Potencial de Desenvolvimento dos Municípios Fluminenses: uma Metodologia Alternativa ao IQM, com Base em Análise Fatorial Exploratória e Análise de *Clusters*

Maria Cecília Prates Rodrigues

Resumo

A análise fatorial e a análise de *clusters*, com base no estudo de técnicas de interdependência, têm se mostrado muito úteis para o entendimento da estrutura das variáveis, dos casos (unidades observacionais) e dos questionários de opinião. Certamente este entendimento pode aprimorar o processo de tomada de decisão, tanto da esfera privada quanto da pública.

Neste artigo, investiga-se a aplicação da análise fatorial exploratória e da análise de *clusters* para a identificação do potencial de desenvolvimento dos municípios fluminenses. Esta investigação é realizada tomando-se por base a mesma matriz dos dados que foi utilizada pela Fundação CIDE(1) para compor, através de metodologia própria, o seu Índice de Qualidade dos Municípios – IQM.

Comprova-se que a principal vantagem da metodologia aqui proposta é fazer com que a análise se torne mais parcimoniosa, de mais fácil interpretabilidade e menos passível de erros nas medidas dos dados.

1. Introdução

A análise multivariada é utilizada para melhor entender (1) a relação de dependência entre dois conjuntos de variáveis: um formado por variáveis dependentes (Y_1, ... Y_j) e outro por variáveis independentes (X_1, ... X_i); e (2) a relação de associação mútua entre um determinado conjunto de variáveis (X_1,....X_p) . Em se tratando de variáveis quantitativas, pode-se dizer que os modelos de regressão múltipla e de MANOVA são exemplos de técnicas multivariadas relacionadas àquele primeiro objetivo; já os modelos de análise fatorial e de análise de *clusters* estão relacionados ao segundo objetivo (Latif, p. 1).

A proposta deste artigo é procurar entender a aplicação da análise fatorial exploratória e da análise de *clusters,* enquanto métodos para a interpretação de dados, de modo a orientar o processo de tomada de decisão em questões relacionadas à Administração. Assim, o que se pretende é, a partir da base de dados utilizada pela Fundação CIDE para compor o seu Índice de Qualidade dos Municípios – IQM, segundo metodologia própria, propor uma forma alternativa de interpretação destes mesmos dados, à luz destes dois modelos multivariados. Como se sabe, o IQM pode servir como um *farol* a orientar tanto os investimentos privados quanto as políticas públicas de desenvolvimento dos municípios fluminenses.

2. Análise Fatorial Exploratória e Análise de *Clusters*: aspectos teóricos

Aaker, Kumar e Day (p. 582-608) apresentam uma explicação bastante elucidativa acerca da aplicação destes modelos. Para eles, a análise fatorial serve para combinar variáveis para criar novos fatores, os chamados fatores latentes; já a análise de *clusters* combina os objetos, ou unidades observacionais, para formar novos grupos. Em ambos os métodos, o que se pratica é a análise de interdependência, a partir da matriz de variância-covariância (ou de correlação) entre as variáveis ou entre as unidades observacionais.

No âmbito da análise fatorial, o que se pretende é a identificação de possíveis associações entre as variáveis observacionais, de modo a se definir a existência de um fator comum (latente) entre elas. Assim, pode-se dizer que a análise fatorial, ou análise do fator

comum, tem como objetivo a identificação de fatores ou *constructos* subjacentes às variáveis observacionais, o que, sem dúvidas, contribui para facilitar sobremaneira a interpretação dos dados. Isto porque, por exemplo, ao invés de buscar entender o comportamento de 20 variáveis observacionais, o analista deverá procurar entender o comportamento de 3 ou 4 fatores latentes, através do comportamento dos seus *scores* fatoriais (que será definido mais adiante, no item 4.1).

Johnson e Wichern (p.396) explicam que em análise do fator comum, as variáveis são grupadas em função de suas correlações. Isto significa que variáveis que compõem um determinado fator devem ser altamente correlacionadas entre si e fracamente correlacionadas com as variáveis que entram na composição do outro fator.

A idéia básica do modelo é a de que cada variável observacional possa ser expressa em termos do fator(es) latente(s). A tabela 1 ilustra o caso em que, após se proceder à análise dos dados, identificou-se a existência de 1 fator latente (F) comum às 5 variáveis observacionais (X_1, ..., X_5). Veja, pelo exemplo, que cada uma das cinco variáveis observacionais é explicada pelo fator comum (F) e pelo respectivo fator específico ou resíduo (e). Os coeficientes do fator (L) são as cargas fatoriais, e representam o grau de associação (ou de correlação, quando a matriz de correlação tiver sido a matriz de associação utilizada) entre a variável e o fator.

A comunalidade é o valor da carga fatorial (L) elevado ao quadrado, e representa o percentual da variância da variável que é explicada pela variância do fator comum. Quanto mais elevada for a comunalidade (varia entre 0 e 1), maior é a evidência de que X é um indicador válido do *constructo* que se deseja medir. Fazendo-se um paralelo com a análise de regressão, L seria o coeficiente de correlação entre a variável dependente e a independente, e L^2 seria o coeficiente de determinação do modelo.

Tabela 1 – O fator latente gerado a partir da análise fatorial

Identificação do fator latente	**Comunalidade**
$X_1 = L_1 F + e_1$	L_1^2
$X_2 = L_2 F + e_2$	L_2^2
$X_3 = L_3 F + e_3$	L_3^2
$X_4 = L_4 F + e_4$	L_4^2
$X_5 = L_5 F + e_5$	L_5^2

Segundo sua finalidade, a análise do fator comum pode ser do tipo exploratória ou do tipo confirmatória. No primeiro caso, o investigador dispõe apenas das variáveis e deseja identificar quantos, e quais, fatores latentes podem ser extraídos do conjunto das variáveis, através das associações entre elas. No segundo caso, como aponta Balassiano (p.1), os fatores já são conhecidos *à priori*, bem como seus indicadores, e o que se pretende é testar a adequação desses fatores.

Como já dito, será adotada a análise fatorial exploratória no âmbito deste artigo. Isso porque, o objetivo aqui será justamente o de *explorar*, ou investigar, a matriz de dados utilizada pela Fundação CIDE em seu IQM (91 municípios *vs* 38 indicadores), para identificar a existência de quantos e quais os fatores latentes existentes. A hipótese subjacente a esta opção é a de que se quer aqui entender *o que os dados estão dizendo*, independente de associações previstas para estes indicadores, com base em modelos teóricos.

Já a análise de grupamentos, ou análise de *clusters*, visa grupar indivíduos (ou unidades observacionais, ou objetos), segundo determinados critérios de distância entre os

respectivos vetores de dados. É bom lembrar que a cada unidade observacional está associado um vetor multidimensional de dados $[x_1, x_2, ... , x_p]$`.

Como afirmam Johnson e Wichern (p.573), a análise de *clusters* corresponde também a uma importante técnica exploratória, que busca identificar uma estrutura de grupamentos "naturais" de modo a avaliar a dimensionalidade dos dados, identificar *outliers* e fornecer interessantes hipóteses acerca de associações – nesse último caso, ela é usada sobretudo para grupar variáveis.

Existem dois métodos de grupamento, o hierárquico e o não-hierárquico. O método hierárquico pode ser do tipo aglomerativo (*bottom-up*) ou divisível (*top-down*). Estabelece-se uma relação de hierarquia entre o objeto (individual) e o conjunto dos objetos (*cluster*). Uma vez incorporado a um grupo, o objeto permanece associado a ele até o final do processo de *clustering*. Neste método, os critérios de grupamento mais utilizados são o da associação simples (baseada nas menores distâncias entre os objetos) e o da associação completa (baseada na maior distância entre os objetos). Cada solução de *cluster* gerada deve ser devidamente interpretada, para que se identifique qual a mais adequada para dar um significado aos dados em questão.

No método não-hierárquico, o processo de *clustering* é mais dinâmico e interativo. Diferentemente do método hierárquico, ele permite que objetos deixem um *cluster* e se associem a outro, de modo que estes reposicionamentos sucessivos possam ir contribuindo para melhorar os resultados finais. O critério não-hierárquico mais utilizado é o *K-means*, em que se define inicialmente o vetor central dos *clusters* (ou centróides) e se busca, em seguida, ir inserindo os objetos mais próximos a eles. Como se vê, ao contrário do método anterior, no *K-means* se estabelece, de antemão, o número de *clusters* com que se quer trabalhar, e se busca testar esta hipótese, a partir do significado encontrado para aquela solução de *cluster*. De certa forma, isto sugere o seu caráter de análise confirmatória.

No item 4, serão vistos os procedimentos seguidos para a análise de *cluster* dos 91 municípios do estado do Rio de Janeiro, segundo seu potencial de desenvolvimento, e se verá também porque o método *K-means* se mostrou o mais adequado.

Enfim, tanto a análise fatorial quanto a análise de *cluster* são modelos multivariados utilizados para melhor entender a associação entre variáveis e entre unidades observacionais. São técnicas antigas, cuja aplicação inicial se deu no campo das ciências sociais e comportamentais. Assim, o conceito dos fatores latentes foi primeiramente lançado por Galton em 1888, e em 1904, Spearman estendeu o conceito para o desenvolvimento da análise fatorial, quando a aplicou para melhor interpretar os testes de inteligência, conseguindo capturar o "fator da habilidade intelectual geral", conhecido como o fator G (Giri, p. 359).

No entanto, é preciso ter bem claro as limitações desses modelos. São métodos com um alto grau de subjetividade, em que várias alternativas de solução são apresentadas, cabendo ao analista optar por uma ou outra solução. Ou seja, como se verá no item 4, a qualidade da solução final fica extremamente dependente da capacidade analítica do pesquisador, na busca de uma explicação que possa ser considerada razoável, segundo a sua lógica.

3. O IQM segundo a metodologia da Fundação CIDE

Em 1998, a Fundação CIDE criou o Índice de Qualidade dos Municípios – IQM, com "*o objetivo de classificar os municípios fluminenses segundo seu potencial e condições apresentadas para o crescimento e o desenvolvimento*". Como esclarece o relatório da Fundação CIDE (p.8), o que se pretende não é medir a qualidade de vida de seus habitantes, mas a forma pela qual cada município se apresenta para receber novos investimentos.

A partir de 67 variáveis, obtidas das mais diversas fontes (como IBGE, CIDE, Secretarias de Estado, MEC, Sebrae, Firjan, Detran, Light, Sindicato dos Bancários) e com datas de

referência variadas (de 1991 a 1998), foram construídos 38 indicadores. De modo a situar o indicador no contexto do município e possibilitar a comparação entre municípios, grande parte desses indicadores são apresentados de forma relativa. Exemplificando, o indicador BAN refere-se às agências bancárias no município, sendo descrito como o "número de agências bancárias dividido pela raiz da população e multiplicado por 100".

De acordo com a metodologia descrita pela Fundação CIDE (p.13-22), os indicadores foram escolhidos em função de sua (1) representatividade, ou capacidade de representar um determinado fenômeno, e (2) disponibilidade e periodicidade de atualização.

Os 38 indicadores foram distribuídos em 7 grupos, conforme apresentados na tabela 2. Foram, também, atribuídos pesos aos indicadores e aos grupos, "*de modo a refletir a importância que se desejou conferir a cada um dos aspectos considerados, tendo em vista a base teórica adotada*". Esclarece-se que a base teórica adotada foi a Teoria das Localidades Centrais, de Christaller, e a Teoria dos Polos de Desenvolvimento, de Perroux. Sobre a definição dos indicadores, ver o anexo 1.

Tabela 2 – IQM: Grupos, indicadores e pesos

Grupos e pesos	**Indicadores e pesos**
i. Centralidade e vantagem locacional - CEN (peso 10)	CVA (10); CON (7); ONI (9); MES (6); CAT (8)
ii. Qualificação da mão-de-obra - QMA (peso 9)	CES (10); ALF (7); PRO (8); TEC (9)
iii. Riqueza e potencial de consumo - RIQ (peso 9)	CRA (9); PIB (10); ENE (7); DEP (8); FMU (6)
iv. Facilidades para negócios - FAC (peso 8)	BAN (9); TEL (8); COR (7); SEB (5); HOT (6); INC (10); INT (4)
v. Infra-estrutura para grandes empreendimentos - IGE (peso 8)	ROD (8); FE (7); AER (5); GAS (6); LIN (9); DIS (10)
vi. Dinamismo - DIN (peso 7)	CRE (10); B24 (7); VEI (7); OPC (9); PIC (10)
vii. Cidadania - CID (peso 6)	ENS (9); LEI (7); SEG (5); CUL (6); JUS (8); DOM (10)

Nota: Os indicadores encontram-se abreviados; sobre a descrição dos indicadores, ver o anexo 1.

Na realidade, o que essas teorias fazem é apontar, de modo bem genérico, a importância da organização espacial segundo centros polarizadores. Não se pode dizer que exista uma relação direta entre o que a teoria propõe e a forma como o IQM foi construído. Isto significa que a equipe da Fundação CIDE traduziu aquela teoria na forma do IQM, entendido como o processo de seleção de indicadores, definição dos grupos, alocação dos indicadores aos grupos e a atribuição de pesos. Outra equipe de outra instituição poderia ter feito outra leitura da teoria bastante distinta, com outros indicadores, outros grupos, outros pesos, etc.... Ou até com os mesmos indicadores, mas outros pesos e outros grupos. Em outras palavras, o que se quer dizer é que a composição do IQM é apenas uma forma subjetiva de percepção da Teoria das Localidades Centrais e da Teoria dos Polos.

Em linhas gerais, a metodologia para o cálculo do IQM segue a metodologia que vem sendo utilizada pelo Programa das Nações Unidas – PNUD para a estimativa do Indice de Desenvolvimento Humano – IDH, desde 1990. O IQM é obtido da média ponderada (pelos respectivos pesos) dos índices calculados para os grupos; estes, por sua vez, são calculados através da média ponderada dos índices para os indicadores, que compõem cada um dos

grupos. Já os índices para cada indicador são obtidos por interpolação linear; ou seja, ao melhor resultado apurado para o indicador é atribuído valor igual a um, e ao pior resultado é atribuído valor igual a zero. Desnecessário dizer que o valor para cada um destes indicadores constituídos varia entre zero e um.

Para a melhor compreensão desta metodologia de formação de índices, veja os três exemplos a seguir (tabela 3):

Tabela 3 - Exemplos de formação de índices

Ex.1 – índice para cada indicador Indice da taxa de alfabetização dos maiores de 15 anos de idade – IALF IALF = ALF – menor (ALF) / maior (ALF) – menor (ALF) Onde ALF é a taxa de alfabetização de um dado município; menor (ALF) é a menor taxa de alfabetização dentre todos os 91 municípios; e maior (ALF) é a maior taxa de alfabetização dentre todos os municípios.
Ex. 2 – Índice setorial (dos grupos) Indice de Qualificação da mão-de-obra – IQMA IQMA = (10.ICES + 7.IALF + 8.IPRO + 9.ITEC) / 34
Ex. 3 – IQM IQM = (10.CEN + 9.IQMA + 9.IRIQ + 8.IFAC + 8.IGE + 7.DIN + 6.CID) / 57

A tabela 4 apresenta o *ranking* dos 91 municípios fluminenses, obtido a partir desta metodologia aplicada pela Fundação CIDE para determinar o potencial de desenvolvimento de cada um dos municípios. Notar que o número de municípios alocados em cada coluna equivale ao número de municípios definidos para cada um dos quatro grupos a partir da análise de *cluster* (item 4), de modo a facilitar ao leitor a comparação dos resultados obtidos a partir desses dois métodos de análise.

Tabela 4 – *Ranking* dos municípios fluminenses, em ordem decrescente do IQM

1) Rio de Janeiro	16) Itaguaí	45) Aperibé	75) Cachoeiras de Macacu
2) Niterói	17) Angra dos Reis	46) Pinheiral	
3) Resende	18) Piraí	47) Arraial do Cabo	
4) Macaé	19) Duque de Caxias	48) Paraíba do Sul	76) Paty do Alferes
5) Volta Redonda	20) Rio das Ostras	49) Cordeiro	77) Bom Jardim
6) Petropólis	21) Araruama	50) Magé	78) Laje do Muriaé
7) Casimiro de Abreu	22) Bom Jesus do Itabapoana		79) Silva Jardim
8) Três Rios		51) Areal	80) Japeri
9) Miguel Pereira	23) Nova Iguaçu	52) Maricá	81) Cardoso Moreira
10) Campos	24) Vassouras	53) Quissamã	82) Duas Barras
11) Itatiaia	25) Nova Friburgo	54) Cantagalo	83) S. José do Vale do R.Preto
12) Cabo Frio		55) S. Pedro d`Aldeia	
13) Barra Mansa	26) Valença	56) Porciúncula	84) Tanguá
14) S. Ant. de Pádua	27) Miracema	57) Eng.Paulo Frontin	85) Trajano de Morais
15) Teresópolis	28) Mangaratiba	58) Conceição de Macabu	86) Sta. Maria Madalena
	29) Barra do Piraí		87) Varre-Sai
	30) Búzios	59) Guapimirim	88) S. Sebastião do Alto
	31) Mendes	60) Saquarema	
	32) Iguaba Grande	61) Rio Claro	89) S. José de Ubá
	33) Rio Bonito	62) Itaboraí	90) Sumidouro
	34) Itaperuna	63) Parati	
	35) Nilópolis	64) Carmo	91) São Francisco do
	36) São Gonçalo	65) Natividade	

	37) Paracambi 38) Queimados 39) Rio das Flores 40) C. Levy Gasparian 41) Itaocara 42) S. João de Meriti 43) Porto Real 44) Belford Roxo	66) Cambuci 67) Quatis 68) Italva 69) Macuco 70) S. João da Barra 71) São Fidélis 72) Sapucaia 73) Seropédica 74) Carapebus	Itabapoana

Feita esta breve descrição sobre a metodologia do IQM aplicada pela Fundação CIDE e sobre os principais resultados encontrados, propõem-se aqui algumas questões para reflexão, o que deverá ser feito ao longo do item 4. São elas:

- Para avaliar o potencial de desenvolvimento dos municípios, seria mesmo necessário este número tão grande de indicadores (38), que pode, inclusive, comprometer a qualidade dos resultados?
- Os 7 grupos constituídos representam, de fato, a melhor forma de grupar estes indicadores? Ou haveria outra alternativa válida, mais parcimoniosa e de fácil interpretação?
- Haveria outra maneira de sistematizar os municípios segundo o seu potencial de desenvolvimento, que fosse coerente e lógica?

4. Uma metodologia alternativa à do IQM

À luz da análise fatorial exploratória, se procurará identificar alguns poucos fatores latentes subjacentes aos 38 indicadores (ou variáveis observacionais) utilizados no IQM – item 4.1. Estes fatores correspondem, no IQM, aos 7 índices setoriais, e têm por finalidade capturar os principais *constructos* relacionados à idéia do potencial de desenvolvimento.

Com base na análise de *clusters*, se procurará uma forma lógica de organizar os municípios, segundo o seu potencial de desenvolvimento – item 4.2. Os *clusters* de municípios a serem formados cumprem a mesma finalidade do *ranking*, em termos do IQM.

O *software* aqui utilizado para a análise dos dados foi o *SPSS 10.0 for Windows*, em sua versão em inglês.

4.1 Aplicação de análise fatorial exploratóra

Inicialmente, é apresentado um roteiro básico para se proceder à análise fatorial exploratória. A seguir, são descritas as principais etapas que foram conduzidas até se chegar à solução que foi aqui considerada a mais adequada.

Roteiro básico

(1) Verificação da adequação dos dados à análise fatorial, através de:

- Análise da matriz de correlação: na matriz, cada indicador deve apresentar correlação elevada com pelo menos alguns indicadores, não necessariamente com todos. Isto significa que este grupo de indicadores correlacionados têm um *constructo* em comum, capturado pelo fator comum. Se a correlação de um determinado indicador for baixa com todos os outros, isto quer dizer que ele não traduz, juntamente com qualquer outro indicador, qualquer idéia em comum. Um valor de correlação pode ser considerado aceitável acima de 0,4.
- Teste KMO: a medida Kaiser-Meyer-Olkin testa a adequação da amostra quanto ao grau de correlação parcial entre as variáveis, que deve ser pequeno. Se isto ocorre, significa que os fatores latentes explicam grande parte da associação entre

as variáveis, e os resíduos estão pouco associados entre si. Valores para o teste KMO iguais ou inferiores a 0,7 indicam que a análise fatorial pode ser inadequada.

- Teste de esfericidade de Bartlett: neste teste, a hipótese inicial (H_0) é que a matriz de correlação é uma matriz-identidade, indicando que o modelo é inadequado. Se por exemplo, para um nível de significância definido em 0,05, a significância (α) encontrada for menor do que 0,05, deve-se rejeitar H_0, concluindo-se, portanto, que o modelo é adequado em função das associações verificadas.

(2) Determinação do número de fatores latentes – alguns critérios básicos podem ser seguidos para a extração dos fatores latentes mais relevantes, tais como: fatores com autovalores (L), associados à matriz de associação, maiores do que 1 (para este critério, a matriz de associação analisada, $\sum$, deve ser a de correlação); a "regra do cotovelo" no *scree plot;* e a variância acumulada igual ou acima de 70%.

(3) Análise da solução fatorial: o valor da comunalidade extraída para as variáveis deve ser razoável (pelo menos, acima de 0,5). Também devem ser elevados os valores das cargas fatoriais obtidos na matriz dos fatores *rotados*, isto é depois da rotação dos eixos, pois são justamente essas cargas que vão auxiliar na interpretação dos fatores.

(4) Interpretação do significado dos fatores: sugere-se a rotação nos eixos justamente para facilitar a interpretação dos fatores. Os métodos de rotação mais utilizados são: (1) Varimax, em que se faz uma rotação ortogonal dos eixos; e (2) Oblimin, em que se promove uma rotação oblíqua nos eixos. A idéia aqui é verificar qual o método de rotação que propicia a interpretação mais plausível para os fatores. Pode ocorrer também que nenhum dos métodos facilite a interpretação; neste caso, deve-se repensar se o desenho da análise fatorial, que se está utilizando, é um procedimento metodológico válido para os dados em questão.

(5) Obtenção dos *scores* fatoriais – os *scores* fatoriais são os valores, para cada unidade observacional, assumidos pelo fator latente. Os valores do *score* fatorial resultam da combinação linear entre cada um dos valores das variáveis observacionais e os respectivos coeficientes do score fatorial (obtidos na matriz dos coeficientes do *score* fatorial). Aaker, Kumar e Day (p. 589) aconselham o uso dos *scores* fatoriais, ao invés das variáveis originais, em análises ou interpretações subsequentes das variáveis.

Principais etapas seguidas:

1ª etapa

Foi aplicada a análise fatorial à base de dados como um todo, ou seja, aos 38 indicadores e 91 municípios.

Não se pode, de forma alguma, concluir pela adequação dos dados, mesmo com os resultados favoráveis do teste KMO (0,759) e do teste de Bartlett (alfa < 0,001). A matriz de correlação mostrou que existem muitos indicadores com correlação bastante baixa, ou praticamente nenhuma, com todos os demais indicadores. Nesta situação, estão indicadores como: AER, CAT, CRE, FMU, ENS, GAS, INC, LEI, LIN, MES, ONI, PIB, PIC e PRO. Isto sugere que vários dos indicadores utilizados se relacionam muito superficialmente com todos os demais indicadores. Neste sentido, eles podem ser considerados indicadores não válidos ou espúrios, não contribuindo para medir o que realmente se deseja.

Além disso, a solução fatorial (método de rotação *Varimax*) se mostrou muito insatisfatória, haja vista, sobretudo, a falta de significado dos fatores, buscado na lógica das cargas fatoriais dos indicadores que os compõem. Além disso, as baixas comunalidades encontradas para grande parte das variáveis sinaliza que os 9 fatores extraídos (com autovalores maiores do que 1) explicam, juntos, muito pouco a variância total dos indicadores; basta ver que a comunalidade é inferior a 0,5 em 14 dos 38 indicadores.

Concluiu-se, portanto, pelo não ajuste do modelo.

2ª etapa

Observando-se o quadro das estatísticas descritivas apuradas na 1ª etapa, pode-se levantar a hipótese de que a baixa qualidade dos dados poderia ser uma das causas para o não-ajuste do modelo acima. Notar que em 22 dos 38 indicadores, o desvio-padrão é igual ou maior do que a média, sinalizando que os dados estão muito dispersos. A prática tem mostrado que, em uma distribuição relativamente homogênea, esta relação entre o desvio-padrão e a média, conhecida como coeficiente de variação, situa-se em até 30%.

Sem dúvida, esta baixa qualidade dos dados é fruto de uma dispersão real dos resultados para os indicadores entre os municípios. Como seria de se esperar, por exemplo, os resultados para o município do Rio de Janeiro são os que mais se distanciam da média da distribuição, em razão do seu próprio dinamismo, atribuído ao fato de ser a sede da capital do Estado. Mas, suspeita-se também que tenha havido falhas na entrada dos dados, o que pode prejudicar muito o tratamento dos dados aqui pretendido, através de análise fatorial e análise de *clusters*. Apenas a título de exemplificação, citam-se alguns exemplos destas possíveis falhas, ou seja, resultados que não parecem condizer com a realidade:

- No que se refere ao indicador MES (percentual de matrículas do ensino superior), a média do estado é de 6,7% e, no entanto, o dado para Seropédica é de 198,3%.
- Para o indicador PIB (PIB per capita em R$ de 1996), o valor de Piraí é o mais elevado, de 44.828, sendo a média do Estado de 4.790.

Considerou-se, portanto, fundamental proceder-se ao ajuste dos dados. O critério adotado foi que os resultados em que os seus correspondentes valores padronizados superassem a mais ou menos 2 desvios-padrões, seriam considerados *missing* (sem dados).

Feito o ajuste dos dados, *rodou-se* novamente a análise fatorial com os dados ajustados, isto é, valores *missing* no lugar dos *outliers* (no SPSS, adotou-se a opção de substituir os *missing values* pela média do indicador). Apesar disso, o modelo da análise fatorial não chegou a apresentar melhora sensível, que levasse à sua aceitação. Inicialmente, em se considerando a condição do autovalor maior do que 1, foram gerados 10 fatores, mas com cargas fatoriais fracas e de difícil interpretabilidade (mesmo utilizando- se a matriz dos fatores *rotados*). Tentou-se reduzir o número de fatores, de modo a melhorar a interpretabilidade dos fatores, chegando-se a 4 fatores: nesta situação extrema, o valor acumulado dos autovalores ficou muito baixo (48%), razão pela qual se decidiu novamente pelo não ajuste do modelo.

3ª etapa: Solução encontrada

Se o ajuste dos dados ainda não fora suficiente para permitir a aplicação do modelo, por que não se partir para o ajuste dos indicadores?

Já que a matriz de correlação, através das baixas correlações detectadas, havia apontado para a existência de indicadores não-válidos, e portanto inadequados, por que não excluir alguns destes indicadores? Mesmo porque a prática tem mostrado que a situação ideal para se aplicar a análise fatorial é a presença de um número não muito elevado de variáveis, porém o maior número possível de unidades observacionais.

Decidiu-se, então, pela utilização de apenas 15 indicadores, ou seja, daqueles que apresentam os níveis mais elevados de correlação com os demais.

Desta vez, o modelo se mostrou relativamente ajustado. Houve "melhora" nas estatísiticas descritivas, no teste KMO, nas comunalidades obtidas e na matriz de correlação. As aspas são propositais, e visam alertar ao leitor que os resultados encontrados ainda ficaram longe de uma solução ideal – apesar do teste KMO e de Bartlett estarem satisfatórios, o

desvio-padrão ainda seguiu sendo maior do que a média em 4 dos 15 indicadores, e 4 dentre as 15 comunalidades extraídas ficaram entre 0,4 e 0,5. Mas, a sensível melhora conseguida foi na interpretabilidade dos dados, pois as cargas fatoriais da matriz *rotada* permitiram, com facilidade, identificar o significado dos três fatores latentes capturados. Senão, veja os resultados na tabela 5:

Tabela 5 – Matriz dos fatores *rotados* – método Varimax

Indicadores	Fatores		
	1	2	3
BANM	,787		
CVAM	,686		
CONM	,676		
DEPM	,652		
B24M	,647		,445
JUSTM	,631		
OPCM	,566		
SEBM	,557		
CRAM		,811	
TELM		,667	
VEIM		,633	
INTM		,487	
ALFM			,830
CESM		,567	,625
DOMM			,558

Método de extração: Fatores principais. A rotação convergiu após 7 interações.
Obs.: O "M" acrescido à abreviatura de cada indicador siginifica que estes indicadores estão com os seus valores *outliers* como *missings.*

Assim, os três fatores capturados conseguem explicar 66% da associação total entre os dados para os 15 indicadores (valor acumulado dos autovalores). Portanto, em análises futuras, ao invés de se trabalhar com os 15 indicadores, poder-se-ia trabalhar apenas com os 3 fatores, ou seja com os *scores* fatoriais gerados, sabendo que se estaria incorrendo em uma perda de 34% na associação entre os dados.

O fator 1 mostra as condições de DINAMISMO do município, e foi constituído pelos seguintes indicadores: agências bancárias, consumo varejista, concessionárias de veículos, depósitos bancários, postos de banco 24 horas, acesso à justiça, operações de crédito efetuadas, existência de balcões Sebrae.

O fator 2 transmite a idéia de POTENCIAL DE CONSUMO do município, e foi formado pelos indicadores: chefes de domicílios com renda elevada, terminais telefônicos, veículos novos e provedores de Internet.

O fator 3, interpretado como CONDIÇÕES DE VIDA da população, foi constituído pelas variáveis: taxa de alfabetização, chefes de domicílios com escolaridade razoável e domicílios em condições adequadas.

A idéia é que estes três fatores juntos possam cumprir o papel do IQM proposto pela Fundação CIDE, qual seja o de dar a idéia do potencial de desenvolvimento do município. Com efeito, quando se *roda* a análise fatorial com os *scores* dos 3 fatores latentes, um fator apenas é extraído, que representa justamente o POTENCIAL DE DESENVOLVIMENTO dos municípios.

4.2 Aplicação da análise de *clusters*

Será feito um breve relato das etapas tentativas para se proceder à análise de *cluster* dos dados, até se chegar à solução que foi considerada a mais adequada.

1ª etapa

Mesmo já sabendo que os dados apresentam uma grande dispersão, resolveu-se tentar a análise de *clusters* com as variáveis observacionais originais, ou seja, sem dar qualquer tipo de tratamento aos *outliers* já identificados. Para essa análise de *clusters*, levou-se em consideração, inicialmente, os 15 indicadores selecionados anteriormente para a análise fatorial. Em seguida, foram considerados apenas os 3 indicadores que apresentaram as cargas fatoriais mais elevadas em cada um dos 3 fatores latentes. Em ambas as tentativas, a solução não foi considerada razoável, ocorrendo uma concentração grande de municípios em determinados grupos.

O que essas tentativas mostraram foi que, já que o objetivo em questão era identificar uma certa dimensionalidade (ou lógica) entre os municípios, e não a identificação de *outliers*, a inclusão destes *outliers* acabou se tornando um elemento perturbador a mais para a análise dos *clusters* de municípios.

2ª etapa

Nessa etapa, procurou-se dar tratamento aos *outliers*, que passaram a ser considerados como *missings*. A melhor solução aqui encontrada, levando-se em consideração os 15 indicadores selecionados anteriormente, foi através do método K-means, utilizando-se a opção (do SPSS) de inclusão dos casos com dados *missing*.

Não houve, desta vez, a super-concentração de municípios em 1 ou 2 *clusters*; a distribuição dos municípios entre os *clusters* foi bem mais equilibrada para os 3 centróides inicialmente definidos. Mas a qualidade do grupamento não foi considerada satisfatória. Assim, pelo que se conhece em termos de nível de desenvolvimento destes municípios, é inconcebível juntar em um mesmo grupo municípios tão díspares como Bom Jesus do Itabapoana, Itaperuna, Natividade, Rio de Janeiro, Resende e Volta Redonda. A qualidade dos grupamentos também não apresentou melhora com a definição de diferentes números de centróides iniciais.

3ª etapa

Nesta etapa, foram feitas tentativas de formação dos *clusters* a partir dos *scores* fatoriais gerados para os 3 fatores latentes, identificados na 3ª etapa do item 4.1. É bom lembrar que aquela solução dos três fatores levou em consideração o tratamento dado pelo SPSS aos *outliers*, em que os valores *missing* foram substituídos pela média do indicador (item 4.1, 2ª etapa).

Tanto pelo método hierárquico ("*between-groups linkages*") como pelo método não-hierárquico *K-means*, as soluções a partir dos *scores* fatoriais não foram consideradas satisfatórias. Uma possível explicação é que os *scores* fatoriais ajudam na interpretação dos dados mas, nesta situação em particular, não serviram para a delimitação dos *clusters* dos casos ou unidades observacionais.

4ª etapa: Solução encontrada

A solução aqui encontrada, considerada a mais adequada, apresenta idéias da etapa 1 e da etapa 2 desta análise de *clusters*. Da etapa 1, a contribuição trazida foi a de trabalhar com os três indicadores com as cargas fatoriais mais elevadas de cada um dos fatores latentes. Com efeito, Aaker, Kumar e Day (p. 596) aconselham que, em algumas situações, o analista pode, e deve, usar uma ou duas variáveis com as cargas mais elevadas na composição do

fator, de modo a representar o fator em coletas de dados ou análises subsequentes. Já da etapa 2, a idéia trazida foi a do tratamento dado aos *outliers* no âmbito do método *K-means.*

Assim, os *clusters* formados tiveram por base os indicadores de relação entre agências bancárias e população (BAN), percentual dos chefes de domicílios com rendimentos acima de 20 salários-mínimos (CRA) e taxa de alfabetização dos maiores de 15 anos de idade (ALF). Isto significa que os municípios foram grupados por um vetor de 3 variáveis observacionais, que são, em princípio, fortemente representativas dos níveis de dinamismo do município, do seu potencial de consumo e das condições de vida de sua população. Estas três variáveis podem, portanto, ser consideradas representativas do potencial de desenvolvimento dos municípios, que é o conceito subjacente aos 3 fatores latentes.

No que se refere aos dados levantados para estes três indicadores, é interessante notar que, no que se refere à taxa de alfabetização, 4 municípios tiveram dados *missing* devido a suas taxas discrepantemente baixas[(2)]. Foram eles: Cardoso Moreira, São Francisco do Itabapoana, Silva Jardim e Sumidouro. Taxa de alfabetização baixa é um dos sinais de baixo potencial de desenvolvimento: apesar do tratamento para dados *missing* do método *K-means* aqui utilizado, pode-se dizer que ele não conseguiu mascarar este atraso, pois estes municípios foram, de fato, alocados no grupo de menor potencial de desenvolvimento. Pela razão oposta, resultados discrepantemente elevados para os indicadores CRA (Niterói e Rio de Janeiro) e BAN (Rio de Janeiro) apresentaram dados *missing* e, também, apesar do tratamento conferido para este tipo de dado, estes municípios foram alocados no grupo de maior potencial de desenvolvimento. Mas, por outro lado, é bom ter claro que a definição e o tratamento aos *outliers* influiu na localização dos municípios dentro do *cluster*, reduzindo suas distâncias em relação ao centróide.

Resumindo, pode-se dizer que a melhor solução foi encontrada a partir da aplicação do método não-hierárquico *K-means*, em que foram considerados os 3 indicadores com as cargas fatoriais mais elevadas de cada um dos fatores; em que 6 municípios (dentre os 91) com dados *missing* foram incluídos nos *clusters;* e em que foram definidos 4 centróides iniciais.

Na realidade, esta solução foi a melhor, porque conseguiu discriminar bem entre os 91 municípios, quanto ao seu potencial de desenvolvimento. Em outras palavras, os resultados encontrados coincidiram, em grande medida, com os resultados esperados. Na tabela 6, estão apresentados os 4 *clusters* com os municípios que os compõem, segundo seu potencial de desenvolvimento.

É importante entender que, em cada um dos quatro grupos, os municípios se encontram em ordem crescente de sua distância em relação ao centróide do grupo, e não em ordem decrescente segundo o potencial de desenvolvimento. Isto significa que quanto mais próximo ao centróide, mais bem adaptado o município se encontra no grupo. E quanto mais afastado do centróide, menos adaptado ao *cluster*, em função do seu vetor de indicadores que pode estar, relativamente ao *cluster*, ou muito bom ou muito ruim – esses municípios mais afastados estão na chamada "linha de transição" entre o *cluster* em que foram inseridos e os *clusters* vizinhos. Exemplificando, e tendo por base os resultados do município para estes 3 indicadores selecionados, pode-se dizer que Macaé, alocado no *cluster* 2, está na área de transição com o grupo 1; enquanto Tanguá, alocado no grupo 3, fica na região de transição com o grupo 4.

Tabela 6 – Os 4 clusters de municípios fluminenses, segundo seu potencial de desenvolvimento

Grupo 1:	**Grupo 2**:	**Grupo 3:**	**Grupo 4:**
Barra Mansa	Areal	Miracema	Silva Jardim
Nova Iguaçu	Paraíba do Sul	Saquarema	Trajano de Morais
São Gonçalo	Maricá	Casimiro de Abreu	S. Francisco do Itabapoana
Barra do Piraí	Queimados	S. Antônio de Pádua	

Resende	Itatiaia	Araruama	Sumidouro
Três Rios	Angra dos Reis	Rio das Ostras	Cardoso Moreira
São João de Meriti	Eng. Paulo Frontin	Itaocara	Paty do Alferes
Volta Redonda	Magé	Bom Jesus do Itabapoana	Laje do Muriaé
Duque de Caxias	Itaguái	Natividade	Cambuci
Petrópolis	Campos dos Goytacazes	Cachoeiras de Macacu	Bom Jardim
Nova Friburgo	Cabo Frio	S. José do Vale do Rio Preto	Varre-Sai
Pinheiral	Seropédica	Parati	Duas Barras
Rio de Janeiro	Miguel Pereira	Quissamã	S. Maria Madalena
Nilópolis	Mangaratiba	S. João da Barra	Rio Claro
Niterói	Mendes	Sapucaia	Porciúncula
	Comendador Levy	Guapimirim	Carapebus
	São Pedro d Aldeia	Armação dos Búzios	S. José de Ubá
	Itaborái	Aperibé	São Sebastião do Alto
	Cordeiro	Paracambi	
	Arraial do Cabo	Cantagalo	
	Valença	Itaperuna	
	Macuco	Piraí	
	Teresópolis	Rio das Flores	
	Porto Real	Carmo	
	Quatis	São Fidélis	
	Macaé	Rio Bonito	
	Belford Roxo	Tanguá	
	Vassouras	Italva	
	Iguaba Grande	Japeri	
		Conceicão de Macabu	

5. Conclusões

Analisando os municípios que compõem os quatro *clusters* identificados, conclui-se que o grupo 1 tendeu a concentrar os (15) municípios com maior potencial de desenvolvimento, que são justamente os da Região Metropolitana do Rio de Janeiro e os da industrializada região sul. No outro extremo, no grupo 4, ficaram os (17) municípios que apresentam atualmente menor potencial de desenvolvimento, que são sobretudo os da região norte e noroeste do Estado. Grande parte dos municípios das regiões serrana, litoral e central foi alocada nos grupos intermediários 2 e 3, sendo que os (29) municípios do grupo 2 tenderam a apresentar melhor situação do que aqueles do grupo 3 (30 municípios).

Assim, no que se refere ao grupamento dos municípios em termos de potencial de desenvolvimento, pode-se dizer que existe um certo grau de comparabilidade entre os resultados obtidos a partir da metodologia do IQM e da aplicação da análise fatorial exploratória e de *cluster*. Tomando-se, por exemplo (tabelas 4 e 6), o caso dos municípios situados em posição extrema, observa-se que 7 dentre os 15 municípios alocados no 1º *cluster* encontram-se também entre os 15 municípios mais bem classificados segundo o IQM. Por outro lado, dentre os 17 municípios do 4º *cluster*, 13 deles estão também entre os 17 municípios piores classificados pela Fundação CIDE. Ou seja, ao todo, nestas duas posições extremas, 62,5% dos municípios são comuns.

A questão que se coloca, portanto, é qual a vantagem de se utilizar uma ou outra metodologia.

Como visto, em ambos os modelos, a subjetividade está presente, porém, de maneira diferente. Quando se usa o IQM, existe uma boa dose de subjetividade na entrada dos dados, ou seja, na definição e na atribuição de pesos aos indicadores e grupos. Por outro lado, quando se trabalha com a análise fatorial e de *clusters*, a subjetividade entra na interpretação dos

dados, ou seja, na análise dos resultados obtidos a partir da associação entre os indicadores e da distância entre as unidades observacionais.

Nesse sentido, a aplicação da análise fatorial mostrou que bastavam apenas 15 indicadores e 3 grupos (os 3 fatores latentes) para capturar a idéia do potencial de desenvolvimento dos municípios. Ou seja, não era preciso número tão grande de indicadores (38) e grupos (7) para a composição do IQM. A vantagem dessa redução no número de indicadores e de grupos é que a análise se torna mais parcimoniosa, menos passível de erros nas medidas dos dados (pois existem menos indicadores a serem incluídos) e de mais fácil interpretabilidade.

Finalmente, chama-se a atenção para três procedimentos interessantes, que foram utilizados no âmbito da análise fatorial e da análise de *clusters* neste estudo empírico:

- Eliminação de indicadores pouco representativos ou pouco válidos do(s) conceito(s) que se deseja capturar, a partir da análise da matriz das correlações do conjunto dos indicadores.
- Tratamento dos dados *outliers*, que passam a ser considerados como valores *missing*, e que atuam como elementos perturbadores da análise.
- Definição dos *clusters* a partir das variáveis observacionais mais *carregadas* em cada fator, e não de todas as variáveis que compõem o fator, como é o mais usual.

6. Referências bibliográficas

AAKER, David. KUMAR, V. DAY, George. **Marketing Research**. John Wiley & Sons, Inc. 6th edition, 1998.

BALASSIANO, Moisés. **Análise Fatorial**. mimeo. FGV, 2000.

CIDE, Centro de Informações e Dados do Rio de Janeiro. **IQM – Índice de Qualidade dos Municípios.** Rio de Janeiro: CIDE, 1998.

GIRI, Narayan. **Multivariate Statistical Analysis**. New York: Marcel Dekker, Inc. 1996.

JOHNSON, Richard. Wichern, Dean. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall. 3rd edtion, 1992.

LATIF, Sumaia Abdei. **A análise fatorial auxiliando a resolução de um problema real de pesquisa de marketing**. São Paulo: Caderno de Pesquisas em Administração, v. 00, nº 0, 2º semestre1994.

Notas finais

(1) A Fundação CIDE é o Centro de Informações e Dados do Rio de Janeiro, ligado à Secretaria de Estado de Planejamento e Controle.

(2) Na realidade, os resultados referentes a estes 4 municípios chegaram a ser excluídos porque ALF apresentou distribuição bastante homogênea (coeficiente de variação igual a 8,4%). Em estudos futuros, deve-se pensar na definição do *outlier* de modo a incorporar diferenças no grau de homogeneidade da distribuição.

Anexo 1

Descrição dos indicadores utilizados no IQM

Código	Indicador
AER	Existência de aeroporto
ALF	Taxa de alfabetização da população de 15 anos ou mais
B24	Pontos de serviços bancários com atendimento 24 horas, em relação à raiz da população
BAN	Agências bancárias dividido pela raiz da população e multiplicado por 100
CAT	Raiz quadrada do Valor Adicionado Fiscal do Comércio Atacadista per capita

CES	Percentual de chefes de domicílios com, pelo menos, 2º grau completo
CON	Concessionárias de veículos no município, em relação à raiz da população, multiplicado por 100
COR	Agência dos Correios multiplicada por 2 mais posto de venda de selos dividido pela raiz da população e multiplicada por 100
CRA	Percentual de chefes de domicílios com renda superior a 20 salários mínimos
CRE	Taxa média geométrica de crescimento anual da população residente, entre 1991 e 1996
CUL	Soma dos números de cinemas, teatros, bibliotecas, dividido pela raiz da população e multiplicado por 100
CVA	Atratividade do município para o suprimento de bens de consumo em geral
DEP	Média dos depósitos bancários em agências do município por habitante
DIS	Existência de distritos, condomínios, polos ou parques industriais
DOM	Percentual médio de domicílios com abastecimento de água adequado, com esgotamento sanitário adequado e com coleta de lixo
ENE	Consumo residencial de energia elétrica por habitante
ENS	Matrículas no ensino básico em relação à população residente em idade escolar
FER	Existêcia ou próximidade de linha férrea
FMU	Capacidade de investimento. Relação entre as despesas de capital com investimentos e a população
GAS	Existência de gasoduto
HOT	Leitos de hotel para cada 1.000 habitantes
INC	Pontuação pela política municipal de incentivos
INT	Provedor de acesso à INTERNET em relação ao tempo de um pulso e à população
JUS	Existência de PROCON e Defensoria pública
LEI	Leitos nas especialidades básicas em hospitais credenciados pelo SUS, para cada grupo de 1.000 habitantes
LIN	Pontuação segundo a existência de linhas de transmissão de energia elétrica
MES	Matrículas em instituições de ensino superior
ONI	Linhas intermunicipais que servem ao município dividido pela raiz da população e multiplicado por 100
OPC	Valor médio das operações de crédito em agências bancárias do município, por habitante
PIB	Estimativa do PIB per capita (renda per capita) do município
PIC	Estimativa da taxa média de crescimento do PIB do município, entre 1990 e 1996
PRO	Conclusões em cursos oferecidos pelo SENAC e pelo SENAI, em relação à população de 15 anos ou mais
ROD	Existência ou proximidade de rodovias de pista dupla
SEB	Existência de Balcão SEBRAE
SEG	Policiais civis e militares para cada grupo de 10.000 habitantes
TEC	Matrícula em cursos técnicos de 2º grau, em relação à população de 15 anos ou mais
TEL	Terminais telefônicos para cada grupo de 1.000 habitantes
VEI	Veículos novos (a partir de 1996) licenciados para cada grupo de 1.000 habitantes

Fonte: Fundação CIDE, IQM - arquivo da Internet, http://www.cide.rj.gov.br, acessado em outubro de 2000.