i*-ésima variável para o *m*-ésimo fator); e_i é o error, que representa as fontes de desvio que atuam apenas sobre x_i

Supõe-se que o valor esperado das variáveis aleatórias x_i , e_i e f_m é igual a zero,

$Mx_i = 0, \quad Me_i = 0 \quad (i = 1, \ldots, p)$

$Mf_m = 0 \quad (k = 1, \ldots, m)$

E que para as matrizes de covariância cumprem-se as condições:

$M[(f_i - Mf_i)(f_j - Mf_j)] = M[f_i\ f_j] = \delta_{ij},$

$M[(e_i - Me_i)(e_j - Me_j)] = M[e_i\ e_j] = v_i\delta_{ij},$

$M[(f_i - Mf_i)(e_j - Me_j) = M[f_i\ e_j] = 0,$

Em que $\delta_{ii} = 1$, $\delta_{ij} = 0\ (\ i\ \neq j)$ é o símbolo de Kronecker. Isto significa que os fatores e os erros consideram não correlacionados.

Da equação (1), deduz-se que, em primeiro lugar que:

$l_{ij} = M[x_i f_j]$

Ou seja, as cargas fatoriais são correlacionadas entre as variáveis y os fatores, e, em segundo lugar:

$$k_{ii} = \sum_{k=1}^{m} l_{ik}^2 + v_i$$

$$k_{ij} = \sum_{k=1}^{m} l_{ik} l_{jk}\ (i\ \neq j)$$

Em que, $\quad k_{ij} = M[(x_i - Mx_i)(x - Mx_j)] = M[x_i\ f_x]$

Os números l_{ik}^2 denominam-se *i*-ésima comunalidades e os números v_i, especificidade. A comunalidade é a parte da variância da variável k_{ii} que é explicada pelos fatores, a especificidade é a parte da variância que não é explicada pelos fatores.

Estas equações podem-se escrever na forma matricial da seguinte forma:

$$K_x = LL' + V \qquad (2)$$

A equação (2) é o modelo principal da análise fatorial. Posto que a matriz de covariância K_x está correlacionada com a matriz de correlação, tem-se que:

$$R_x = \left\| \frac{k_{ij}}{\sqrt{Dx_i Dx_j}} \right\|, \qquad K_x = \left\| k_{ij} \right\| \qquad (3)$$

As equações (2) – (3) gera a representação da matriz de correlação mediante a matriz $L = \| l_{ik} \|$ das cargas e a matriz diagonal $V = \left\| v_i \delta_{ij} \right\|$ dos elementos próprios positivos v_i.

Indica-se que a matriz $LL' = LIL'$, em que I é a matriz identidade. Isto significa que a matriz LL' é o resultado da transformação da matriz I em uma nova base. Por isso, seus valores próprios são positivos.

2.2 **Resolução das equações do modelo principal**

Para encontrar os fatores reduz-se a busca das cargas fatoriais l_{ik} e os erros e_i segundo os valores dados $x_1, \dots, x_p$. Na análise fatorial não se resolve a equação (1), se não a equação (2). No primeiro membro desta equação tem p(p+1)/2 quantidades conhecidas (os elementos da matriz de covariância). No segundo membro, pm + p=n(m + 1) incógnitas: l_{ik}, v_i. Por isso, m + 1 > (p + 1)/2. Sabe-se que p é o número de variáveis aleatórias e m é o número de fatores buscados, então a solução não é única.

Na maior parte das vezes os termos análise de fatores comuns e análise de componentes principais são utilizados como sinônimos, porém não o são, mesmo não havendo grande diferença em seus métodos e resultados, afinal ambos são métodos de análise fatorial. A diferença conceitual importante aqui é que na análise de componentes principais a variância a ser considerada para a extração dos fatores é a variância total, e na análise fatores comuns considera-se apenas a variância comum entre as variáveis.

Em seguida, deve-se realizar o procedimento de extração de fatores em análise fatorial exploratória. Nem todos os fatores são aproveitáveis numa análise fatorial e há controvérsia sobre os critérios que determinam quando um fator é estatisticamente importante. A determinação do número de fatores pode ser facilitada por meio da análise do gráfico de sedimentação (scree plot), técnica advogada por Cattell (1966).

Os autovalores são, por sua vez, medidas do comprimento dos autovetores na elipse, ou da figura tridimensional elipsóide – se considerarmos correlações multivariadas. Portanto, ao analisarmos os autovalores de um conjunto de dados, consegue-se conhecer de que forma as variâncias da matriz de correlações estão distribuídas. Em outras palavras, é possível visualizar as grandezas da figura elipsóide formada na distribuição espacial das variáveis. O autovalor é calculado pela soma dos quadrados dos carregamentos de cada variável para a variável latente representada pelo fator obtido. O entendimento do conceito de autovalor é facilitado ao lembrarmos dois pontos importantes: 1) o carregamento (loading) é o valor do coeficiente de correlação entre a variável e o fator obtido; e 2) o quadrado do coeficiente de qualquer correlação é igual a porcentagem da variância de uma variável que é explicada pela outra. (FIELD, 2009).

Portanto, quando se analisa os autovalores de um conjunto de dados, consegue-se conhecer de que forma as variâncias da matriz de correlações estão distribuídas. Em outras palavras, os autovalores representam o quanto da variância é explicada pelo fator.

Outra opção para verificação do número de fatores a serem mantidos é a análise paralela. Bases de dados com as mesmas características da base a ser analisada são geradas aleatoriamente para terem seus autovalores comparados. Os fatores são mantidos se seus autovalores forem maiores que os autovalores das bases com dados randômicos.

Se a análise fatorial revela a unidimensionalidade do instrumento de avaliação, a utilização da teoria clássica dos testes para o cálculo da discriminação do item torna-se mais aceitável quando calculada por meio de coeficiente de correlação item-total e também é um dos pressupostos da teoria de resposta ao item. Outros métodos para avaliação da dimensionalidade incluem a análise de componentes principais da matriz de correlações inter-itens, empregada na teoria clássica dos testes; e a análise de componentes principais dos resíduos do modelo de regressão logística, empregado na teoria de resposta ao item.

Em seguida, deve-se realizar o procedimento de extração de fatores em análise fatorial exploratória. Nem todos os fatores são aproveitáveis numa análise fatorial e há controvérsia sobre os critérios que determinam quando um fator é estatisticamente importante. A determinação do número de fatores pode ser facilitada por meio da análise do gráfico de *scree plot*, técnica advogada por Cattell (1966).

2.3 Rotação e interpretação das componentes principais

A aplicação de um método de rotação tem como objetivo principal a transformação dos coeficientes das componentes principais retidas numa estrutura simplificada (THURSTONE, 1947). Seja $C_{(p\ x\ m)}$ a matriz de vectores $\tilde{a}_j^*$, antes da rotação, e $\tilde{B}_{(p\ x\ m)}$ a matriz de vectores $\tilde{b}_j$, depois da rotação:

$$\mathbf{C} = [\ \tilde{a}_1^*\ \ \tilde{a}_2^*\ \dots\ \tilde{a}_m^*]$$

$$\mathbf{B} = [\ \tilde{b}_1\ \ \tilde{b}_2\ \dots\ \tilde{b}_m]$$

Em que **m** é o número de componentes principais retidas e

$$\mathbf{B} = \mathbf{C}\ .\ \mathbf{G}$$

sendo **G** uma matriz de dimensão (m x m)

Para que **B** seja uma estrutura simplificada é necessário que se verifiquem as seguintes condições (HARMAN, 1976):

1. Cada linha de **B** deverá conter pelo, menos um zero, ignificando que cada variável deverá estar não correlacionada com pelo menos uma componente depois da rotação;

2. Cada coluna de **B** deverá conter pelo menos q zeros;

3. Para cada par de colunas de **B**, as variáveis com coeficientes nulos numa das colunas não os deverão ter na outra coluna; quando m ≥ 4, cada coluna deverá ter um maior número de coeficientes nulos do que não nulos. Estas condições tentam garantir a independência dos vectores depois da rotação.

Com exceção do caso em que apenas são retidas duas componentes principais (m = 2), os métodos de rotação são demasiado complicados e só a utilização de software apropriado torna acessível a sua aplicação. Diversos algoritmos foram propostos para rotação das componentes principais, seja pelos métodos ortogonais, seja por métodos oblíquos.

Quando se compreende que a dimensão (ou fator) é um eixo de classificação no qual as variáveis estão posicionadas, compreende-se a importância de se conhecer os métodos de rotação fatorial. Não cabe o detalhamento dos métodos de rotação no escopo deste artigo. No entanto, deve-se esclarecer que a rotação escolhida para extração dos fatores depende principalmente do grau de inter-relação que se supõe para seus fatores. Se você deseja encontrar fatores independentes, os métodos de rotação ortogonal, como o varimax, são preferidos em detrimento daqueles de rotação oblíqua, como o promax e o direct oblimin - mais apropriados para fatores correlacionados entre si. O método de VARIMAX foi proposto por Kaiser (1958).

2.4 As componentes principais de uma matriz de correlações

A alternativa mais utilizada é calcular as componentes principais a partir das variáveis padronizadas com variância unitária, o que corresponde exatamente a aplicar a análise das componentes principais à matriz de correlações **R.** O procedimento matemático é exatamente o mesmo e as componentes principais passam a ser os vectores próprios passam a ser os vectores próprios associados aos valores próprios de **R**. A soma dos valores próprios passa a ser **p**, o número de variáveis e a variância total das variáveis padronizadas. A proporção de variância explicada pela componente Y_j passa então a ser $\frac{\lambda_j}{p}$.

Alguns autores sugerem que, para que um modelo de análise fatorial possa ser adequadamente ajustado aos dados, é necessário que a matriz de correlação inversa R_{pxp}^{-1} seja próxima da matriz diagonal (RENCHER, 2002).

2.5 Contrastes no modelo fatorial

Uma medida de adequação que é fundamentada nesse princípio é o coeficiente KMO proposto inicialmente Kaiser (1970).

2.5.1 Estatística da Kaiser-Meyer-olkin (KMO)

Esta estatística compara as correlações entre as variáveis. Seja r_{ij} o coeficiente de correlação observado entre as variáveis i e j e a_{ij} o coeficiente d correlação parcial entre as mesmas variáveis que é simultânea , uma estimativa das correlações entre os fatores. Os a_{ij} deverão estar próximos de zero uma vez que se pressupõe que os factores são ortogonais entre si.

A estatística KMO é definida da seguinte maneira:

$$KMO = \frac{\sum_i \sum_j R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 \ + \ \sum i \neq j}$$

Em que R_{ij} é a correlação amostral entre as variáveis $X_i \, e \, X_j, \ e \ Q_{ij}$ é a correlação parcial entre $X_i \ \ e \ \ X_j$. A correlação parcial entre duas variáveis é a correlação que existe entre elas quando todas as outras (p-2) variáveis são consideradas como constantes (Kachigan, 1991; Wichern, 2002). Quando as correlações parciais Q_{ij} são próximas de zero, o coeficiente KMO está próximo de 1, ou, equivalentemente, a matriz R_{pxp}^{-1} está próxima da matriz diagonal. Kaiser; Rice (1977) indicam que, para a adequacidade de ajuste de um modelo de análise fatorial, o valor de KMO deve ser maior ou igual a 0,8. Na literatura, existem também sugestões de se trabalhar com faixas de validade de KMO (Pereira, 1999). Assim, um KMO na faixa de 0,9 seria excelente enquanto que um coeficiente na faixa de 0,5 seria péssimo e exigiria medidas de correção nos dados amostrais através de exclusão de variáveis dentre as p avaliadas, ou inclusão de novas variáveis.

Uma discussão sobre o uso de KMO é apresentada em Rencher (2002), mostrando que o usuário deve ser cauteloso ao utilizar apenas o valor de KMO como ponto de partida para ajustar ou não um modelo de análise fatorial ortogonal aos dados. O coeficiente KMO está disponível nos software SPSS e SAS, dentre outros.

2.5.2 Teste de esfericidade de Bartlett para a matriz de correlação

O ajuste de um modelo de análise fatorial aos dados pressupõe que as variáveis-resposta sejam correlacionadas 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. Este teste, chamado de esfericidade de Bartlett. As hipóteses testadas são:

$H_0 : P_{pxp} = I_{pxp}$

$H_0 : P_{pxp} \neq I_{pxp}$,

Em que I_{pxp} é a matriz de correlação teórica das p-variáveis. A estatística de teste T é definida por:

$$T = -\left[n - \frac{1}{6}(2\,p + 11)\right]\left[\sum_{=1}^{p} \ln(\widehat{\lambda_\imath})\right]$$

onde $l_n(.)$ denota a função logaritmo neperiano de $\widehat{\lambda_\imath}$, = 1,2,...,p são os autovalores da matriz de correlação amostral R_{pxp}. Sob a hipótese nula e n grande, a estatística T tem uma distribuição aproximadamente Qui-quadrado com $\frac{1}{2}p\ (p-1)$ graus de liberdade. Para que o modelo de análise fatorial possa ser ajustado, o teste de Bartlett deve rejeitar a hipótese nula. É importante lembrar que a aplicação do teste de Bartlett requer que as variáveis envolvidas na análise tenham distribuição normal p-variada.

2.5.3 Estimador do número de fatores

Quantos fatores devem-se ser considerados? Propõem-se o seguinte método estatístico. Comprova-se a hipótese nula H_o: precisam-se de m fatores versus a hipótese alternativa: Não precisam-se de m fatores.

A estatística de teste é a seguinte (Maxwell, 1971):

$$\chi^2 \sim \left[(p-1) - \frac{1}{6}(2n+5) - \frac{2}{3}m\right] \sum_{i<j} \frac{(a_{ii} - \hat{c}_{ij})^2}{(\hat{v}_i \hat{v}_j)} \qquad (4)$$

Em que $\hat{c}_{ij}$ e $\hat{c}$ são os resultados da estimação da matriz de covariância, a matriz de cargas L e a matriz V (ou seja, o resultado da solução de, por exemplo, o modelo fatorial principal).

Estabelece o nível de significância 100 $\alpha\%$ e encontra-se o valor da χ^2_{calc} utilizando a fórmula (4). Comprova-se este valor com o valor tabulado, χ^2_{tab} com número de graus de liberdade igual a $[(p-m)^2 - (p+m)]/2$. Se $\chi^2_{calc} < \chi^2_{tab}$, então a hipótese H_o á aceita. Caso contrário a hipótese H_o é rejeitada e o modelo fatorial precisa-se $m+1$ fatores, ao mesmo.

3 RESULTADOS E DISCUSSÕES

Observa-se na Tabela 2 a matriz de correlação das variáveis vista na Tabela 1. Foi possível visualizar várias correlações altas, por exemplo, a variável Estrada e a variável Gás, bem como a variável Cidade e Gás. Os pares de variáveis que apresentaram baixas correlações foram as seguintes: deslocamento e cidade, estrada e cidade. Isso é confirmado pelo determinante da matriz que é usado como índice que permite saber se existem correlações altas na matriz que permite extrair fatores.

Tabela 2: Matriz de correlações

	Cidade	Estrada	Peso	Cilindro	Deslocamento	Gás	No_x
Cidade	1,000						
Estrada	**0,928**	1,000					
Peso	**-0,871**	**-0,874**	1,000				
Cilindro	**-0,799**	**-0,833**	**0,802**	1,000			

Desloca-mento	-0,672	-0,670	**0,711**	**0,894**	1,000		
Gás	**-0,938**	**-0,966**	**0,906**	**0,884**	**0,742**	1,000	
Nox	-0,317	0-,145	0,202	0,452	0,432	0,291	1,000

Nesse caso o determinante foi igual a 0,000, indicando a existência de altas correlações na matriz de correlação, fortalecendo a ideia de usar análise de redução dimensão.

Tabela 3: Teste KMO e Bartlett's

Medida de adequação amostral de Kaiser-Meyer-Olkin – KMO		0,795
Prova de esfericidade de Bartlett	χ^2 aproximado	172,518
	Grau de liberdade	21
	p-valor	0,000

Os testes de KMO e Bartlett indicam qual é o grau de suscetibilidade ou o ajuste dos dados à análise fatorial.

A prova de esfericidade de Bartlett contrasta a hipótese de que os elementos fora da diagonal principal (as correlações) da matriz de correlações são zero. O KMO contrasta se as correlações parciais entre as variáveis são pequenas, toma valores entre 0 e 1, e indica que a análise fatorial é tão mais adequada quanto maior seja seu valor. Neste caso obtivemos um KMO = 0,795, sendo assim aceitável a adequação da análise fatorial ao conjunto de dados. Pelo teste de significância de Bartlett a hipótese nula é rejeitada, pois o valor da estatística é 172,518 com um valor p = 0,000. Confirmando também a adequação do modelo fatorial na análise dos dados.

Uma vez que os dados são apropriados para uma análise fatorial, segue-se em frente com a extração dos fatores.

Tabela 4. Comunalidades

	Inicial	Extração
Cidade	1,000	0,891

Estrada	1,000	**0,953**
Peso	1,000	0,885
Cilindro	1,000	0,912
Deslocamento	1,000	0,775
Gás	1,000	**0,966**
Nox	1,000	**0,936**

Primeiramente tem-se a fatoração inicial extraída pelos eixos principais apresentadas pelas comunalidades, ou seja, é a proporção da variância de cada variável explicada pelos fatores comuns ou pelas componentes principais. A comunalidade de uma variável é a soma das cargas fatoriais associadas a ela elevada ao quadrado. No caso de componentes principais, quando se retém todas as variáveis, a comunalidade é sempre 1. Nesse caso se adotar uma solução de dois fatores que explica mais de 95% das variâncias das variáveis Gás e Estrada (96,6% e 95,3% respectivamente).

Tabela 5. Variância total explicada

Componente	Autovalores iniciais			Soma das saturações ao quadrado da extração		
	Total	% da variância	% acumulado	Total	% da variância	% acumulado
1	5,303	**75,758**	75,758	5,303	75,758	75,758
2	1,014	**14,488**	**90,246**	1,014	14,488	**90,246**
3	0,390	5,570	95,816			
4	0,150	2,145	97,961			
5	0,089	1,267	99,228			
6	0,035	0,502	99,730			
7	0,019	0,270	100,000			

Observa-se na tabela 5 a porcentagem da variância explicada de cada componente e quais são as componentes que foram extraídas (aquelas cujos autovalores superam a unidade). Da totalidade das variáveis analisadas (100%), a componente 1 explica 75,758% e a componente 2 explica 14,488, componentes estas consideradas para análise. Somando a primeira componente com a segunda tem-se 90,246% da variabilidade das variáveis originais, ou seja, 90,246% da variância total é explicada pelas duas primeiras componentes.

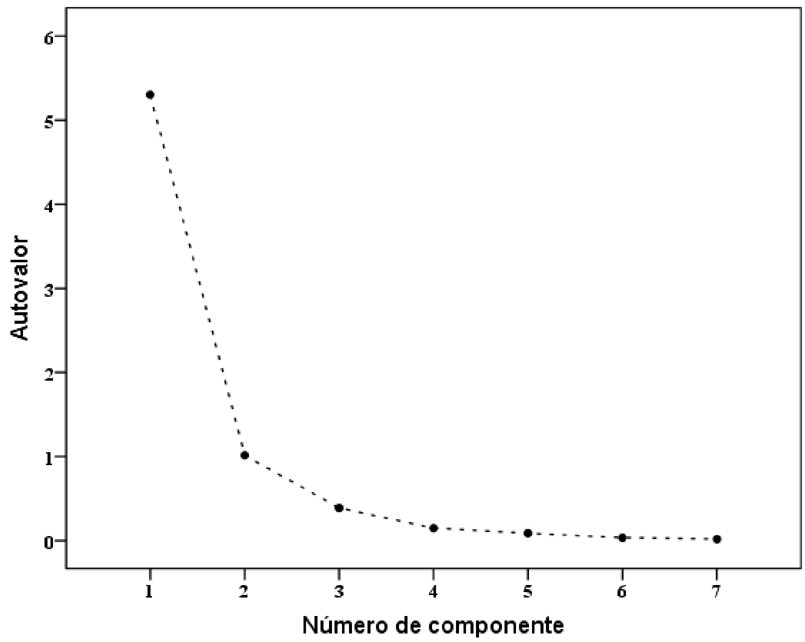

Figura 1: Gráfico de sedimentação (scree plot)

Observa-se na figura 1 a variância explicada por cada componente principal, na ordem do maior para o menor. A análise visual deste gráfico centra-se na busca de um ponto de inflexão no gráfico, o que habitualmente é produzido por valores abaixo de 1. A visão do gráfico de sedimentação justifica a seleção de dois fatores, algo que já havia procedido ao fixar o critério de autovalor na unidade.

Tabela 6. Matriz de componentes

	Componente	
	1	2
Cidade	-0,936	0,121
Estrada	-0,934	0,282
Peso	0,918	-0,203
Cilindro	0,943	0,150
Deslocamento	0,847	0,240
Gás	0,974	-0,133
Nox	0,394	0,884

Conforme a Tabela 6 apresenta as cargas de cada uma das variáveis em cada um dos fatores, de modo que os fatores com pesos mais elevados indicam uma relação mais estreita com as variáveis.

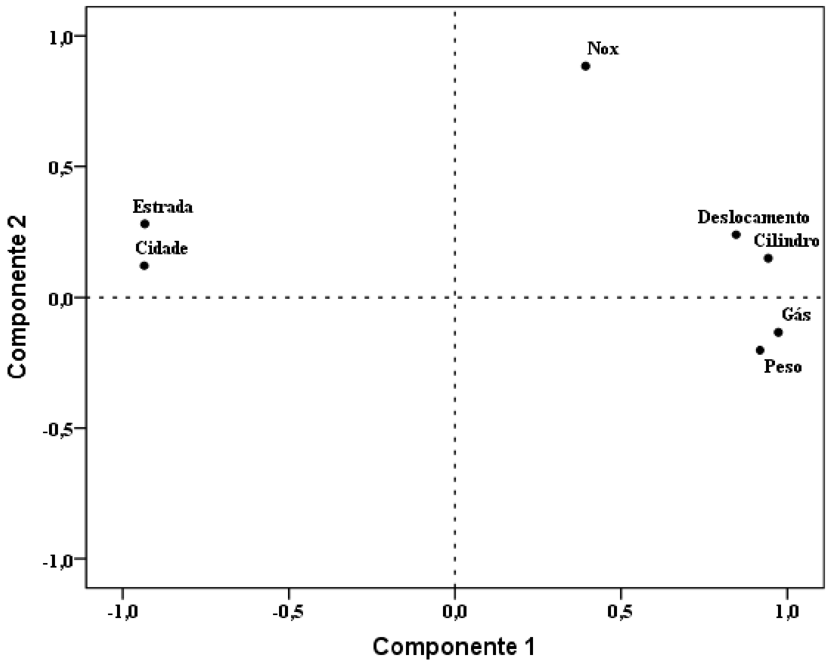

Figura 2: Gráfico da matriz de componentes

Observa-se na Figura 2 a explicação dos fatores subjacentes, possibilitando interpretar o primeiro fator (Deslocamento, Cilindro, Gás, Peso e NOX), ou seja, esta componente explica mais a característica do carro, e o segundo fator como sendo a quantidade de óxidos de nitrogênios na estrada e na Cidade.

Até agora os cálculos foram realizados sem nenhuma rotação dos fatores. Agora quando não se espera que os fatores estejam correlacionados, escolhe-se uma rotação ortogonal em que os eixos, depois da rotação passam a ser perpendiculares. O método Varimax sugerida por Kaiser (1953) é o mais utilizado. O qual tem intuito de "espalhar" os quadrados das cargas de cada fator o máximo possível. Permitindo assim encontrar grupos de coeficientes grandes e desprezíveis em qualquer coluna da matriz de cargas rotacionais.

Tabela 7. Matriz de componentes rotacionados

	Componente	
	1	2

Cidade	-0,925	-0,186
Estrada	-0,975	-0,034
Peso	0,935	0,104
Cilindro	0,845	0,446
Deslocamento	0,725	0,499
Gás	0,965	0,187
NOX	0,088	0,964

Os valores da matriz de componente rotacionados indicam que as variáveis (Cidade, Estrada, Peso, Cilindro, Deslocamento e Gás) estão altamente relacionadas com a primeira componente, sendo que segunda componente elas são baixas. Já variável NOX está relacionada com a segunda componente e sua correlação é quase zero na primeira componente.

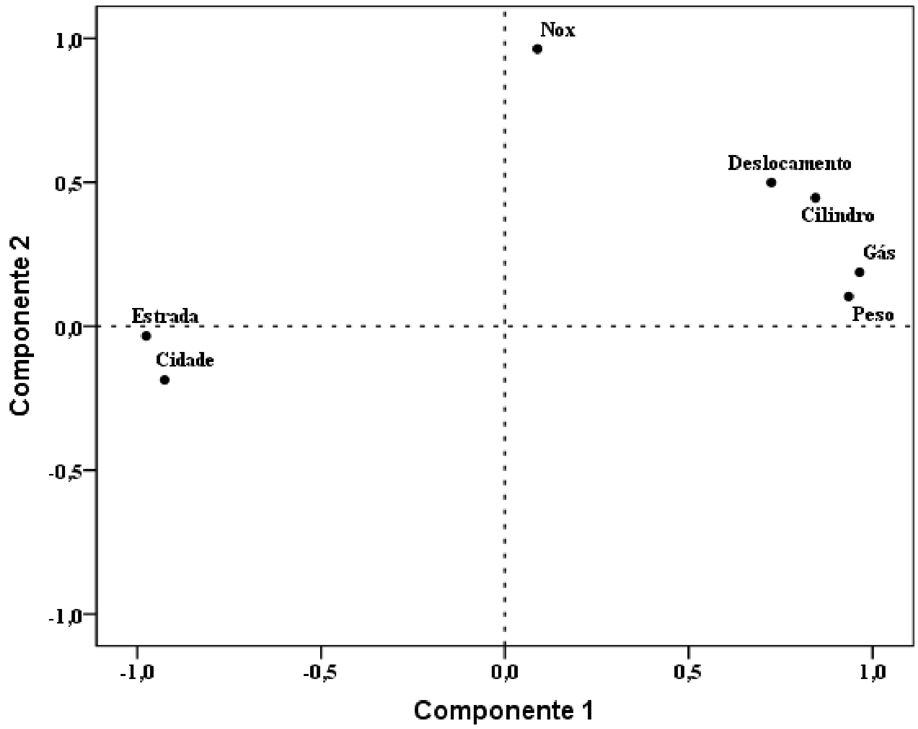

Figura 3: Gráfico da matriz de componentes no espaço rotacionado

Observa-se na Figura 3 o gráfico das cargas (pesos) das componentes no espaço rotacionado. As variáveis situam-se mais próximas da componente que em que estão mais relacionadas. A rotação mostrou pouca diferença nos dois fatores, assim a mesma conclusão da Figura 2 pode ser considerada.

4 **CONCLUSÃO**

Conclui-se que foi possível visualizar várias correlações altas, como por exemplo, Estrada com Gás e deslocamento e cidade. Enquanto que poucos pares de variáveis tiveram correlações baixas. O valor do determinante muito próximo de zero confirmou que existe correlações fortes entre várias variáveis. A prova de esfericidade de Bartlett indicou um valor de KMO = 0,795, sendo assim aceitável a adequação da análise fatorial ao conjunto de dados. Pelo teste de significância de Bartlett a hipótese nula é rejeitada, pois o valor da estatística é 172,518 com um valor p = 0,000. Confirmando também a adequação do modelo fatorial na análise dos dados

Observou-se que o primeiro fator pode ser explicado pelas seguintes variáveis: Deslocamento, Cilindro, Gás, Peso e NOX, ou seja, considera mais a característica do carro. Enquanto que o segundo fator é influenciado pela quantidade de óxidos de nitrogênios na estrada e na Cidade. Depois de realizada a rotação verificou-se que a variável NOX está relacionada com a segunda componente e sua correlação é quase zero na primeira

Não se pretende exaurir tal assunto. O intuito deste trabalho foi realizar uma simples aplicação de análise fatorial, demostrando e interpretando gráficos, bem como tabelas e figuras objetivando um aprofundamento na interpretação desta importante técnica de redução de dimensão.

REFERÊNCIAS

BARTLETT, M.S.(1947), "\Multivariate analysis", Journalof the Royal Statistical Society, Séries B, vol. 9: 176- 197.

CATTELL, R. B. (1966), "The scree test for the number of factores", Multivariate Behaviour Research, 1: 245-76.

CHATFIELD, C. e A. J. Collins (1980), Introduction to Multivariate Analysis, Chapman & Hall.

DILON, W. R. e M. Goldstein (1984), Multivariate Analysis. Methods and Applications, John Wiley & Sons.

FISHER (1936), "The use of measurements in taxonomic problems", Annals of Eugenics, vol. 7: 176-184.

GUTS, A.K; FROLOVA, Yu.V; Páutova, L.A. Método matemático en la Sociología. Série: Sinergética: del pasado al future. Moscú: Krasand, 2013.

HAIR, J. F., R. E. Anderson, R. L. Tatham e W. C. Black (1992), Multivariate Data Analysis With Readings, Maxwell Macmillan International.

HARMAN, H. H. (1976), Modern Factor Analisys, University of Chicago Press.

KAISER, H. F. (1958), "The Varimax Criterion for Analytic Rotation in Factor Analysis", Psychometrica, 23.

MANLY, B. F. J. (1986), Multivariate Statistical Methods, A Primer, Chapman & Hall.

MARDIA, M. V., J. T. Kent e J. M. Bibby (1982, 3ª ed.), Multivariate Analysis, Academic Press.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

MORRISON, D.F. (1976), Multivariate Statistical Methods, McGraw Hill.

PEARSON, K. (1901), Fundou a Revista Biometrika, fundou o Departamento de Estatística na University College London .

REIS, E. Título: Estatística Multivariada Aplicada. Edições Sílabo, Lda., 2ª Edição Lisboa, Abril de 2001

RENCHER, A. C. (1995), Methods of Multivariate Analysis, Wiley.

SPEARMAN, C. E. Psicólogo Inglês. Pioneiro da Análise fatorial e coeficiente de correlação de postos. Art. Analise fatorial da inteligência 1904.Wipedia

TRURSTONE. (1947), Multiple Factor Analysis, Univ. of Chicago Press.

WILKS, S S. (1932), “Certain generalizations in the Analysis of Variance”, Bior 24: 471-494.