Presidente da República
Luiz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Paulo Bernardo Silva

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor-Executivo
Sérgio da Costa Côrtes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Wasmália Socorro Barata Bivar

Diretoria de Geociências
Luiz Paulo Souto Fortes (em exercício)

Diretoria de Informática
Luiz Fernando Pinto Mariano

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Sérgio da Costa Côrtes (interino)

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 68 número 228 janeiro/junho 2007

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 68, n. 228, p. 1-1395, jan./jun. 2007

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Francisco Louzada-Neto
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Helem Ortega da Silva - Coordenação de Métodos e Qualidade - DPE/COMEQ/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2004.

Capa
Renato J. Aguiar – Coordenação de *Marketing*/CDDI/IBGE

Ilustração da Capa
Marcos Balster – Coordenação de *Marketing*/CDDI/IBGE

Revista brasileira de estatística/IBGE, - v.1, n.1 (jan./mar.1940), - Rio de Janeiro:IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística.

Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).

Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev.98) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor.....5

Artigos

Uma avaliação da eficiência do gasto público municipal no Brasil.....7
Fernando César Brito Santos
Francisco Cribari–Neto
Maria da Conceição Sampaio de Souza

Modelos fatoriais para retornos de ativos.....57
Luís Gustavo do Amaral Vinha
Chang Chiann

Estimação a partir de amostras periódicas não sobrepostas: a experiência francesa.....91
Andréa Diniz da Silva
Jean-Michel Durr

Comparação de desempenho de modelos da família ARCH.....111
Emerson Herbert Amorim
Luis Aparecido Milan

Política editorial.....137

Nota do Editor

Neste primeiro volume da RBEs do ano de 2007 temos quatro artigos interessantes. O primeiro artigo, de autoria de Fernando César Brito Santos, Francisco Cribari–Neto e Maria da Conceição Sampaio de Souza, avalia a eficiência do gasto público municipal no Brasil a partir de três técnicas diferentes. O segundo artigo, de autoria de Luís Gustavo do Amaral Vinha e Chang Chiann, apresenta modelos fatoriais para retornos de ativos. O terceiro artigo, de autoria de Andréa Diniz da Silva e Jean-Michel Durr, propõe um procedimento de estimação a partir de amostras periódicas não sobrepostas como uma alternativa ao modelo corrente de enumeração completa da população. O caso francês é enfocado. O quarto artigo, de autoria de Emerson Herbert Amorim e Luis Aparecido Milan, apresenta uma revisão de alguns modelos paramétricos da "família ARCH", mostra sua aplicabilidade na estimação da volatilidade de ativos financeiros e compara os seus desempenhos.

Aproveito a oportunidade para agradecer a colaboração de todos os Editores Associados e revisores do periódico, bem como à equipe do IBGE.

Uma excelente leitura.

Francisco Louzada-Neto
Editor Responsável

Uma avaliação da eficiência do gasto público municipal no Brasil

Fernando César Brito Santos[*]
Francisco Cribari-Neto[**]
Maria da Conceição Sampaio de Souza[***]

Resumo

Este artigo fornece uma visão geral de como os municípios brasileiros têm-se comportado com relação ao gerenciamento de recursos públicos. O objetivo central é mensurar o impacto que a idade do município exerce na eficiência dessas cidades, além de indicar alguns aspectos considerados relevantes para a otimização dos gastos municipais. As avaliações foram realizadas a partir de três técnicas: DEA (análise envoltória de dados), regressão clássica e regressão quantílica. A técnica DEA se mostrou eficaz para a análise aqui proposta, porém não foi capaz de identificar o efeito que fatores externos poderiam provocar na eficiência dos municípios do Brasil. Daí a utilização dos métodos de regressão supramencionados, sendo a regressão quantílica mais elucidativa por não se limitar às estimativas das medidas de tendência central. Os resultados revelaram que variáveis como idade dos municípios e recebimento de *royalties* sempre devem ser levadas em consideração em uma avaliação municipal. Em particular, os resultados revelam que o intenso processo de criação de novos municípios existente no Brasil desde a promulgação da Constituição de 1988 é detrimental para a eficiência do gasto público municipal.

Palavra-chave: DEA, regressão clássica, regressão quantílica, serviços públicos municipais.

[*] Endereços para correspondências: Deptº de Estatística e Informática - CCT, Univ. Católica de Pernambuco - Rua do Príncipe, 526, Boa Vista, Recife/PE, 50050-900 – tel.: (81) 2119-4175 - *e-mail*: fcesar@dei.unicap.br.

[**] Francisco Cribari-Neto – Deptº CCEN, Universidade Federal de Pernambuco - Cidade Universitária, Recife/PE, 50740-540 – tel.: (81) 2126-7425; fax: (81) 2126-8422 - *e-mail*: cribari@ufpe.br.

[***] Deptº de Economia, Universidade Nacional de Brasília - Campus Universitário Darcy Ribeiro, s/nº, Brasília/DF, 70910-900 – tel./fax: (61) 3307-2498; fax: (61) 3340-2311 - *e-mail*: mcss@unb.br.

1. Introdução

Uma avaliação estatística do gasto público municipal no Brasil foi realizada por Sampaio de Sousa, Cribari–Neto e Stošić (2005). Contudo, estes autores não levaram em consideração a idade do município como fator condicionante. Esta variável pode influenciar a eficiência dos municípios brasileiros, uma vez que municípios novos supostamente não têm, ainda, uma estrutura administrativa e gerencial adequada.

A princípio, conjectura-se que em uma cidade recém-fundada a falta de estrutura e de recursos pode levá-la a ser pouco eficiente, ou até mesmo ineficiente no que se refere à administração dos recursos municipais. O objetivo deste artigo é avaliar o impacto da criação de novos municípios sobre a eficiência do gasto público municipal no Brasil.

Na avaliação foram utilizadas três técnicas: DEA (análise envoltória de dados), regressão clássica e regressão quantílica. A técnica DEA permite computar fronteiras não-paramétricas de eficiência para os municípios brasileiros. Os eficientes são utilizados para a construção matemática da fronteira de eficiência e servem de referência para os municípios que estão afastados dessa fronteira, considerados ineficientes.

Devido à natureza determinística dos modelos não-paramétricos, as ineficiências encontradas podem não ser explicadas apenas pela incompetência administrativa ou pela inexistência de incentivos e planejamentos satisfatórios para assegurar um funcionamento adequado das municipalidades, mas também pela presença de observações influentes (*outliers*), erros de medida, omissão de variáveis, heterogeneidade dos dados e outras discrepâncias estatísticas (ver Sampaio de Sousa, Cribari–Neto e Stošić, 2005). Adicionalmente, as ineficiências podem ser conseqüência da existência de fatores externos, tais como condições naturais e climáticas, fatores políticos, características demográficas e socioeconômicas, dentre outras.

Este trabalho se concentra na obtenção da estimativa da eficiência técnica de cada município, utilizando sua dependência de fatores exógenos com a finalidade de avaliar a relação entre medidas obtidas pelo método DEA, através dos modelos CCR e BCC, e medidas expurgadas obtidas a partir dos modelos de regressão clássica e quantílica.

Este artigo é composto por seis seções, incluindo esta introdução. Na segunda seção é feita uma sucinta abordagem das técnicas utilizadas: DEA, esquema de reamostragem *jackstrap*, regressão quantílica e regressão clássica. Em seguida, são definidos os modelos utilizados na avaliação objeto deste artigo e, mais adiante, as seções 4 e 5 exibem uma descrição dos dados e os resultados, acompanhados de suas respectivas análises. Finalmente, a sexta seção apresenta um resumo das principais conclusões.

2. Técnicas Utilizadas

2. 1. DEA - análise envoltória de dados

Desenvolvida por Charnes, Cooper e Rhodes (1978), a técnica DEA consiste em determinar para cada unidade tomadora de decisão (DMU) a razão máxima entre a soma ponderada dos produtos e a soma ponderada dos insumos, condicionada à disponibilidade dos recursos e à quantidade gerada de produtos. Neste caso, os termos de ponderação são as incógnitas a serem determinadas pelo modelo. Dessa forma, a técnica DEA é identificada como um modelo de programação matemática fracional que pode ser transformado em um modelo de programação linear (Charnes, Cooper e Rhodes, 1978).

2.1.1. O modelo CCR

Sejam x_{ij} a quantidade disponível do insumo i usada na atividade j, y_{kj} a quantidade do produto k gerada na atividade j, v_i o peso dado ao insumo utilizado na atividade j e u_k o peso dado ao produto gerado na atividade j. O modelo CCR é dado por:

1. Modelo de Programação Fracional

$$\max_{u,v} Z_n = \frac{\sum_{k=1}^{K} u_k y_{kn}}{\sum_{i=1}^{I} v_i x_{in}}$$

sujeito a

$$\frac{\sum_{k=1}^{K} u_k y_{kj}}{\sum_{i=1}^{I} v_i x_{ij}} \leq 1, \quad j = 1,2,\ldots,n,\ldots,J,$$

$$u_k \geq 0 \quad (k = 1,2,\ldots,K), \quad v_i \geq 0 \quad (i = 1,2,\ldots,I).$$

2. Modelo de Programação Linear

$$\max_{u,v} W_n = \sum_{k=1}^{K} u_k y_{kn}$$

sujeito a

$$\sum_{k=1}^{K} u_k y_{kj} - \sum_{i=1}^{I} v_i x_{ij} \leq 0, \quad j = 1,2,\ldots,J,$$

$$\sum_{i=1}^{I} v_i x_{in} = 1,$$

$$u_k \geq 0 \quad (k = 1,2,\ldots,K), \quad v_i \geq 0 \quad (i = 1,2,\ldots,I).$$

De maneira mais direta, o modelo CCR de programação linear acima visa a determinar os pesos que maximizam o produto final gerado, W_n, para cada DMU tomada como referência, com os mesmos ou menos recursos. Isto significa que existem n problemas de programação linear para serem resolvidos. No que se refere às restrições, elas foram introduzidas no modelo para assegurar que nenhuma DMU pode estar além da "fronteira de eficiência" e para garantir que os pesos são positivos.

Convém salientar que para cada modelo de programação linear, chamado primal ou primitivo, existe um outro modelo, denominado dual, que é composto pelos mesmos coeficientes do primal, porém dispostos de maneira diferente. Neste caso, o dual do

modelo CCR de programação linear é dado por

$$\min_{\theta,\lambda} \theta$$

sujeito a

$$\theta x_{in} - \sum_{j=1}^{J} \lambda_j x_{ij} \geq 0, \qquad i = 1,2,\ldots,I,$$

$$\sum_{j=1}^{J} \lambda_j y_{kj} \geq y_{nk}, \quad k = 1,2\ldots,K,$$

$$\lambda_j \geq 0, \qquad j = 1,2,\ldots,J.$$

Aqui, verifica-se que o objetivo é determinar os pesos que minimizam o uso dos recursos disponíveis para cada DMU, com a finalidade de gerar a mesma quantidade de produtos ou mais. Quanto às restrições, pode-se notar que o vetor de intensidade, λ, representa retornos constantes de escala.

2.1.2. O modelo BCC

Em algumas situações a obtenção de medidas relativas de eficiência deve levar em consideração desajustes estruturais de longo prazo e, portanto, o uso de retornos constantes de escala não é adequado. Logo, nessas circunstâncias, o modelo CCR não deve ser utilizado. Banker, Charnes e Cooper (1984) fizeram uma pequena alteração no modelo CCR, introduzindo uma nova restrição àquele modelo, que permitisse a existência de retorno variável de escala. Neste caso, o conjunto factível de atividades é formado por todas as combinações convexas das atividades observadas (disponíveis). Portanto, tem-se retornos crescentes para níveis baixos de produção e retornos decrescentes para níveis altos.

Assim, foi criado o modelo BCC:

$$\min_{\theta,\lambda} \theta$$

sujeito a

$$\theta x_{in} - \sum_{j=1}^{J} \lambda_j x_{ij} \geq 0, \qquad i = 1,2,\ldots,I,$$

$$\sum_{j=1}^{J} \lambda_j y_{kj} \geq y_{nk}, \quad k = 1,2,\ldots,K,$$

$$\lambda_j \geq 0, \qquad j = 1,2,\ldots,J,$$

$$\sum_{j=1}^{J} \lambda_j = 1.$$

Devido ao fato de que a técnica DEA é um método não-estocástico, a "fronteira de eficiência" gerada por este método é passível de erros de medida e exposta ao questionamento das propriedades estatísticas de seus resultados. Uma forma de superar estas dificuldades seria a combinação da técnica DEA com outras metodologias, como, por exemplo, o esquema *jackstrap*, que combina outros dois esquemas de reamostragem (*jackknife* e *bootstrap*) para eliminar a influência de *outliers* (pontos influentes) e possíveis erros de medida contidos nos dados.

Segundo Sampaio de Sousa e Stošić (2005), o método DEA é extremamente atrativo, pois não exige conhecimento prévio da relação funcional existente entre as variáveis de entrada e saída e não impõe pesos estatísticos arbitrários às variáveis. Por outro lado, ainda segundo Sampaio de Sousa e Stošić (2005), como o método é baseado no conceito de fronteira de produção, um simples erro no conjunto de dados (erro de medida) ou uma unidade com excepcional desempenho (*outlier*) pode comprometer seriamente a análise, visto que os resultados para as unidades restantes ficam alterados na direção de valores de pouca eficiência, a distribuição de freqüência da eficiência torna-se altamente assimétrica e a escala de eficiência global muda para uma forma não-linear.

Vários autores, como Banker e Gifford (1988), Seaver e Triantis (1989) e Wilson (1993, 1995), desenvolveram técnicas considerando os efeitos de *outliers* e de erros de medida. Contudo, as aproximações por eles propostas dependem fortemente de inspeção manual, o que é virtualmente impossível quando o conjunto de dados é muito grande.

A definição de *outlier* derivada do conceito de alavancagem usado por Cribari-Neto e Zarkos (2004) aplicado à técnica DEA utiliza o impacto produzido no resultado das eficiências DEA para todas as outras DMUs, quando uma dada DMU é removida do conjunto de dados. Assim, a identificação das DMUs influentes é feita a partir das medidas de alavancagem, que podem ser calculadas da seguinte maneira:

(i) Aplica-se o método DEA ao conjunto original de dados, obtendo-se, assim, o conjunto de medidas de eficiência $\{\theta_j \mid j = 1,2,\ldots,J\}$.

(ii) Remove-se cada DMU do conjunto de dados, uma após a outra, e para cada DMU removida recalcula-se o conjunto de medidas de eficiência, agora denotado por $\{\theta^*_{jr} \mid j = 1,2,\ldots,J; j \neq r\}$, onde $r = 1,2,\ldots,J$ indica a DMU removida.

(iii) Calcula-se a medida de alavancagem para a j-ésima DMU como sendo o desvio padrão das medidas de eficiência antes e depois da remoção da DMU, ou seja:

$$\ell_r = \sqrt{\frac{\sum_{j=1; j\neq r}^{J} (\theta^*_{jr} - \theta_j)^2}{J-1}}. \qquad (1)$$

Como pode-se notar, a obtenção de ℓ_r exige a resolução de $J(J-1)$ modelos de programação linear, o que pode ser computacionalmente muito intensivo e até proibitivo para J muito grande. Surge, assim, a necessidade de uma técnica que diminua a intensidade dos cálculos.

2.1.3. O Esquema *Jackstrap*

Desenvolvido por Sampaio de Sousa e Stošić (2005), este esquema objetiva reduzir o impacto das DMUs influentes, usando o conceito de *outliers* dado por Cribari-Neto e Zarkos (2004). Isto pode ser realizado através de um processo em duas fases.

1. Primeira fase: Nesta fase é implementado um algoritmo em três passos para calcular as medidas de alavancagem, ou seja:

(i) Seleciona-se aleatoriamente um subconjunto de L DMUs, geralmente 10% de J, e calculam-se as medidas de alavancagem de acordo com (1), agora denotadas por $\tilde{\ell}_{j1}$, onde o número 1 no índice indica o primeiro subconjunto gerado;

(ii) Repete-se o passo (i) um número grande, B, de vezes, obtendo-se $\tilde{\ell}_{jb}$, onde $b = 1,2,\ldots,B$. Neste caso, encontram-se BL subconjuntos de medidas de alavancagem. Assim, cada DMU, em média, é selecionada $m_j = BL/J$ vezes; e

(iii) Calcula-se a alavancagem média para cada DMU através da expressão

$$\tilde{\ell}_j = \frac{\sum_{b=1}^{m_j} \tilde{\ell}_{jb}}{m_j}$$

e a alavancagem média global

$$\tilde{\ell} = \frac{\sum_{j=1}^{J} \tilde{\ell}_j}{J}.$$

2. Segunda fase: Aqui, é utilizado o esquema *bootstrap* para reduzir a probabilidade de um *outlier* ser selecionado, usando as medidas de alavancagem obtidas na primeira fase. Sampaio de Sousa e Stošić (2005) implementaram este esquema a partir das distribuições de probabilidade linear, inversa, exponencial e *da heaviside step function*.

   Por levar em consideração o tamanho da amostra, neste trabalho foi usada a distribuição *heaviside step function*, assim definida:

$$P(\tilde{\ell}_j) = \begin{cases} 1, & \text{se} \quad \tilde{\ell}_j < \tilde{\ell} \log J, \\ 0, & \text{se} \quad \tilde{\ell}_j \geq \tilde{\ell} \log J. \end{cases}$$

Assim, mesmo usando o esquema *jackstrap* para tornar o método DEA mais robusto, é necessário intenso cálculo computacional. Uma alternativa metodológica para a mensuração da eficiência seria o uso de modelos paramétricos, ou seja, o uso de modelos estatísticos de fronteira estocástica. Contudo, este tipo de modelo exige

informações prévias sobre sua forma funcional além de impor fortes condições para sua aplicação.

2. 2. Regressão quantílica

Esta classe de regressão foi introduzida por Koenker e Basset (1978) como uma forma de ampliar o método clássico de estimação em um modelo de regressão, que nos fornece a média condicional, para uma técnica de estimação de várias curvas, através dos quantis condicionais, que caracterizam toda a distribuição condicional da variável de interesse, dado um conjunto de regressores.

A solução de problemas de programação linear consiste em otimizar (maximizando ou minimizando) uma determinada função objetivo, condicionada a certas restrições. O modelo de regressão quantílica pode ser expresso na forma de programação linear, o que exige a definição de uma função objetivo.

Seja o seguinte modelo linear, expresso matricialmente:

$$Y = X\beta + \varepsilon,$$

onde Y é um vetor de variáveis aleatórias de tamanho $n \times 1$, X é uma matriz $n \times p$ de regressores de posto $p\,(p < n)$, β é um vetor $p \times 1$ de parâmetros desconhecidos e ε é um vetor $n \times 1$ de erros independentes e identicamente distribuídos. A função de distribuição de Y é dada por

$$P(Y \leq y_i) = P[(X\beta + \varepsilon) \leq y_i] = P(\varepsilon \leq y_i - x_i\beta) = F_\varepsilon(y_i - x_i\beta),$$

onde x_i representa a i-ésima linha da matriz X. Se β for um escalar e $x_i \equiv 1$, tem-se como caso especial o modelo de locação.

A função quantil condicional de Y dado X pode ser assim definida:

$$Q_\tau(Y \mid X) = X\beta + F_\varepsilon^{-1}(\tau) = X\beta(\tau), \qquad (2)$$

onde $\beta(\tau) = \beta + F_{\varepsilon}^{-1}(\tau)u$, sendo $u' = (1,0,\ldots,0) \in \Re^p$ e sendo F_{ε}^{-1} a função quantil dos erros.

De forma geral, o τ-ésimo quantil amostral de Y no modelo de locação é uma solução do problema

$$\min_b \left\{ \sum_{i:y_i \geq b} \tau |y_i - b| + \sum_{i:y_i < b} (1-\tau)|y_i - b| \right\}, \quad b \in \Re.$$

O análogo para o modelo linear do τ-ésimo quantil é definido de maneira semelhante. Assim, a solução para $\beta(\tau)$ em (2), $\hat{\beta}(\tau)$, denominado quantil de regressão, é

$$\min_{\beta} \; n^{-1} \left\{ \sum_{i:y_i \geq x_i\beta} \tau |y_i - x_i\beta| + \sum_{i:y_i < x_i\beta} (1-\tau)|y_i - x_i\beta| \right\},$$

que pode ser reescrito como

$$\min_{\beta} \; n^{-1} \sum_{i=1}^{n} \rho_{\tau}(y_i - x_i\beta),$$

onde $\rho_{\tau}(e_i)$, a função perda, é dada por

$$\rho_{\tau}(e_i) = (\tau - I(e_i < 0))e_i = \begin{cases} \tau e_i, & se \quad e_i \geq 0, \\ (\tau - 1)e_i, & se \quad e_i < 0, \end{cases}$$

sendo $e_i = y_i - x_i\beta$ e sendo $I(\cdot)$ a usual função indicadora. Essa função é conhecida como *check function*.

Dessa forma, as estimativas de $\beta(\tau)$ são obtidas minimizando a média da "função perda". Esta otimização pode ser atingida através das técnicas de programação linear, bastando para isto introduzir $2n$ variáveis artificiais $\{v_i^{(1)}, v_i^{(2)} : i = 1,2,\ldots,n\}$ para representar as partes positiva e negativa do vetor de resíduos.

Assim, o modelo de regressão quantílica tem a seguinte representação na forma de programação linear, denominada primal:

- Função objetivo:

$$\min_{\beta} \; n^{-1}\left\{ \sum_{i:y_i \geq x_i\beta} \tau \,|\, y_i - x_i\beta \,| + \sum_{i:y_i < x_i\beta} (1-\tau)\,|y_i - x_i\beta\,| \right\}.$$

- Restrições:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j(\tau) + e_i \;,$$

com $\beta_j(\tau)$ e e_i sem restrição de sinal.

2.3. Regressão clássica

2.3.1. O método dos mínimos quadrados ordinários

Seja o seguinte modelo linear, expresso em forma matricial:

$$Y = X\beta + e \,,$$

onde Y é um vetor de variáveis aleatórias de tamanho $n \times 1$, X é uma matriz $n \times p$ de regressores de posto $p\,(p < n)$, β é um vetor $p \times 1$ e e é um vetor $n \times 1$ de erros, sendo p a quantidade de parâmetros de regressão.

Supondo que os valores de X não são aleatórios, o cálculo da esperança de Y é

$$E(Y) = E(X\beta + e) = X\beta + E(e) \,.$$

Admitindo que a média dos erros é nula, vem

$$E(Y) = X\beta \,.$$

Dessa forma, o erro pode ser definido como sendo $e = Y - E(Y) = Y - X\beta$.

O método dos mínimos quadrados consiste em minimizar a soma dos quadrados dos erros, ou seja,

$$S = e'e = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta \,.$$

Derivando com relação a β e igualando a 0, encontram-se as estimativas dos parâmetros do modelo em questão, ou seja:

$$\hat{\beta} = (X'X)^{-1}X'Y .$$

Note que as variáveis que compõem a matriz X não devem apresentar colinearidade perfeita, pois isto tornaria a matriz $(X'X)$ singular.

Considerando que a variância dos erros é constante e que o erro de uma observação é não-correlacionado com o erro de outras observações, isto é, $\text{var}(e) = \sigma^2$, onde $0 < \sigma^2 < \infty$, e $\text{cov}(e_i, e_j) = 0 \; \forall \, i \neq j$, a matriz de covariância de Y é

$$\text{cov}(Y) = E\{[Y - E(Y)][Y - E(Y)]'\} = E(ee') = \sigma^2 I_n ,$$

onde I_n é a matriz identidade de ordem n.

2.3.2. Propriedades dos estimadores de mínimos quadrados ordinários

Os estimadores de mínimos quadrados ordinários são não-viesados:

$$E(\hat{\beta}) = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta .$$

A matriz de covariância de $\hat{\beta}$ é dada por

$$\begin{aligned} \text{cov}(\hat{\beta}) &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\} = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[(X'X)^{-1}X'ee'X(X'X)^{-1}] = (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}. \end{aligned}$$

Como a estimativa de mínimos quadrados ordinários de σ^2 é $\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-p)$, onde $\hat{e} = [I_n - X(X'X)^{-1}X']Y$ é o vetor de resíduos, a estimativa de $\text{cov}(\hat{\beta})$ é

$$\hat{\text{cov}}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}.$$

Uma observação importante é que $\hat{\beta}$ é função linear de Y. Logo, se Y tem distribuição normal, $\hat{\beta}$ também tem distribuição normal.

Os estimadores de mínimos quadrados ordinários são consistentes:

$$\begin{aligned} EQM(\hat{\beta}) &= Var(\hat{\beta}) + [Viés(\hat{\beta})]^2 = \\ &= \sigma^2 (X'X)^{-1}, \end{aligned}$$

onde EQM denota erro quadrático médio. Assim,

$$\lim_{n\to\infty} EQM(\hat{\beta}) = 0.$$

Logo,

$$\hat{\beta} \xrightarrow{p} \beta \text{ quando } n \to \infty,$$

onde $\xrightarrow{p}$ denota convergência em distribuição.

Assim, os estimadores de mínimos quadrados ordinários são lineares, não-viesados e consistentes. Como estes estimadores satisfazem ao teorema de Gauss-Markov, então eles se enquadram na classe dos estimadores lineares não-viesados de variância mínima. Contudo, convém salientar que esta propriedade só se verifica sob as suposições já citadas.

3. Modelos

Uma avaliação estatística do gasto público municipal no Brasil foi realizada por Sampaio de Sousa, Cribari–Neto e Stošić, em 2005. Contudo, esse trabalho não levou em consideração a idade do município, variável esta que pode influenciar o resultado da eficiência dos municípios brasileiros, uma vez que municípios novos, supostamente, não têm, ainda, uma estrutura administrativa adequada. Esta conjectura é abordada neste artigo, que faz a avaliação da eficiência técnica dos municípios brasileiros considerando a idade dos mesmos.

Através da técnica DEA, os municípios brasileiros são classificados como eficientes ou ineficientes no que se refere ao gasto público. Os eficientes são utilizados para a construção matemática da fronteira de eficiência e servem de referência para os municípios que estão afastados dessa fronteira, considerados ineficientes.

Devido à natureza determinística dos modelos não-paramétricos, as ineficiências encontradas podem não ser explicadas apenas pela incompetência administrativa ou pela inexistência de incentivos e de planejamentos satisfatórios para assegurar um funcionamento adequado das municipalidades, mas, também, pela presença de observações influentes (*outliers*), erros de medida, omissão de variáveis, hetero-

geneidade dos dados e outras discrepâncias estatísticas (ver Sampaio de Sousa, Cribari-Neto e Stošić, 2005).

Adicionalmente, as ineficiências podem ser conseqüência da existência de fatores externos sobre os quais os municípios não têm controle, tais como: condições naturais e climáticas, fatores políticos, características demográficas e socioeconômicas, dentre outras. Neste contexto, a regressão clássica e a regressão quantílica podem ser utilizadas para predizer a eficiência dos municípios brasileiros em função de fatores condicionantes.

Dessa forma, este trabalho também se concentra na obtenção da estimativa da eficiência técnica de cada município, utilizando sua dependência de fatores exógenos com a finalidade de avaliar a relação entre as medidas obtidas pelo método DEA, através dos modelos CCR e BCC, e as medidas preditas pelos fatores condicionantes obtidas a partir dos modelos de regressão clássica e quantílica. Por conseguinte, de posse das eficiências preditas pelos modelos de regressão exclusivamente em função de fatores exógenos, foram expurgados os efeitos destes fatores das eficiências DEA, obtendo-se, assim, medidas mais "puras" de eficiência.

3.1. Os modelos DEA

Visando ao tratamento de *outliers* e erros de medida, Sampaio de Sousa, Cribari-Neto e Stošić (2005) utilizaram o esquema *jackstrap* para mensurar as medidas de eficiência a partir da técnica DEA, tanto no modelo CCR quanto no BCC.

As variáveis que compõem estes modelos são identificadas em dois grupos:

- Insumos, entradas ou *inputs:* são os recursos utilizados pelos municípios. Aqui, são utilizados:
  - –gasto atual (X_1) - esta variável está relacionada com o custo total agregado;
  - –número de professores (X_2) - usado para mostrar a eficiência dos serviços em educação;
  - –número de hospitais e postos de saúde (X_3) - indicador de gastos com serviços de saúde; e

- proporção de mortalidade infantil (X_4) - esta variável foi colocada no modelo como insumo pelo fato de que ela é indicadora de eficiência dos serviços de saúde.

- Produtos finais/acabados, saídas ou *outputs*: são os serviços oferecidos pelos municípios. As variáveis que compõem este grupo foram cuidadosamente selecionadas de um conjunto de serviços disponibilizados pelos municípios, levando em consideração a consonância com o grupo de insumos. Estas variáveis são:
  - população total residente (Y_1) - esta variável visa à avaliação dos serviços administrativos;
  - população alfabetizada (Y_2), número de estudantes matriculados nos ensinos fundamental e médio (Y_3), freqüência dos estudantes por escola (Y_4), número de estudantes aprovados por escola (Y_5), número de estudantes na série correta por escola (Y_6) - estas cinco variáveis estão relacionadas à eficiência dos serviços educacionais; e
  - número de domicílios com água potável (Y_7), número de domicílios com acesso ao sistema de esgoto (Y_8) e número de domicílios com acesso à coleta de lixo (Y_9) - estas variáveis avaliam as condições de saúde e moradia.

No cômputo da fronteira de eficiência, o modelo utilizado foi o de minimização de custos (*input oriented*). Essa escolha foi motivada pelo fato de que, no Brasil, existe uma crescente pressão para que os gestores municipais reduzam seus custos e imponham maior racionalidade ao uso dos recursos públicos. Essas pressões culminaram, em 2001, com a criação da Lei de Responsabilidade Fiscal - LRF, que fixa regras estritas para o controle de gastos municipais. Nesse contexto, a busca por menores custos, para um dado nível de serviço público, é melhor tratada mediante o uso do modelo insumo-orientado.

3.2. Os Modelos Lineares de Regressão Clássica

3.2.1. Definição dos Modelos

Devido ao fato de que a variável dependente assume valores no intervalo [0, 1], ou seja, $(0 < \theta \leq 1)$, o que caracteriza uma restrição ou uma censura à variável, o modelo de regressão mais adequado seria o *tobit*, também conhecido como modelo de regressão censurada. Desenvolvido por James Tobin (1958), *tobit* é um modelo de regressão onde fatores ligados à mensuração dos dados impedem a observação de valores da variável resposta ao longo de toda a sua extensão possível. Isto significa que a variável resposta assume apenas determinados valores, ficando os demais censurados. Esta censura introduz uma correlação entre o termo de erro e as covariáveis de um modelo de regressão linear clássica, violando uma das condições necessárias à aplicação do método dos mínimos quadrados. Os parâmetros do modelo *tobit* geralmente são estimados pelo método da máxima verossimilhança. Contudo, para sua aplicação é necessário assumir normalidade e homoscedasticidade, pois, caso contrário, as estimativas dos parâmetros seriam inconsistentes.

O uso do método dos mínimos quadrados ordinários em modelos lineares de regressão onde a variável dependente é censurada também gera estimativas inconsistentes. Segundo Greene (1981), este problema se agrava na medida em que a proporção de observações censuradas na amostra aumenta. Então, pode-se notar que o problema da inconsistência não traz grandes conseqüências quando o número de observações censuradas é pequeno.

Os estimadores de mínimos quadrados ordinários são não-viesados, consistentes, e assintoticamente normais mesmo na presença de heteroscedasticidade de forma desconhecida, o que não ocorre com os estimadores de máxima verossimilhança utilizados no modelo *tobit*. Este é mais um argumento em favor do método dos mínimos quadrados, uma vez que se pode usar estimadores consistentes da matriz de covariância de $\hat{\beta}$ (HCCMEs). Além do mais, tem sido mostrado na literatura que fazendo uma transformação logarítmica na variável dependente, os estimadores de mínimos quadrados ordinários continuam não-viesados e passam a ser consistentes quando a

variável dependente assume apenas valores estritamente positivos.

Diante do exposto, a escolha do modelo para realizar as estimativas da eficiência técnica dos municípios do Brasil recai no modelo de regressão linear clássica através do método de mínimos quadrados, com especificação *semilog*, ou seja,

$$\ln\theta = X\beta + e.$$

Neste modelo, o coeficiente de inclinação mede a variação proporcional ou relativa ocorrida em θ para cada variação absoluta verificada em X. Assim, multiplicando-se a variação relativa em θ por 100, obtém-se a variação percentual ou a taxa de crescimento ocorrida em θ para uma variação absoluta em X.

Outro modelo aqui estudado é definido em função da variável distância entre os municípios. Anselin (1988), com o objetivo de levar em consideração a dependência espacial de θ, ou seja, visando eliminar os efeitos da correlação espacial, propôs o modelo

$$(I - \rho W)\theta = X\beta + e \Rightarrow \theta = \rho W\theta + X\beta + e\text{, com } e = \lambda We + u\text{.}$$

Como o modelo especificado é o *semilog*, nós usamos como resposta $\ln\theta$ e não θ. Assim,

$$\ln\theta = \rho W\theta + X\beta + e$$

onde I é a matriz identidade de ordem n e W é uma matriz $n \times n$ que controla a existência dos efeitos de vizinhança entre as cidades; aqui ρ mede a correlação espacial, ou seja, se for diferente de zero, indica que a eficiência estimada para um determinado município é diretamente afetada pela eficiência de seus vizinhos. O parâmetro λ captura a correlação espacial entre os erros e u é um novo termo para o erro.

Igualmente a Sampaio de Sousa, Cribari-Neto e Stošić (2005), aqui W tem duas formas: na primeira, denotada por W_1, o elemento (i, j) de W_1 é igual a 1 se a distância

entre os municípios i e j não excede a 50 quilômetros e é igual a zero caso contrário; na segunda, denotada por W_2, o elemento (i,j) de W_2 é igual ao quociente entre a distância entre os municípios i e j e a distância máxima encontrada, o que resulta em uma medida entre zero e 1. Este critério para a definição de W permite que a avaliação seja realizada através de uma medida binária (*dummy*) e de uma medida relativa e, conseqüentemente, a partir de dois modelos distintos:

$$\ln\theta = \rho W_1\theta + X\beta + e \quad \text{e} \quad \ln\theta = \rho W_2\theta + X\beta + e\,.$$

3.3. Os modelos de regressão quantílica

Considerando a heterogeneidade dos dados, é possível que as estimativas condicionadas pela média, obtidas a partir do modelo linear de regressão tradicional através dos estimadores de mínimos quadrados sofram influência dessa variabilidade. A regressão quantílica oferece mecanismos para que as estimativas sejam condicionadas pela mediana, que não é afetada por valores extremos ou, ainda, se for do interesse, por outros quantis.

Os modelos de regressão quantílica, aqui utilizados, seguem o mesmo padrão adotado nos modelos de regressão linear clássica descritos acima.

3.4. Variáveis regressoras para composição dos modelos de regressão

A variável resposta é a eficiência técnica obtida pelo método DEA e as covariáveis são identificadas e classificadas de acordo com suas características, da seguinte forma:

- socioeconômicas: despesas com pessoal (E_1), despesa com pessoal ativo (E_2), percentual de domicílios cujo chefe ganha até um salário mínimo (E_3), rendimento médio (E_4), rendimento mediano (E_5), participação no Projeto Alvorada (E_{22}) e rendimento de *royalties* de água (ROY);
- políticas: partido político dos prefeitos - PFL (E_6), PMDB (E_7), PSDB (E_8), PT (E_9), PPS (E_{10}), PPB (E_{11}), PTB (E_{12}), PDT (E_{13}), outros partidos (E_{14});

- administrativas: prédios que pagaram IPTU em 1998 (E_{15}), índice de atualização do cadastro predial (E_{16}), participação em consórcios intermunicipais (E_{17}), grau de informatização (E_{18}) e poder de decisão dos conselhos municipais (E_{19});
- indicadoras de economia de escala: densidade demográfica (E_{20}) e taxa de urbanização (E_{21});
- indicadoras de localização: Polígono das Secas (E_{23}), região metropolitana (RM_2), capitais dos estados (CAP) e distância entre os municípios (W_1 *e* W_2); e
- outras variáveis: idade do município (E_{25}), municípios turísticos (MT), população total (POP), *royalties* I (R_1), *royalties* II (R_2), receitas correntes (RC), transferências correntes (TC) e receita líquida (RL).

Dentre as variáveis citadas acima algumas são do tipo *dummy*, ou seja, assumem o valor 1 se há a característica desejada e o valor 0 caso contrário. Especificamente para as variáveis distância entre os municípios e idade dos municípios, as características desejadas são menores ou igual a 50 quilômetros e oito anos, respectivamente. Esta escolha prende-se ao fato de que um município muito novo, no máximo com duas administrações, tende a ser ineficiente, devido à falta de estrutura, e que a eficiência de um determinado município pode influir na eficiência de seus vizinhos próximos.

No que se refere às variáveis *royalties* I (R_1) e *royalties* II (R_2), elas são do tipo *dummy*. R_1 assume o valor 1 se o município recebia qualquer quantia a título de *royalties* e o valor 0 caso contrário. Já R_2 é igual a 1 se o município recebia mais de 10% da sua receita tributária a título de *royalties* e é igual a 0 caso contrário.

4. Os dados - uma breve descrição

A base dos dados utilizada é composta por 4 755 municípios, selecionados dentre 5 264. Alguns municípios deixaram de ser incluídos nos dados porque não apresentavam as informações completas.

Todas as informações são referentes ao ano de 2 000, exceto com relação à variável "prédios que pagaram IPTU", cujas informações são de 1998. Estes dados foram obtidos junto aos órgãos: Secretaria do Tesouro Nacional, Instituto Brasileiro de Geografia e Estatística, Ministério da Educação e Cultura, Instituto de Pesquisas Econômicas e Aplicadas, e através do Programa das Nações Unidas para o Desenvolvimento.

A Tabela 1 exibe medidas estatísticas que fornecem uma visão global a respeito do comportamento das variáveis quantitativas utilizadas como regressores nos modelos de regressão linear clássica e quantílica. Como era de se esperar, estas variáveis se mostram heterogêneas, principalmente a receita líquida (RL), com coeficiente de variação 1 606,32%. Apenas duas, índice de atualização do cadastro predial (E_{16}) e taxa de urbanização (E_{21}), com coeficientes de variação 13,35% e 39,76%, respectivamente, apresentam certa homogeneidade. Algumas dessas medidas revelam situações interessantes e até curiosas, como listado a seguir:

- A cidade de Salto Veloso (SC) foi a que menos gastou com pessoal, R$ 915,78, enquanto São Paulo (SP) foi a que apresentou maior gasto, R$ 1 800 263 284,00. No que tange à despesa com pessoal ativo, a situação é semelhante: R$ 809,19 gastos por Salto Veloso e R$ 1 776 983 284,00 por São Paulo;
- O Município de Cantanhede (MA) foi o que apresentou o menor rendimento médio, cerca de R$ 114,85, e Santana de Parnaíba (SP), o maior, R$ 2 583,57;

Tabela 1 - Medidas descritivas de covariáveis que compõem os modelos de regressão linear clássica e quantílica

Variável	Medida			
	Mínimo	1º Quartil	Média	Mediana
E_1	915,78	1 054 762,49	5 982 257,94	1 796 975,85
E_2	809,19	880 172,53	5 151 351,47	1 477 156,66
E_3	0,00	3,20	11,96	8,38
E_4	114,85	260,47	435,90	400,27
E_5	70,00	151,00	240,25	200,00
E_{15}	0,00	26,50	1 856,16	180,00
E_{16}	0,00	0,89	0,91	0,97
E_{20}	0,10	12,00	89,94	25,00
E_{21}	0,00	40,04	58,36	58,43
POP	795,00	5 449,00	31 550,86	10 977,00
RC	0,00	1 951 458,00	14 599 047,37	3 370 762,56
TC	0,00	1 868 988,31	9 717 298,06	3 059 835,00
RL	0,00	31 453,20	4 881 749,32	187 842,00

Variável	Medida			
	3º Quartil	Máximo	Desvio padrão	Coeficiente de Variação (%)
E_1	3 529 655,01	1 800 263 284,00	39 239 479,25	655,93
E_2	2 968 204,91	1 776 983 284,00	35 950 452,13	697,88
E_3	17,63	62,93	11,02	92,12
E_4	561,01	2 583,57	214,91	49,30
E_5	300,00	1 000,00	109,03	45,38
E_{15}	754,50	917 523,00	18 513,67	997,42
E_{16}	0,97	1,00	0,12	13,35
E_{20}	50,10	11 608,80	459,00	510,36
E_{21}	77,21	100,00	23,21	39,76
POP	22 145,00	10 434 252,00	196 876,14	624,00
RC	7 215 022,70	7 753 358 748,00	141 570 626,36	969,72
TC	6 370 433,26	3 259 088 539,00	64 334 223,03	662,06
RL	765 562,83	4 494 270 209,00	78 416 908,62	1 606,3

- Em Santa Tereza (RS) não havia domicílios onde o chefe ganhava um salário mínimo ou menos. Na contramão estavam Barro Alto (GO), Nina Rodrigues (MA) e Iracema (CE), onde em cerca de 63% de seus domicílios o chefe da família ganhava até um salário mínimo;
- Para 705 municípios nenhum prédio pagou IPTU em 1998 e 648 não tinham esta informação disponível. No entanto, o índice médio de atualização do cadastro predial era de 0,909;
- No que se refere à população, a cidade de Borá (SP) tinha apenas 795 habitantes, quando a média era de 31 551 habitantes por cidade e metade dos municípios tinha menos de 11 000 habitantes; e
- 944 municípios não tinham receitas correntes e nem transferências correntes e 945 não tinham receita líquida.

Quanto às variáveis do tipo *dummy*, a Tabela 2 mostra o número de municípios com a característica desejada. Os dados são referentes ao ano de 2000. Por exemplo, 891 municípios tinham prefeitos filiados ao PFL, 1 089 ao PMDB e 167 ao PT; 1 876 cidades participavam de consórcios intermunicipais; 3 124 estavam informatizadas e em 2 075 os conselhos municipais tinham poder de decisão. Além disso, 2 001 municípios participavam do Projeto Alvorada, 1 212 pertenciam à região do Polígono das Secas e 807 tinham, no máximo, oito anos de existência.

Tabela 2 - Número de municípios com a característica requerida pela variável *dummy*

Variável *dummy*	Número de municípios
E_6	891
E_7	1 089
E_8	824
E_9	167
E_{10}	140
E_{11}	*536*
E_{12}	*336*
E_{13}	*250*
E_{14}	*522*
E_{17}	*1 876*
E_{18}	*3 124*
E_{19}	**2 075**
E_{22}	**2 001**
E_{23}	**1 212**
RM_2	**325**
E_{25}	**807**
ROY	**390**
MT	**350**
CAP	**26**
R_1	**697**
R_2	**474**

5. Resultados

Os resultados apresentados, a seguir, foram obtidos usando o *software* livre R, que é composto por pacotes desenvolvidos para a análise, manipulação e exibição gráfica de dados. Na verdade, o R é uma implementação gratuita da linguagem de programação de alto nível S. Uma de suas vantagens é a flexibilidade em permitir a criação de novas funções e a modificação de muitas funções internas (ver **http//www.r-project.org**).

5.1. Definição dos modelos ajustados

5.1.1. Modelos de regressão linear clássica

1. A Partir do DEA-CCR

Em virtude dos dados apresentarem heteroscedasticidade, conforme Tabela 3, que exibe as estatísticas dos testes de Koenker e Breusch-Pagan, os testes de significância foram realizados a partir de estimadores consistentes da matriz de covariância de $\hat{\beta}$ (HCCMEs), especificamente através dos estimadores HC3 e HC4 (ver Davidson & MacKinnon (1993) e Cribari-Neto (2004)).

Tabela 3 - Testes de heteroscedasticidade de Koenker e Breusch-Pagan para os três modelos de regressão linear clássica, onde a variável resposta é a eficiência técnica gerada pelo método DEA-CCR

Modelo	Teste					
	Koenker			Breusch-Pagan		
	LM_K	gl	p -valor	LM_{PB}	gl	p -valor
Sem correlação espacial	152,91	24	2×10^{-16}	210,01	24	2×10^{-16}
Com correlação espacial (W_1)	158,82	24	2×10^{-16}	217,64	24	2×10^{-16}
Com correlação espacial (W_2)	153,84	24	2×10^{-16}	209,27	24	2×10^{-16}

Após uma análise das variáveis disponíveis, ficou evidente que, devido à natureza das mesmas, algumas delas apresentavam elevadas correlações entre si e, conseqüentemente, foram excluídas dos três modelos de regressão linear clássica: o que mensura a eficiência técnica dos municípios sem levar em conta a correlação espacial e aqueles que consideram tal correlação. O Quadro 1 mostra as variáveis que foram excluídas, o motivo da exclusão e os modelos dos quais elas foram retiradas.

Uma constatação importante é que o modelo que não leva em conta a correlação espacial, através da variável distância entre os municípios, apresentou a covariável região metropolitana RM_2 como não-significativa no nível de 5%, enquanto os modelos que foram definidos com a correlação espacial apresentaram as covariáveis W_1EFIC e W_2EFIC como significativas no nível de 5%. Como os modelos com e sem correlação espacial diferem exatamente quanto ao uso dos regressores RM_2, W_1EFIC e W_2EFIC e

como RM_2 é não-significativo, então o modelo sem correlação espacial foi descartado, permanecendo apenas os dois modelos que envolvem as duas matrizes que definem as distâncias entre os municípios, o que é composto por W_1, com dezesseis covariáveis, e o que é formado por W_2, que tem 14 regressores.

Quadro 1 - Variáveis excluídas dos modelos de regressão linear clássica, onde a variável resposta é a eficiência técnica gerada pelo DEA-CCR.

Variável excluída	Motivo da exclusão	Modelo da variável
E_1	Não-significativa	Os três
E_2	Correlacionada com E_1	Os três
E_3	Não-significativa	Sem correlação espacial e com correlação espacial (W_2)
E_5	Correlacionada com E_4	Os três
E_6	Não-significativa	Os três
E_8	Não-significativa	Os três
E_9	Não-significativa	Os três
E_{10}	Não-significativa	Os três
E_{11}	Não-significativa	Os três
E_{12}	Não-significativa	Os três
E_{14}	Incompleta	Os três
E_{15}	Período diferente de 2000	Os três
CAP	Não-significativa	Com correlação espacial (W_2)
MT	Não-significativa	Os três
POP	$FIV > 5$	Os três
R_1	Correlacionada com R_2	Os três
RC	$FIV > 5$	Os três
RL	Correlacionada com E_1	Os três
RM_2	Não-significativa	Sem correlação espacial
ROY	Correlacionada com R_2	Os três
TC	$FIV > 5$	Os três

Assim, o modelo de regressão linear clássica, que tem como variável resposta a eficiência gerada pelo método DEA-CCR selecionado, é aquele composto pelas 16 covariáveis cuja descrição é exibida na Tabela 4. Esta decisão foi tomada com base no

fato de que quanto mais informações forem utilizadas para a composição de um modelo, mais confiáveis tenderão a ser os resultados obtidos. Os resultados revelam que:

- A vinculação partidária dos prefeitos dos municípios é significativa apenas no caso das cidades cujo administrador é filiado ao PMDB (E_7) e ao PDT (E_{13}). As cidades gerenciadas pelo PMDB têm sua eficiência técnica média diminuída em torno de 3,455%, enquanto as administradas pelo PDT a diminuição da eficiência média é ainda maior: 4,676%, considerando, em cada caso, as demais covariáveis constantes;
- Igualmente aos partidos políticos, os municípios turísticos (MT) não conseguem fazer com que sua vocação para o turismo os tornem mais eficientes, visto que esta variável não é significativa no nível de 5%. Certamente esta variável deve receber mais atenção dos administradores municipais, uma vez que, sendo bem trabalhada, pode ser uma fonte geradora de importantes recursos, que bem aplicados poderiam tornar o turismo em uma variável significativamente positiva na prestação de serviços à população;

Tabela 4 - Descrição do modelo de regressão linear clássica selecionado, onde a variável resposta é a eficiência técnica gerada pelo método DEA-CCR

Covariável	Estimativa do coeficiente	Estatística teste		
		t	*quasi-t*	
			HC3	HC4
Intercepto	$-9{,}967\times10^{-1}$	−22,593	−22,583	−22,551
$W_1 EFIC$	$4{,}579\times10^{-3}$	5,724	5,792	5,787
E_3	$1{,}396\times10^{-3}$	2,054	1,972	1,972
E_4	$-8{,}186\times10^{-5}$	−2,385	−2,172	−2,149
E_7	$-3{,}455\times10^{-2}$	−3,211	−3,230	−3,234
E_{13}	$-4{,}676\times10^{-2}$	−2,308	−2,218	−2,218
$E_{16}^{(1)}$	$-7{,}190\times10^{-2}$	−1,912	−2,047	−2,046
E_{17}	$-5{,}659\times10^{-2}$	−5,825	−5,727	−5,733
E_{18}	$8{,}493\times10^{-2}$	7,864	7,639	7,647
E_{19}	$3{,}745\times10^{-2}$	4,121	4,136	4,141
E_{20}	$4{,}827\times10^{-5}$	4,326	3,721	3,313
E_{21}	$5{,}411\times10^{-3}$	21,941	21,164	21,160
E_{22}	$5{,}062\times10^{-2}$	3,390	3,145	3,146
E_{23}	$-6{,}788\times10^{-2}$	−4,838	−4,698	−4,702
E_{25}	$-8{,}823\times10^{-2}$	−6,844	−5,897	−5,904
CAP	$1{,}867\times10^{-1}$	2,830	2,927	2,762
R_2	$-1{,}526\times10^{-1}$	−9,868	−8,704	−8,706

(1) **Não-significativa no nível de 5% pelo teste *t*, mas significativa no mesmo nível normal pelos testes *quasi-t*.**

- Como era esperado, os municípios localizados no Polígono das Secas, E_{23}, tendem a ser mais ineficientes do que os demais, uma vez que, estando sujeitos a condições climáticas adversas, têm dificuldades de direcionar recursos para suprir as demandas da comunidade;
- A variável que mais contribui para o aumento da eficiência média dos municípios é capitais dos estados (CAP),que, mantendo os demais regressores constantes, aumenta a eficiência média em 18,67%, seguida do grau de informatização do município (E_{18}), que proporciona crescimento médio de 8,493%. Nos dois casos se verifica o esperado. Primeiro, as capitais geram mais recursos, devido ao fato

de serem mais desenvolvidas em determinados segmentos, como o industrial e o tecnológico. Segundo, a informatização do município pode diminuir custos e agilizar processos, o que aumenta a eficiência;

- A conjectura central deste artigo se confirmou, uma vez que a idade dos municípios, E_{25}, é a segunda variável significativa que mais influi negativamente. Quando as demais variáveis independentes são mantidas constantes, a eficiência média das cidades que têm menos de oito anos (cerca de duas gestões) é diminuída em aproximadamente 8,823%. Isto reforça a idéia de que a falta de estrutura administrativa e de experiência em gestão pública causa perdas aos municípios, tornando-os ineficientes. Assim, esta variável, que é o foco principal deste trabalho, indica que a criação de novos municípios, como vem acontecendo no Brasil, pode trazer mais problemas do que soluções, visto que a tendência é que estas cidades não tenham bom desempenho no tocante à administração dos gastos públicos;
- Os municípios que recebem *royalties* (R_2) não conseguem canalizar tal receita em prol da melhoria dos serviços públicos; muito pelo contrário, esta é a variável que mais influi na diminuição da eficiência dessas cidades, apresentando decréscimo médio de 15,26%. Isto sugere que esta receita adicional gera gastos de maior vulto devidos ou indevidos e, por conseguinte, administrações ineficientes; e
- Dentre as covariáveis significativas, as que menos exercem influência na eficiência média dos municípios são rendimento médio (E_4) e densidade demográfica (E_{20}), que caminham em direções opostas, ou seja, para cada real que aumenta, em média, a renda do trabalhador, a eficiência fica reduzida em menos de 0,01%, enquanto cada pessoa adicional por metro quadrado incrementa a eficiência média em menos de 0,05%. Estes resultados sugerem que as municipalidades mais empobrecidas tendem a administrar melhor seus escassos recursos. Esta suposição torna-se mais forte em virtude do percentual de domicílios cujo chefe ganha até um salário mínimo (E_3) ser positivamente significativo.

2. A Partir do DEA-BCC

Nesta seção, a análise realizada é semelhante àquela dos modelos de regressão linear clássica, onde a variável resposta é a eficiência técnica gerada pelo método DEA-CCR. Em síntese, a Tabela 5 mostra que aqui os dados também apresentaram heteroscedasticidade; o Quadro 2 exibe as variáveis excluídas dos modelos e a Tabela 6 apresenta o modelo selecionado.

O modelo de regressão linear clássica gerado pelo DEA-BCC confirma a maioria dos resultados obtidos pelo modelo de regressão linear clássica, onde a variável resposta é a eficiência gerada pelo DEA-CCR. Contudo, são constatadas as seguintes divergências:

Tabela 5 - Testes de heteroscedasticidade de Koenker e Breusch-Pagan para os três modelos de regressão linear clássica, onde a variável resposta é a eficiência técnica gerada pelo método DEA-BCC

Modelo	Teste					
	Koenker			Breusch-Pagan		
	LM_K	gl	p -valor	LM_{PB}	gl	p -valor
Sem correlação espacial	52,72	24	6×10^{-4}	68,68	24	3×10^{-6}
Com correlação espacial (W_1)	50,60	24	1×10^{-3}	66,68	24	7×10^{-6}
Com correlação espacial (W_2)	51,77	24	8×10^{-4}	67,61	24	5×10^{-6}

Quadro 2 - Variáveis excluídas dos modelos de regressão linear clássica, onde a variável resposta é a eficiência técnica gerada pelo DEA-BCC

Variável excluída	Motivo da exclusão	Modelo da variável
E_1	Não-significativa	Os três
E_2	Correlacionada com E_1	Os três
E_4	Não-significativa	Os três
E_5	Correlacionada com E_4	Os três
E_6	Não-significativa	Os três
E_8	Não-significativa	Os três
E_9	Não-significativa	Os três
E_{10}	Não-significativa	Os três
E_{12}	Não-significativa	Os três
E_{13}	Não-significativa	Os três
E_{14}	Incompleta	Os três
E_{15}	Período diferente de 2000	Os três
E_{16}	Não-significativa	Sem correlação espacial
E_{22}	Não-significativa	Os três
CAP	Não-significativa	Os três
MT	Não-significativa	Os três
POP	$FIV > 5$	Os três
R_1	Correlacionada com R_2	Os três
RC	$FIV > 5$	Os três
RL	Correlacionada com E_1	Os três
RM_2	Não-significativa	Sem correlação espacial
ROY	Correlacionada com R_2	Os três
TC	$FIV > 5$	Os três
W_2EFIC	Não-significativa	Com correlação espacial (W_2)

- As variáveis E_4, E_{13}, E_{20}, E_{22} e CAP passaram de significativas, no nível de 5%, no modelo gerado pelo DEA-CCR, para não-significativas no modelo relativo ao DEA-BCC, enquanto o regressor E_{11} deixou de ser não-significativo e passou a ser significativo, no nível de 5%, no modelo gerado pelo DEA-BCC; e
- O regressor índice de atualização do cadastro predial (E_{16}), significativo nos dois modelos, surpreendentemente aponta efeitos sobre eficiências em direções opostas, ou seja, no modelo gerado pelo DEA-CCR a eficiência diminui em

7,190%, enquanto no modelo obtido pelo DEA-BCC a eficiência aumenta em 9,771%.

Tabela 6 - Descrição do modelo de regressão linear clássica selecionado, onde a variável resposta é a eficiência técnica gerada pelo método DEA-BCC

Covariável	Estimativa do coeficiente	Estatística teste		
		t	*quasi-t*	
			HC3	HC4
Intercepto	$-9{,}489\times10^{-1}$	-22,128	-21,137	-21,126
W_1EFIC	$6{,}380\times10^{-3}$	7,942	7,788	7,796
E_3	$3{,}377\times10^{-3}$	5,575	5,813	5,816
E_7	$-3{,}518\times10^{-2}$	-3,130	-3,027	-3,030
E_{11}	$-6{,}286\times10^{-2}$	-4,203	-4,036	-4,040
E_{16}	$9{,}771\times10^{-2}$	2,516	2,369	2,366
E_{17}	$-3{,}666\times10^{-2}$	-3,676	-3,621	-3,625
E_{18}	$2{,}499\times10^{-2}$	2,334	2,295	2,298
E_{19}	$2{,}814\times10^{-2}$	3,004	3,017	3,020
E_{21}	$1{,}719\times10^{-3}$	6,924	6,877	6,884
E_{23}	$-5{,}930\times10^{-2}$	-4,497	-4,680	-4,683
E_{25}	$-4{,}042\times10^{-2}$	-3,064	-2,910	-2,913
R_2	$-8{,}032\times10^{-2}$	-5,099	-4,770	-4,772

Adicionalmente às análises realizadas até então, os resultados obtidos a partir dos dois modelos indicam que não há nenhuma relação direta entre rendas mais altas e melhoria da prestação de bons serviços públicos à sociedade; que o poder de decisão dos conselhos municipais inibe a ação da corrupção e o abuso no uso dos recursos públicos; e que a participação dos municípios em consórcios, ao contrário do esperado, diminui a eficiência da municipalidade envolvida.

Além disso, é importante destacar que a variável idade do município, E_{25}, mantém-se significativa, no nível de 5%, e continua influenciando negativamente a eficiência dos municípios com menos de oito anos de fundação. Entretanto, no modelo obtido a partir do DEA-BCC, esta influência é menor: $-4{,}042\%$.

5.1.2. Modelos de regressão quantílica

A escolha dos regressores para a composição do modelo de regressão quantílica foi baseada na análise que definiu os modelos de regressão linear clássica, ou seja, os regressores que foram considerados estatisticamente significativos no nível de 5% nos modelos de regressão clássica foram os selecionados para formar os dois modelos de regressão quantílica: aquele onde a variável dependente é a eficiência técnica obtida pelo método DEA-CCR, e o outro onde esta mesma variável é gerada pelo DEA-BCC.

Cada modelo foi estimado levando-se em consideração os quantis 0.10 (décimo percentil), 0,25 (primeiro quartil), 0,50 (mediana), 0,75 (terceiro quartil) e 0,90 (nonagésimo percentil).

Uma vez que há heteroscedasticidade, foi usada a estimativa *Huber sandwich* da matriz de covariância do estimador dos parâmetros de regressão.

1. A Partir do DEA-CCR

A Tabela 7 exibe as estimativas dos coeficientes do modelo de regressão quantílica ajustado, para os cinco quantis adotados, com suas respectivas estatísticas de teste *t*.

Nela pode-se verificar que:

- A variável que mede a eficiência dos municípios com base na circunvizinhança, W_1EFIC, é significativa em todos os quantis estudados e seu efeito é mais intenso entre os municípios menos eficientes;
- O recebimento de *royalties* (R_2) também é significativo em todos os quantis, porém, ao contrário da variável W_1EFIC, seu impacto é negativo e seu efeito mais intenso se dá entre os municípios que apresentam desempenhos mais fracos, isto é, nos quantis inferiores;
- Participação dos municípios em consórcios (E_{17}), poder de decisão dos conselhos municipais (E_{19}), densidade demográfica (E_{20}) e taxa de urbanização (E_{21}) são regressores que se mostram significativos e com resultados razoavelmente

homogêneos em todos os quantis. Isto indica que o impacto causado por cada uma dessas variáveis é aproximadamente o mesmo, independentemente do município estar entre os mais eficientes ou não; e

- As demais covariáveis são não-significativas em pelo menos um dos quantis avaliados. Neste rol está a idade do município (E_{25}), que não é significativa no modelo referente ao quantil 0,75. Contudo, ela apresenta uma evolução até surpreendente. Ela causa um forte impacto negativo ($-17{,}64\%$) no menor quantil e este impacto é suavizado à medida que os quantis aumentam, chegando a provocar impacto favorável à eficiência (5,29%) no quantil mais elevado. Estes resultados sugerem que cidades criadas a partir de municípios considerados pouco eficientes, ou mesmo ineficientes, tendem a prestar serviços de má qualidade à sociedade, ao passo que fazer a cisão em um município que tem boa estrutura administrativa e experiência na gestão dos fundos municipais pode gerar duas cidades com potencial de prestar bons serviços às suas populações. Aqui, supõe-se, então, que há uma transferência de conhecimentos ao novo município.

Tabela 7 - Estimativas dos coeficientes de regressão quantílica, onde a variável dependente é a eficiência técnica gerada pelo DEA-CCR

Covariável	Estimativa dos coeficientes segundo o quantil				
	0,10	0,25	0,50	0,75	0,90
Intercepto	−1,3042	−1,1699	−0,9542	−0,8076	−0,6047
t	−15,1646	−21,1651	−24,7058	−17,2045	−10,8028
W_1EFIC	0,0079	0,0044	0,0038	0,0039	0,0022
t	6,4583	4,2067	4,3408	4,0419	2,1555
E_3	0,0026	0,0026	0,0014	−0,0002	−0,0017
t	2,2871	2,6411	1,9132	−0,3201	−1,8017
E_4	−0,0001	−0,00006	−0,0001	−0,0001	−0,0001
t	−1,5984	−1,5296	−3,4023	−3,8032	−2,6686
E_7	−0,0374	−0,0138	−0,0311	−0,0347	−0,0552
t	−2,0292	−0,9713	−2,7354	−2,5241	−3,9550
E_{13}	−0,0653	−0,0537	−0,0497	−0,0345	−0,0397
t	−2,1975	−2,1468	−1,7658	−1,4497	−0,9495
E_{16}	−0,2230	−0,1682	−0,1170	−0,0206	0,0356
t	−2,8716	−3,5596	−4,5570	−0,5493	0,7733
E_{17}	−0,0568	−0,0628	−0,0631	−0,0660	−0,0688
t	−3,6540	−4,7689	−5,9284	−5,4801	−5,1238
E_{18}	0,1396	0,1369	0,1024	0,0497	0,0131
t	7,6768	9,1696	8,5500	3,8304	0,7591
E_{19}	0,0432	0,0371	0,0309	0,0380	0,0449
t	2,9197	3,0814	3,1884	3,3961	3,4699
E_{20}	0,00004	0,00003	0,00006	0,00004	0,00005
t	4,1627	6,5961	3,7487	4,1308	5,2509
E_{21}	0,0055	0,0057	0,0059	0,0058	0,0053
t	13,6593	17,2133	21,2814	19,6145	15,3219
E_{22}	0,0362	0,0726	0,0499	0,0528	0,0436
t	1,5514	6,4741	3,0151	3,0052	1,8001
E_{23}	−0,0443	−0,0777	−0,0595	−0,0628	−0,0776
t	−1,8231	−4,0903	−3,9286	−3,9969	−3,3532
E_{25}	−0,1764	−0,1388	−0,1010	−0,0332	0,0529
t	−8,7231	−6,5593	−5,6627	−1,5304	2,1500
CAP	0,2162	0,1920	0,1655	0,2249	0,1337
t	1,9378	6,5126	4,1809	2,6378	2,3540
R_2	−0,2391	−0,1715	−0,1450	−0,1052	−0,1046
t	−5,9280	−8,2110	−8,4196	−5,3424	−4,3370

Com o objetivo de comparar os resultados obtidos pelos modelos de regressão clássica, que estima a média condicional, com os resultados gerados pelos modelos de regressão quantílica, que estimam os quantis condicionais, o modelo referente ao quantil

da mediana foi reestimado sem as covariáveis não-significativas. O resultado está disposto na Tabela 8.

Com exceção das variáveis percentual de domicílios cujo chefe ganha até um salário mínimo (E_3) e PDT (E_{13}), que se mostraram não-significativas no modelo de regressão quantílica estimado a partir do DEA-CCR, os resultados obtidos pela média são semelhantes aos gerados pela mediana, notadamente no que se refere ao sentido da eficiência técnica estimada, ou seja, todas as covariáveis comuns aos dois modelos apresentaram os mesmos sinais.

Assim sendo, considerando as medidas de eficiência técnica geradas pelo DEA-CCR como variável dependente, as análises realizadas no modelo de regressão linear clássica são pertinentes aqui, no modelo de regressão quantílica, quando o quantil considerado é a mediana.

Tabela 8 - Descrição do modelo de regressão quantílica definitivo para o quantil 0,50, onde a variável resposta é a eficiência técnica gerada pelo método DEA-CCR

Covariável	Estimativa dos coeficientes	Estatística	
	Pontual	Intervalar[1]	teste[2]
Intercepto	−0,9256	−0,9825 ; −0,8481	−26,1100
W_1EFIC	0,0035	0,0019 ; 0,0049	4,1502
E_4	−0,0002	−0,00021 ; −0,00009	−4,2636
E_7	−0,0270	−0,0468 ; −0,0093	−2,3879
E_{16}	−0,1092	−0,1706 ; −0,0657	−3,9836
E_{17}	−0,0643	−0,0866 ; −0,0430	−6,0351
E_{18}	0,0997	0,0771 ; 0,1189	8,3578
E_{19}	0,0312	0,0147 ; 0,0501	3,2119
E_{20}	0,00006	0,00003 ; 0,00008	4,7192
E_{21}	0,0056	0,0053 ; 0,0061	21,7112
E_{22}	0,0569	0,0317 ; 0,0837	3,5920
E_{23}	−0,0527	−0,0822 ; −0,0300	−3,5359
E_{25}	−0,1038	−0,1319 ; −0,0671	−5,8404
CAP	0,1697	0,0881 ; 0,3213	3,6965
R_2	−0,1406	−0,1636 ; −0,1081	−7,5350

[1] **Intervalo de 95% de confiança calculado pelo método da inversão de testes *rankscore*.**

[2] **Calculada considerando variâncias estimadas robustas.**

2. A Partir do DEA-BCC

Igualmente ao item anterior, a análise de significância dos regressores dos modelos de regressão quantílica cuja variável dependente é a eficiência gerada pelo DEA-BCC, foi realizada com base nos dados constantes na Tabela 9, que exibe os coeficientes estimados para os quantis 0,10, 0,25, 0,50, 0,75 e 0,90, respectivamente.

Os resultados revelam que:

- As variáveis W_1EFIC, E_3, E_{21} e $,E_{23}$ são significativas e seus resultados apresentam pouca variabilidade ao longo dos quantis. Isto também foi verificado no modelo gerado pelo DEA-CCR, sendo que apenas a variável E_{21} pertence aos dois conjuntos com estas características;

Tabela 9 - Estimativas dos coeficientes de regressão quantílica, onde a variável dependente é a eficiência técnica gerada pelo DEA-BCC

Covariável	Estimativa dos coeficientes segundo o quantil				
	0,10	0,25	0,50	0,75	0,90
Intercepto	−1,3397	−1,1817	−0,8959	−0,7460	−0,5347
t	−22,8301	−18,9529	−18,0256	−28,4843	−8,4547
W_1EFIC	0,0084	0,0070	0,0054	0,0059	0,0052
t	6,2035	7,0312	5,9451	6,3346	3,7605
E_3	0,0049	0,0044	0,0024	0,0030	0,0020
t	5,4065	6,0233	3,5695	5,1724	2,1360
E_7	−0,0976	−0,0519	−0,0245	−0,0066	−0,0136
t	−4,7714	−3,0950	−1,9607	−0,4833	−0,6613
E_{14}	−0,1313	−0,0851	−0,0637	−0,0440	−0,0534
t	−6,8149	−4,4400	−3,3362	−3,1734	−2,7131
E_{16}	0,0780	0,1193	0,0708	0,1056	0,0852
t	1,4976	2,0461	1,5816	6,0747	1,5694
E_{17}	−0,0248	−0,0457	−0,0322	−0,0369	−0,0495
t	−1,7314	−3,2330	−2,7637	−3,5719	−2,8870
E_{18}	0,0363	0,0395	0,0318	0,0169	0,0052
t	2,5123	2,8232	2,5532	1,4909	0,2973
E_{19}	0,0566	0,0487	0,0216	0,0214	0,0007
t	4,0161	3,8380	1,9940	2,0344	0,0409
E_{21}	0,0013	0,0017	0,0016	0,0019	0,0024
t	3,3520	5,0158	5,6013	7,1828	5,6759
E_{23}	−0,0370	−0,0641	−0,0410	−0,0748	−0,0777
t	−2,5279	−3,9004	−2,7633	−5,0641	−3,7161
E_{25}	−0,0975	−0,0718	−0,0398	−0,0189	0,0120
t	−6,0074	−3,8038	−2,4412	−1,0075	0,5166
R_2	−0,1590	−0,1000	−0,0608	−0,0536	−0,0839
t	−5,6889	−4,1028	−2,8159	−2,9184	−3,9424

- Os partidos políticos, analogamente aos outros modelos já analisados, não conseguem converter suas práticas ideológicas em prêmios à eficiência dos seus municípios. Especificamente neste modelo, PMDB (E_7) e PPB (E_{11}), produzem seus piores resultados nos quantis inferiores e os menos ruins nos quantis superiores, salientando que a variável PMDB não é significativa nos modelos para os quantis 0,75 e 0,90, enquanto PPB é significativa em todos os quantis; e
- A idade do município (E_{25}) se mostrou significativa apenas nos três quantis inferiores. Neste caso, fica confirmado que é um risco criar cidades a partir de municípios que não conseguem oferecer bons serviços à sua coletividade; e
- O recebimento de *royalties* (R_2), aqui, apesar de apresentar resultados proporcionalmente menores do que no modelo de regressão quantílica gerado pelo DEA-CCR, tem efeito similar nos dois modelos.

A Tabela 10 apresenta as estimativas do modelo de regressão quantílica na mediana com todos os regressores significativos, no nível de 5%. Sua obtenção tem como objetivo a comparação dos seus resultados com os obtidos a partir da média condicional. Verifica-se que:

Tabela 10 - Descrição do modelo de regressão quantílica definitivo para o quantil 0,50, onde a variável resposta é a eficiência técnica gerada pelo método DEA-BCC

Covariável	Estimativa dos coeficientes	Estatística	
	Pontual	Intervalar[1]	teste[2]
Intercepto	−0,8116	−0,8631 ; −0,7758	−29,4227
W_1EFIC	0,0055	0,0041 ; 0,0069	6,1019
E_3	0,0025	0,0017 ; 0,0036	3,9013
E_7	−0,0270	−0,0481 ; −0,0121	−2,2131
E_{11}	−0,0649	−0,1010 ; −0,0380	−3,4798
E_{17}	−0,0383	−0,0573 ; −0,0218	−3,3278
E_{18}	0,0300	0,0120 ; 0,0460	2,4535
E_{21}	0,0015	0,0012 ; 0,0021	5,2312
E_{23}	−0,0499	−0,0687 ; −0,0316	−3,4866
E_{25}	−0,0384	−0,0622 ; −0,0154	−2,4031
R_2	−0,0613	−0,0899 ; −0,0270	−3,0778

[1] **Intervalo de 95% de confiança calculado pelo método da inversão de testes *rankscore*.**
[2] **Calculado considerando variâncias estimadas robustas.**

- No tocante ao sentido da eficiência técnica, os resultados obtidos pela regressão linear clássica e pela regressão quantílica são similares. Quanto às proporcionalidades, os resultados apresentam pequenas variações. Assim, as análises realizadas no modelo condicionado pela média servem para explicar o ocorrido no modelo obtido pela mediana. Este fato também se verifica quando a variável dependente é proveniente do DEA-CCR; e
- Um resultado inesperado aconteceu com a variável poder de decisão dos conselhos municipais, E_{19}, que era significativa e tornou-se não-significativa, quando da retirada da variável E_{16} do modelo.

5.2. Eficiência técnica estimada

Para fins comparativos com os modelos lineares de regressão clássica, os modelos de regressão quantílica apresentados nesta seção são referentes apenas ao quantil da mediana.

Como os modelos selecionados têm especificação *semilog*, foi necessário exponenciar os resultados para que as eficiências técnicas estimadas fossem obtidas.

Dessa forma, para os modelos de regressão clássica e quantílica identificadas, respectivamente, por $\hat{\theta}$ e $\tilde{\theta}$, tem-se:

$$\hat{\theta} = e^{X\hat{\beta}} \text{ e } \tilde{\theta} = e^{X\tilde{\beta}},$$

onde $\hat{\beta}$ e $\tilde{\beta}$ são os vetores dos parâmetros estimados através da regressão linear clássica e da regressão quantílica, levando em consideração os fatores condicionantes.

Retirando-se das eficiências técnicas obtidas pelos métodos DEA-CCR e DEA-BCC a influência dos fatores condicionantes, obtêm-se as estimativas das medidas de eficiência expurgadas, cujo cálculo foi realizado por

$$\hat{\hat{\theta}} = \theta - \hat{\theta} \quad e \quad \tilde{\tilde{\theta}} = \theta - \tilde{\theta}$$

para os modelos de regressão linear clássica e quantílica, respectivamente. Como pode-se notar, essas medidas de eficiência apresentam valores no intervalo [-1,1], escala incompatível com aquela resultante dos demais modelos. Para contornar esta dificuldade, esses resultados foram normalizados e, assim, as eficiências técnicas expurgadas passaram a ter a escala [0,1], comum nos outros modelos. A normalização foi feita a partir das expressões

$$\hat{\hat{\theta}}^* = \frac{\hat{\hat{\theta}} - \min\left(\hat{\hat{\theta}}\right)}{\max\left(\hat{\hat{\theta}} - \min\left(\hat{\hat{\theta}}\right)\right)} \quad e \quad \tilde{\tilde{\theta}}^* = \frac{\tilde{\tilde{\theta}} - \min\left(\tilde{\tilde{\theta}}\right)}{\max\left(\tilde{\tilde{\theta}} - \min\left(\tilde{\tilde{\theta}}\right)\right)}$$

para as medidas obtidas através da regressão linear clássica e regressão quantílica, respectivamente.

5.2.1. A partir do método DEA-CCR

Como era de se esperar, os resultados fornecidos pela técnica DEA-CCR estão no intervalo [0, 1], sendo o valor mínimo 0,0947. Metade dos municípios brasileiros tem eficiência máxima em torno de 0,5031, abaixo da eficiência média que é de 0,5222, e

dentre os 75% municípios menos eficientes, o que apresenta melhor desempenho tem eficiência máxima de apenas 0,6257. Estes resultados estão expostos na Tabela 11.

No que se refere aos modelos de regressão linear, tanto no tradicional quanto no quantílico, as medidas de eficiência expurgadas obtidas se concentram mais em torno da média, ou seja, os municípios menos eficientes apresentam melhor desempenho e os mais eficientes têm pior desempenho, quando comparadas com as medidas de eficiência geradas pelo modelo DEA-CCR (Tabela 11).

Tabela 11 - Medidas descritivas da eficiência técnica estimadas pelos modelos DEA-CCR, regressão clássica e regressão quantílica na mediana[1]

Medida descritiva	Eficiência técnica		
	Bruta (DEA-CCR)	Expurgada	
		Regressão clássica	Regressão quantílica
Mínimo	0,0947	0,0000	0,0000
1º Quartil	0,3952	0,3755	0,3897
Mediana	0,5031	0,4522	0,4655
Média	0,5222	0,4676	0,4809
3º Quartil	0,6257	0,5412	0,5546
Máximo	1,0000	1,0000	1,0000
Desvio padrão	0,1757	0,1340	0,1340
Coeficiente de variação	0,3365	0,2866	0,2787

[1] Regressão clássica e quantílica onde a variável dependente é a eficiência gerada pelo DEA-CCR.

Com alusão aos municípios e segundo as Tabelas 12 e 13, as medidas de eficiência dadas pelo DEA-CCR revelam que somente 85 deles são eficientes, o que representa 1,79% dos municípios estudados. Este resultado é ainda mais extremo quando a análise é feita a partir dos modelos de regressão, que apontam apenas um município plenamente eficiente: São João das Missões (MG), representando 0,02%.

Pode-se ainda verificar que 2 142 cidades, representando 45,05% dos municípios, têm suas eficiências estimadas pelo DEA-CCR no intervalo $[0{,}4\,;\,0{,}6)$. Pela ótica da regressão clássica e da quantílica, este percentual aumenta para 52,47% e 55,25%, significando 2 495 e 2 627 municípios, respectivamente, conforme mostram as Tabelas 12 e 13.

Tabela 12 - Quantidade de municípios de acordo com as medidas de eficiência técnica estimadas pelos modelos DEA-CCR, regressão clássica e regressão quantílica na mediana[1]

Eficiência técnica[2]	Modelo		
	DEA-CCR	Regressão clássica	Regressão quantílica
0,0 |— 0,2	32	34	28
0,2 |— 0,4	1 208	1 505	1 308
0,4 |— 0,6	2 142	2 495	2 627
0,6 |— 0,8	996	626	679
0,8 |— 1,0	292	94	112
1,0	85	1	1

[1] Regressão clássica e quantílica onde a variável dependente é a eficiência gerada pelo DEA-CCR.

[2] Para os modelos de regressão: eficiência técnica expurgada.

Tabela 13 - Percentual de municípios de acordo com as medidas de eficiência técnica estimadas pelos modelos DEA-CCR, regressão clássica e regressão quantílica na mediana [1]

Eficiência técnica[2]	Modelo		
	DEA-CCR	Regressão clássica	Regressão quantílica
0,0 |— 0,2	0,67	0,71	0,59
0,2 |— 0,4	25,40	31,65	27,51
0,4 |— 0,6	45,05	52,47	55,25
0,6 |— 0,8	20,95	13,17	14,28
0,8 |— 1,0	6,14	1,98	2,35
1,0	1,79	0,02	0,02

[1] Regressão clássica e quantílica onde a variável dependente é a eficiência gerada pelo DEA-CCR.

[2] Para os modelos de regressão: eficiência técnica expurgada.

Uma análise mais geral pode ser realizada por intermédio de visualização gráfica. As Figuras 1 e 2 exibem os histogramas das eficiências estimadas pelos três métodos aqui discutidos com suas respectivas densidades estimadas. Como já citado anteriormente, nota-se uma grande semelhança entre as previsões geradas pelos dois métodos.

Figura 1 – Histograma das medidas de eficiência geradas pelo DEA-CCR

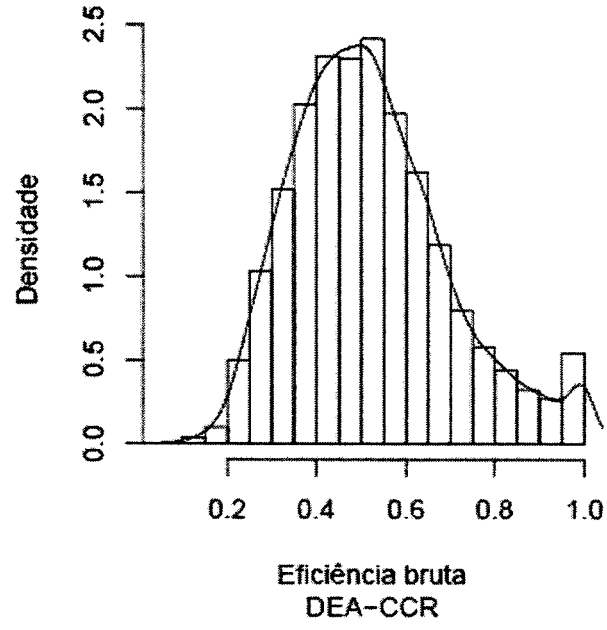
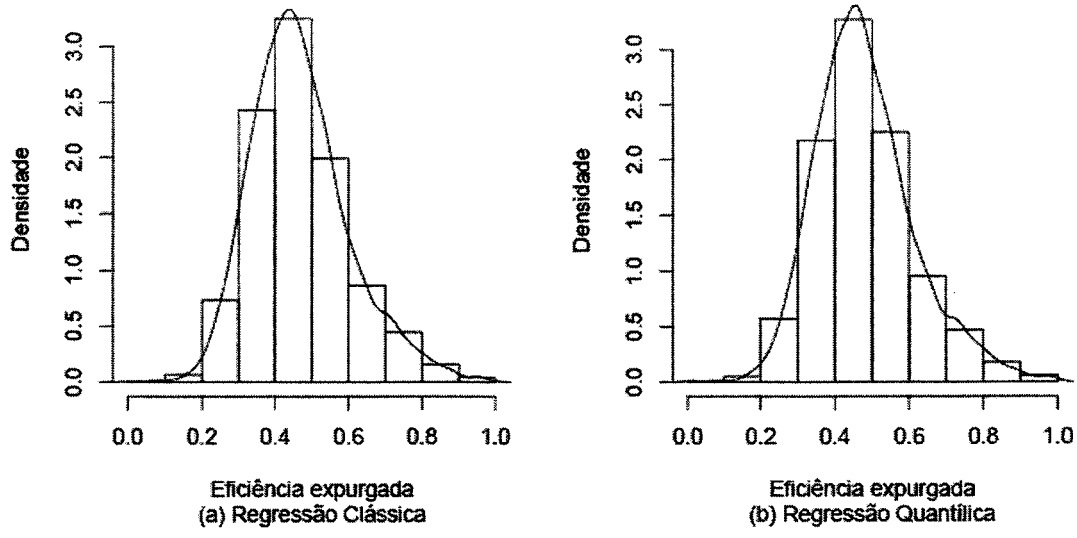


Figura 2 – Histograma das medidas de eficiência expurgadas geradas pelos modelos de regressão clássica e quantílica na mediana, a partir do DEA-CCR



5.2.2. A Partir do método DEA-BCC

Os resultados fornecidos pela técnica DEA-BCC, de maneira geral, são muito parecidos com os gerados pelo DEA-CCR. Isto é constatado pelos dados contidos nas Tabelas 14, 15 e 16, a partir dos quais se verifica que as duas técnicas fornecem resultados semelhantes. Situação análoga se dá com os modelos de regressão clássica e quantílica na mediana.

Tabela 14 - Medidas descritivas da eficiência técnica estimadas pelos modelos DEA-BCC, regressão clássica e regressão quantílica na mediana[1]

Medida descritiva	Eficiência técnica		
	Bruta (DEA-BCC)	Expurgada	
		Regressão clássica	Regressão quantílica
Mínimo	0,0774	0,0000	0,0000
1º Quartil	0,4093	0,3515	0,3505
Mediana	0,5073	0,4468	0,4448
Média	0,5249	0,4636	0,4624
3º Quartil	0,6198	0,5545	0,5543
Máximo	1,0000	1,0000	1,0000
Desvio padrão	0,1652	0,1597	0,1608
Coeficiente de variação	0,3148	0,3444	0,3479

[1] Regressão clássica e quantílica onde a variável dependente é a eficiência gerada pelo DEA-BCC.

Tabela 15 - Quantidade de municípios de acordo com as medidas de eficiência técnica estimadas pelos modelos DEA-BCC, regressão clássica e regressão quantílica na mediana[1]

Eficiência técnica(2)	Modelo		
	DEA-BCC	Regressão clássica	Regressão quantílica
0,0 |— 0,2	26	122	126
0,2 |— 0,4	1 078	1 672	1 661
0,4 |— 0,6	2 281	2 123	2 125
0,6 |— 0,8	1 041	655	659
0,8 |— 1,0	250	182	183
1,0	79	1	1

[1] Regressão clássica e quantílica onde a variável dependente é a eficiência gerada pelo DEA-BCC.
[2] Para os modelos de regressão: eficiência técnica expurgada.

No que se refere à quantidade de municípios conforme as faixas de eficiência adotadas neste trabalho, as duas técnicas DEA conduzem às mesmas conclusões, uma vez que os resultados são parecidos.

Tabela 16 - Percentual de municípios de acordo com as medidas de eficiência técnica estimadas pelos modelos DEA-BCC, regressão clássica e regressão quantílica na mediana[1]

Eficiência técnica(2)	Modelo		
	DEA-BCC	Regressão clássica	Regressão quantílica
0,0 |— 0,2	0,55	2,57	2,65
0,2 |— 0,4	22,67	35,16	34,93
0,4 |— 0,6	47,97	44,65	44,69
0,6 |— 0,8	21,89	13,77	13,86
0,8 |— 1,0	5,26	3,83	3,85
1,0	1,66	0,02	0,02

[1] **Regressão clássica e quantílica onde a variável dependente é a eficiência gerada pelo DEA-BCC.**

[2] **Para os modelos de regressão: eficiência técnica expurgada.**

Esta análise também pode ser realizada através de inspeção visual. As Figuras 3 e 4 exibem os histogramas com suas respectivas densidades estimadas para os modelos aqui tratados. Nelas é possível verificar a similaridade existente, não só entre os métodos de regressão, mas também entre as duas técnicas DEA.

Finalmente, a Tabela 17 mostra os coeficientes de correlação linear calculados para verificar a relação existente entre os modelos de regressão clássica e quantílica, onde o quantil de interesse é a mediana. Considerando a variável dependente como sendo a eficiência técnica obtida a partir do DEA-CCR e DEA-BCC, os resultados foram 0,9971 e 0,9979, respectivamente, indicando que a correlação linear existente é quase perfeita. Estes resultados ratificam todas as análises comparativas feitas até então, inclusive aquelas realizadas através da visualização gráfica, ou seja, as duas técnicas geram, de fato, estimativas semelhantes.

Figura 3 – Histograma das medidas de eficiência geradas pelo DEA-BCC

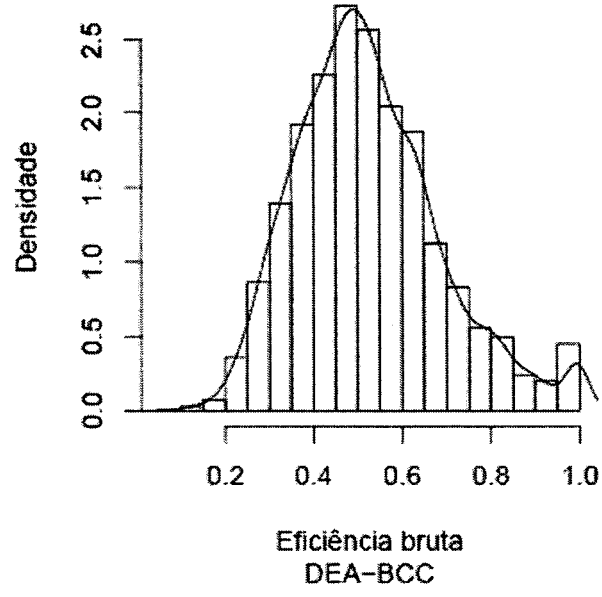


Figura 4 – Histograma das medidas de eficiência expurgadas pelos modelos de regressão clássica e quantílica na mediana, a partir do DEA-BCC

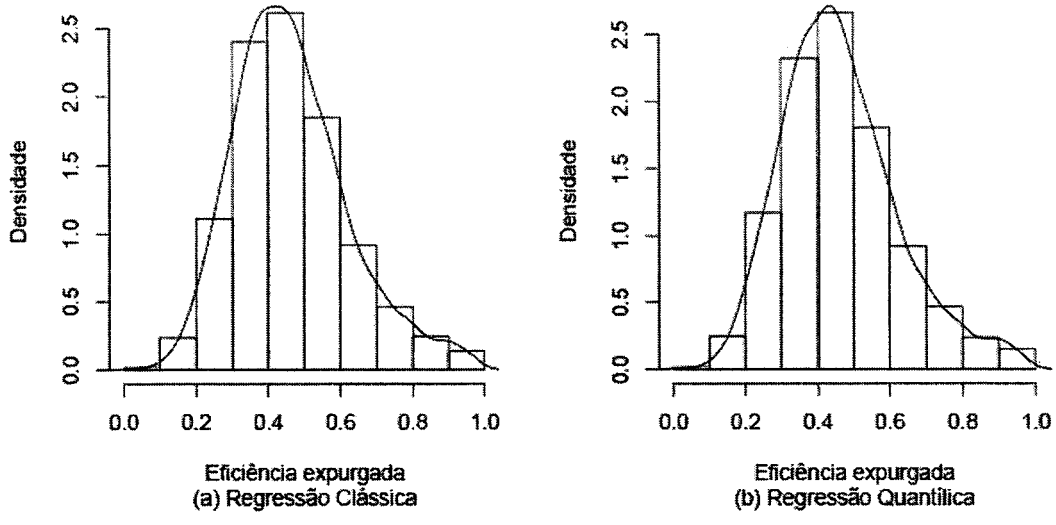


Tabela 17 - Coeficientes de correlação linear entre as medidas de eficiência técnica expurgadas, obtidas pelos métodos de regressão clássica e quantílica para a mediana

Variável dependente	Coeficientes de correlação linear
Eficiência gerada pelo DEA-CCR	0,9971
Eficiência gerada pelo DEA-BCC	0,9979

6. Conclusões

O presente artigo teve como objetivo a avaliação estatística da eficiência do gasto público municipal no Brasil, principalmente no que tange ao impacto da idade do município e dos fatores externos à referida eficiência. Para tal, foi utilizado como base o trabalho desenvolvido por Sampaio de Sousa, Cribari-Neto e Stošić (2005), que abordaram a análise envoltória de dados (DEA-CCR e DEA-BCC), métodos de reamostragens (*jackknife*, *bootstrap* e *jackstrap*), além de regressão clássica e regressão quantílica.

As principais conclusões estão sumariadas a seguir:

- Os métodos não-paramétricos DEA-CCR e DEA-BCC se mostraram eficazes no cálculo das eficiências técnicas de municipalidades brasileiras, principalmente com o uso do esquema *jackstrap*. Por sua vez, essas medidas de eficiência se revelaram bastantes úteis, como variáveis dependentes, nos modelos paramétricos aqui utilizados: regressão clássica e regressão quantílica;
- De uma forma geral, os modelos de regressões clássica e quantílica na mediana, obtidos a partir do DEA-CCR e do DEA-BCC, apresentaram resultados semelhantes. Contudo, a regressão quantílica, por atuar em diferentes quantis, mostrou-se mais elucidativa, uma vez que a análise dos seus resultados transcendeu os limites das medidas de centralidade;

- Os fatores exógenos exercem fortes influências nas eficiências geradas pelos modelos estimados através dos fatores condicionantes, chegando a serem responsáveis, em alguns casos, por mais de 50% da eficiência dos municípios;
- Dos indicadores de localização, o efeito da vizinhança entre as cidades e o efeito da localização no Polígono das Secas foram estatisticamente significativos em todos os modelos analisados. O primeiro exerce um efeito positivo e o segundo, como esperado, causa impacto negativo;
- Os municípios mais pobres tendem a administrar melhor seus escassos recursos. Esta afirmativa tem suporte no percentual de domicílios cujo chefe ganha até um salário mínimo e na renda média do trabalhador, que são estatisticamente significativos;
- Os indicadores administrativos são importantes na avaliação do município. Quanto maior o poder dos conselhos municipais e quanto maior o grau de informatização do município, maior é a tendência dessas cidades prestarem bons serviços à comunidade. Isto porque os conselhos fiscalizam a aplicação dos recursos, diminuindo a corrupção e o abuso na distribuição de verbas, enquanto o uso de recursos computacionais agiliza processos e diminui custos. Entretanto, os consórcios municipais, ao contrário do que se esperava, têm impacto negativo, o que sugere uma reformulação dos mesmos para que o objetivo de minimização de gastos volte a ser atingido;
- A estratégia administrativa para o recebimento de *royalties* deve ser revista pelos municípios, visto que esta receita afeta negativamente a eficiência administrativa em todos os modelos estudados, inclusive nos diversos quantis da regressão quantílica; e
- Os resultados relativos à idade do município, objeto central desse trabalho, confirmam que a falta de estrutura administrativa e de experiência em gestão pública onera os municípios jovens, tornando-os ineficientes. Os resultados mostram ainda que cidades criadas a partir de municípios pouco eficientes, ou mesmo ineficientes, tendem a operar ineficientemente, enquanto a divisão

de municípios que têm boa estrutura gerencial e experiência na gestão dos fundos municipais gera novas municipalidades com potencial para oferecer bons serviços às suas populações. Assim, fica caracterizado o risco que há ao se criar cidades sem levar em consideração os critérios de eficiência aqui tratados.

Referências bibliográficas

ANSELIN, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

BANKER, R.D., CHARNES, A. e COOPER, W.W. (1984). Some models for estimating technical and scale efficiencies in data envelopment analisys. *Management Science*, 30:1078-1092.

BANKER, R.D. e GIFFORD, J.L. (1988). A relative efficiency method for the evaluation of public health nurse productivity. Mimeo.

CHARNES, A., COOPER, W.W. e RHODES, E. (1978). Measuring efficiency of decision making units. *European Journal of Operational Research*, 1:429-44.

CRIBARI-NETO, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, 45:215-233.

CRIBARI-NETO, F. e ZARKOS, S. (2004). Leverage-adjusted heteroskedastic bootstrap methods. *Journal of Statistical Computation and Simulation*, 74:215-232.

DAVIDSON, R. & MACKINNON, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

GREENE, W.H. (1981). On the asymptotic bias of the ordinary least squares estimator of the tobit model. *Econometrica*, 49:505-513.

KOENKER, R. E BASSET, G. (1978). Regression quantiles. *Econometrica*, 46:33-50.

SAMPAIO DE SOUSA, M.C., CRIBARI-NETO, F. e Stošić, B. D. (2005). Explaining DEA technical efficiency scores in an outlier corrected environment: the case of public services in Brazilian municipalities. *Brazilian Review of Econometrics*, 25:289-315.

SAMPAIO DE SOUZA, M.C. e STOŠIĆ, B.D. (2005). Technical efficiency of the Brazilian municipalities: correcting nonparametric frontier measurements for outliers. *Journal of Productivity Analysis*, 24:155-179.

SEAVER, B.L. e TRIANTIS, K.P. (1989). The implications of using messy data to estimate production-frontier-based technical efficiency measures. *Journal of Business and Economic Statistics*, 7:49-59.

TOBIN. J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26:24-36.

WILSON, P. (1993). Detecting influential observations in data envelopment analysis. *Journal of Productivity Analysis*, 6:27-45.

WILSON, P. (1995). Detecting influential observations in deterministic non-parametric frontiers models. *Journal of Business and Economic Statistics*, 11:319-323.

Abstract

This article addresses the issue of evaluating the efficiency of public spending at Brazilian municipalities. The chief goal of the article is to measure the impact of the municipalities' ages on such efficiency. To that end, we have used the dataset compiled by Sampaio de Sousa, Cribari-Neto and Stošić (2005) together with further information from "Programa das Nações Unidas para o Desenvolvimento". The empirical analysis was carried out using three approaches: DEA (data envelopment analysis), classical linear regression and quantile regression. The regression techniques were employed in order to identify the exogenous factors that explain the differences in efficiency. Our results show that young municipalities tend to be less efficient than well established ones.

Modelos fatoriais para retornos de ativos

Luís Gustavo do Amaral Vinha[*]
Chang Chiann[*]

Resumo

A estrutura de covariância entre os retornos dos ativos representa uma informação fundamental para a formação de carteiras de investimentos. Os fatores responsáveis pelo risco sistemático dos ativos podem ser identificados através dos modelos de fatores, o que possibilita uma estimação robusta das covariâncias entre os retornos. O objetivo do presente estudo é identificar fatores fundamental e estatístico que influenciam os retornos de ativos do mercado brasileiro. A geração dos fatores estatísticos e o estudo da influência desses fatores nos retornos foram feitos através da análise fatorial tradicional e pela análise assintótica de componentes principais. Modelos de regressão *cross-section* foram utilizados no estudo da importância dos fatores fundamentais. Entre os modelos estudados, os modelos de fatores estatísticos apresentaram os melhores resultados.

[*] Endereço para correspondência: Departamento de Estatística IME – USP – CP. 66281, CEP 05315-970 – *e-mail*: luisgav@isp.edu.br e *e-mail:* chang@ime.usp.br.

1. Introdução

A formação de carteiras de investimentos procura minimizar o risco para um determinado valor de retorno esperado ou maximizar o retorno para um determinado nível de risco que se deseja assumir. A covariância entre os retornos dos ativos é uma informação fundamental para a otimização de uma carteira, logo o desempenho da carteira depende da estimação dessas covariâncias. Os modelos fatoriais representam uma alternativa para a estimação, através desses modelos é possível verificar a existência de fatores responsáveis pelo risco sistemático dos ativos, o que possibilita uma estimação robusta das covariâncias entre os retornos.

De forma geral os modelos fatoriais para retornos de ativos podem ser divididos em três tipos: macroeconômico, fundamental e estatístico. No modelo macroeconômico os fatores são relacionados com variáveis que representam o cenário econômico do país como, por exemplo, taxa de inflação e produção industrial. No modelo fundamental os fatores são relacionados com características observáveis das empresas como valor de mercado e alavancagem. No modelo estatístico os fatores não são diretamente observáveis e são obtidos a partir de retornos observáveis através de métodos estatísticos.

Um dos modelos fatoriais mais importantes e mais discutidos na literatura é o modelo *Capital Asset Pricing Model* - CAPM. O CAPM preconiza que todo o risco sistemático dos ativos é capturado apenas pelo índice de mercado. As diversas hipóteses simplificadoras assumidas neste modelo geraram inúmeras críticas, para maiores detalhes ver Grinold e Kahn (2000). Ross (1976) apresenta uma alternativa ao CAPM, o modelo *Arbitrage Pricing Mode* - APT, nesse modelo assume-se que o risco sistemático dos ativos é multidimensional, ou seja, existem outros fatores além do índice de mercado que influenciam os movimentos dos ativos.

Em relação aos fatores macroeconômicos, Schor, Bonomo e Valls (2002) estudaram a importância destes fatores sob os retornos de dez carteiras de empresas do mercado brasileiro classificadas por setor de atividade. Neste estudo, os fatores

relacionados com as variáveis macroeconômicas foram estatisticamente significativos para a maioria das carteiras.

Os modelos fatoriais estudados têm em comum a mesma formulação geral

$$R_{it} = a_i + \beta_{1i} f_{1t} + \beta_{2i} f_{2t} + \ldots + \beta_{ki} f_{kt} + \varepsilon_{it}, \qquad (1)$$

onde

- R_{it} é o retorno real ou o excesso de retorno do ativo i ($i = 1, \ldots, N$) no período t ($t = 1, \ldots, T$);
- α_i é o intercepto do modelo;
- f_{kt} é a realização do fator k ($k = 1, \ldots, K$) no período t;
- β_{ki} é a sensibilidade do ativo i em relação ao fator k; e
- ε_{it} é o erro ou risco específico do ativo i no período t.

Assume-se que os erros têm média zero e são não correlacionados com as realizações dos fatores, isto é,

$E(\varepsilon_{it}) = 0$ para todo i e t,

$Cov(f_{kt}, \varepsilon_{it}) = 0$ para todo k, i e t.

Além disso, supõe-se que os erros não apresentam correlação contemporânea e serial, ou seja,

$$Cov(\varepsilon_{it}, \varepsilon_{js}) = \begin{cases} \sigma_i^2, \textit{ para } t = s \textit{ e } i = j; \\ 0, \textit{ caso contrário.} \end{cases}$$

O objetivo do estudo é identificar os fatores estatístico e fundamental que influenciam os retornos dos ativos, determinar as sensibilidades dos ativos em relação aos fatores e quantificar a variabilidade total dos retornos explicada pelos fatores.

2. Modelo de fatores estatísticos

No modelo fatorial estatístico os fatores não são diretamente observáveis e são extraídos dos retornos observáveis, utilizando métodos estatísticos como, por exemplo,

análise fatorial. A análise fatorial tradicional é utilizada quando o tamanho da série temporal observada T é maior que o número de ativos N. Quando $T < N$, a análise assintótica de componentes principais (Connor e Korajczyk, 1986) é a técnica mais apropriada.

2.1. Análise fatorial tradicional

O modelo de análise fatorial é aplicado ao estudo do comportamento dos retornos. Seja $\boldsymbol{R}_t$ um vetor dos retornos de N ativos com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Omega}$, e que retornos mensais de ativos não apresentam correlação serial, o modelo de análise fatorial na forma matricial é dado por

$$\boldsymbol{R}_t = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T,$$

onde

- $\boldsymbol{R}_t$ é o vetor $(N \times 1)$ dos retornos dos ativos no período t;
- $\boldsymbol{\mu}$ é o vetor $(N \times 1)$ dos retornos médios;
- $\boldsymbol{B}$ é a matriz $(N \times K)$ das cargas fatoriais;
- $\boldsymbol{f}_t$ é o vetor $(K \times 1)$ das realizações dos fatores comuns ou fatores estatísticos; e
- $\boldsymbol{\varepsilon}_t$ é o vetor $(N \times 1)$ dos erros ou fatores específicos.

Suposições:

(S1) $E(\boldsymbol{f}_t) = \boldsymbol{0}$ e $Var(\boldsymbol{f}_t) = \boldsymbol{I}_K$;

(S2) $E(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}$ e $Var(\boldsymbol{\varepsilon}_t) = \boldsymbol{D} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix}$; e

(S3) $Cov(\boldsymbol{f}_t ; \boldsymbol{\varepsilon}_t) = \boldsymbol{0}$, para qualquer t.

Cargas fatoriais

As cargas fatoriais podem ser interpretadas a partir da covariância entre os fatores e os retornos dos ativos. Considere a covariância existente entre o retorno R_{it} e o fator f_{jt} para um determinado período t,

$$Cov(R_{it}; f_{jt}) = Cov(\mu_i + \beta_{i1}f_{1t} + \beta_{i2}f_{2t} + ... + \beta_{ij}f_{jt} + ... + \varepsilon_{it}; f_{jt}).$$

Pelas suposições (S1) e (S3) apresentadas anteriormente tem-se

$$Cov(R_{it}; f_{jt}) = Cov(\beta_{ij}f_{jt}; f_{jt}) = \beta_{ij},$$

ou seja, as cargas fatoriais são as covariâncias entre os retornos e os fatores.

Comunalidade e especificidade

A matriz de covariâncias dos retornos pode ser decomposta em

$$\Omega = BB' + D,$$

de forma que os K fatores estão presentes na composição das variâncias dos retornos e das covariâncias entre eles. A variância do retorno do ativo i pode ser expressa por

$$\sigma_{Ri}^2 = Var(R_{it}) = \sum_{j=1}^{K} \beta_{ji}^2 + \sigma_i^2 = h_i^2 + \sigma_i^2 .$$

Pode-se verificar que h_i^2 representa a parcela da variância de R_{it} explicada pelos fatores comuns e σ_i^2 representa a parcela não explicada pelos fatores. Denomina-se *comunalidade* a parcela da variância explicada pelos fatores e *especificidade* a parcela não explicada.

Estimação

Os métodos mais importantes para obtenção dos fatores na análise fatorial são: método da máxima verossimilhança e o método das componentes principais. O método da máxima verossimilhança pode ser utilizado sob a suposição de normalidade conjunta dos retornos, enquanto o método das componentes principais não requer qualquer suposição acerca da distribuição. Para maiores detalhes ver Johnson e Wichern (1998).

Rotação dos fatores

A formulação do modelo de análise fatorial apresenta um resultado interessante com relação à matriz $\boldsymbol{B}$. Seja $\boldsymbol{B}^{*}=\boldsymbol{BH}$, onde $\boldsymbol{H}$ é uma matriz ortogonal[1], tem-se

$$\boldsymbol{B}^{*}\boldsymbol{B}^{*\prime}+\boldsymbol{D}=\boldsymbol{(BH)\,(BH)}'+\boldsymbol{D}=\boldsymbol{BHH'B}+\boldsymbol{D}=\boldsymbol{BB}'+\boldsymbol{D}.$$

Portanto, se $\boldsymbol{B}$ satisfaz a relação acima, existem infinitas matrizes que também satisfazem a relação. Esse resultado, em termos práticos, permite a criação de novos fatores, os fatores rotacionados. Em muitos casos, os fatores encontrados em uma solução inicial não são facilmente interpretáveis, mas, pelo resultado apresentado, é possível propor rotações que geram fatores mais fáceis de serem interpretados.

Determinação do número de fatores

O número de fatores a serem utilizados no modelo é uma parte muito importante da análise. A escolha equivocada do número de fatores pode ocasionar a perda de informação importante, ou a inclusão de algum ruído. Existem vários critérios para determinação do número de fatores na literatura, nesse estudo foram utilizados: porcentagem da variabilidade explicada, *scree plot*, critério de Kaizer e o método proposto por Lambert et al (1990) que utiliza o método de reamostragem *Bootstrap* para determinar intervalos de confiança para os autovalores da matriz de correlações dos retornos. Para maiores detalhes ver Vinha (2006).

---

[1] Ou seja, $\boldsymbol{H'H}=\boldsymbol{HH'}=\boldsymbol{I}$.

2.2. Análise assintótica de componentes principais

A análise assintótica de componentes principais, proposta e desenvolvida por Connor e Korajczyk (1986), é utilizada quando o número de ativos é maior que o tamanho da série temporal observada ($N > T$). Essa técnica é baseada na matriz de covariâncias ($T \times T$) dada por

$$\Omega^T = (1/N)\boldsymbol{R}'\boldsymbol{R},$$

onde $\boldsymbol{R}$ é uma matriz $N \times T$ dos retornos dos N ativos nos T períodos.

Uma das vantagens desse método é que quando o número de ativos N é muito maior que o número de períodos T são calculados apenas os T autovetores da matriz Ω^T enquanto na análise de componentes principais tradicional são calculados os N autovetores da matriz Ω.

Connor e Korajczyk (1986) demonstram que as K primeiras componentes principais da matriz Ω^T são aproximadamente equivalentes aos K fatores da análise fatorial quando o número de ativos N é grande. A demonstração deste resultado é apresentada no Apêndice 1. Portanto, neste caso, os autovetores da matriz Ω^T são os estimadores dos fatores estatísticos.

Refinamento da estimação dos fatores

A estimação dos fatores citada acima não considera a heterocedasticidade dos erros, por essa razão os mesmos autores propuseram um refinamento com o objetivo de aumentar a eficiência do método (Connor e Korajczyk, 1988). A modificação sugerida propõe o cálculo dos autovetores dos retornos reescalonados, um procedimento análogo ao uso do método de mínimos quadrados ponderados em modelos de regressão. Este procedimento é composto pelos seguintes passos:

1. Estimar os fatores f_t através do cálculo dos autovetores de Ω^T; e

2. Para cada ativo estimar β_i e σ_i^2 através das regressões em séries temporais dadas por

$$R_{it} = \alpha_i + \beta_i' \hat{f}_t + \varepsilon_{it},$$

onde $\beta_i' = [\beta_{1i}, \beta_{2i}, ..., \beta_{Ki}]$ e $\hat{f}_t = [\hat{f}_{1t}, \hat{f}_{2t}, ..., \hat{f}_{Kt}]$.

Utilizar as estimativas das variâncias para gerar a matriz de covariâncias dos resíduos

$$\hat{\boldsymbol{D}} = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & 0 \\ 0 & \hat{\sigma}_2^2 & 0 \\ 0 & 0 & \hat{\sigma}_N^2 \end{bmatrix}.$$

3. Determinar a matriz $N \times T$ dos retornos reescalonados

$$\boldsymbol{R}^* = \hat{\boldsymbol{D}}^{-1/2} \boldsymbol{R},$$

e recalcular a matriz de covariâncias

$$\hat{\boldsymbol{\Omega}}^{T^*} = \frac{1}{N} \boldsymbol{R}^{*'} \boldsymbol{R}^*;$$

4. Reestimar os fatores através do cálculo dos autovetores de $\hat{\boldsymbol{\Omega}}^{T^*}$.

Determinação do número de fatores

A determinação do número de fatores para a análise assintótica de componentes principais é apresentada por Connor e Korajczyk (1993). O procedimento é baseado na comparação dos resíduos do modelo com K fatores com os resíduos do modelo com $K+1$ fatores. O número de fatores é determinado através dos seguintes passos:

1. Dadas as séries dos retornos e dos $K+1$ fatores estimados, são geradas as séries dos resíduos dos modelos

$$R_{it} = \alpha_i + \boldsymbol{\beta}_i' \hat{f}_t + \varepsilon_{it},$$

$$R_{it} = \alpha_i + \boldsymbol{\beta}_i' \hat{f}_t + \beta_{K+1,i} f_{K+1,t} + \varepsilon_{it}^{*}.$$

2. Com os resíduos calculados acima, calcular os resíduos quadrados com correção do número de graus de liberdade

$$\hat{\sigma}_{it}^2 = \frac{\hat{\varepsilon}_{it}^2}{1-(K+1)/T - K/N},$$

$$\hat{\sigma}_{it}^{*2} = \frac{\hat{\varepsilon}_{it}^{*2}}{1-(K+3)/T - (K+1)/N}.$$

3. Calcular a diferença dos resíduos quadrados de períodos pares e ímpares

$$\hat{\Delta}_s = \hat{\mu}_{2s-1} - \hat{\mu}_{2s}^{*}, \quad s = 1,\ldots,T/2,$$

onde

$$\hat{\mu}_t = \frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}_{it}^2,$$

$$\hat{\mu}_t^{*} = \frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}_{it}^{*2}.$$

Como resultado tem-se o vetor das diferenças

$$\hat{\Delta} = (\hat{\Delta}_1, \hat{\Delta}_2, \ldots, \hat{\Delta}_{T/2})'.$$

4. Calcular a média e a variância das diferenças

$$\overline{\Delta} = \frac{2}{T}\sum_{s=1}^{T/2} \hat{\Delta}_s,$$

$$\hat{\sigma}_{\Delta}^2 = \frac{2}{T-2}\sum_{s=1}^{T/2} (\hat{\Delta}_s - \overline{\Delta})^2.$$

5. Calcular a estatística t para verificar se a diferença média é estatisticamente maior que zero,

$$t = \frac{\overline{\Delta}}{\hat{\sigma}_{\Delta}}.$$

Se, para um determinado K, a média das diferenças é estatisticamente maior que zero significa que os erros do modelo com K fatores são, em média, maiores que os erros do modelo com $K+1$ fatores. Logo o novo fator é importante e deve ser incorporado ao modelo. Denomina-se K^* o número de fatores indicado para o modelo onde $K^* = \inf\{K$ tal que a hipótese de que a diferença média é igual a zero não é rejeitada$\}$.

3. Modelo de fatores fundamentais

Com esse modelo é possível estudar a influência de fatores relacionados com as características das empresas nos retornos dos ativos. Os fatores fundamentais são obtidos a partir de características específicas e observáveis das empresas. A metodologia utilizada foi desenvolvida primeiramente por Bar Rosenberg e explorada por Grinold e Kahn (2000). Os modelos utilizados são chamados de modelos do tipo "BARRA" onde as características observáveis dos ativos são tratadas como as cargas dos fatores. As realizações dos fatores são não observáveis e são estimadas através de regressões *cross-section*.

Diversas características das empresas podem ser consideradas para a construção do modelo de fatores fundamentais, nesse estudo são utilizadas as variáveis: valor de mercado, valor contábil, lucro/preço, alavancagem financeira[2] e variáveis de classificação.

[2] Os dados disponíveis da alavancagem financeira das empresas são trimestrais, logo foi necessária uma interpolação para estimar os dados mensais. A estimação dos valores mensais foi feita através do método *Locally Weighted Scatter Plot Smoothing* - LOWESS, para maiores detalhes ver Cleveland (1979).

Modelos tipo "BARRA"

No modelo "BARRA" de fatores relacionados com os tipos de indústrias são geradas diversas variáveis de classificação para identificar os tipos de indústrias ao qual as empresas estudadas pertencem. Os termos de sensibilidade β_{ki} em (1) nesse caso são dados por

$$\beta_{ki} = \begin{cases} 1, \textit{ se o ativo } i \textit{ pertence à categoria } k; \\ 0, \textit{ caso contrário.} \end{cases}$$

O modelo, apresentado na forma de regressões *cross-section* para cada período t, é dado por

$$\boldsymbol{R}_t = \boldsymbol{\beta}_1 f_{1t} + \ldots + \boldsymbol{\beta}_K f_{Kt} + \varepsilon_t,\ t = 1, \ldots, T, \qquad (2)$$

onde

- $\boldsymbol{R}_t$ é um vetor $(N \times 1)$ dos retornos no período t;
- $\boldsymbol{\beta}_k$ é o vetor $(N \times 1)$ das cargas do fator k;
- f_{kt} é a realização do fator k no período t;
- ε_t é um vetor $(N \times 1)$ dos erros, com média zero e matriz de covariâncias $\boldsymbol{D}$.

A matriz $\boldsymbol{D}$ é diagonal com os termos σ_i^2 ao longo da diagonal, logo os erros podem ser heterocedásticos e não apresentam correlação entre os ativos.

(2) pode ser representado na forma matricial

$$\boldsymbol{R}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t,$$

onde

- B é a matriz N × K das cargas dos fatores; e
- f_t é o vetor das realizações dos fatores.

Os termos f_t podem ser estimados pelo método dos mínimos quadrados ordinários

$$\hat{f}_{t,MQO} = (B'B)^{-1}B'R_t.$$

Esse estimador não é viesado, mas não é eficiente devido ao fato de que os erros são heterocedásticos. Pode-se utilizar, então, o método de mínimos quadrados ponderados para obter o estimador eficiente de f_t. Assumindo que as variâncias σ_i^2 são conhecidas, o estimador é dado por

$$\hat{f}_{t,MQP} = (B' D^{-1} B)^{-1} B' D^{-1} R_t .$$

Como as variâncias σ_i^2 não são conhecidas, o estimador apresentado acima não pode ser utilizado. Uma alternativa é o método de mínimos quadrados ponderados factíveis que utiliza uma estimativa da matriz de covariâncias. O novo estimador é dado por

$$\hat{f}_{t,MQPF} = (B' \hat{D}^{-1} B)^{-1} B' \hat{D}^{-1} R_t , \qquad (3)$$

onde os elementos da diagonal da matriz $\hat{D}$ são estimados a partir dos resíduos do modelo ajustado por mínimos quadrados ordinários da seguinte forma

$$\hat{\sigma}_i^2 = \frac{1}{T-1} \sum_{t=1}^{T} \left(\hat{\varepsilon}_{it,MQO} - \bar{\varepsilon}_{i,MQO} \right)^2 ,$$

$$\bar{\varepsilon}_{i,MQO} = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_{it,MQO} .$$

Dessa forma são geradas as séries dos K fatores para os T períodos. O mesmo modelo pode ser utilizado não somente com variáveis de classificação, mas também para quaisquer outros atributos das empresas, nesse estudo foram utilizadas as variáveis contábeis.

4. Aplicação

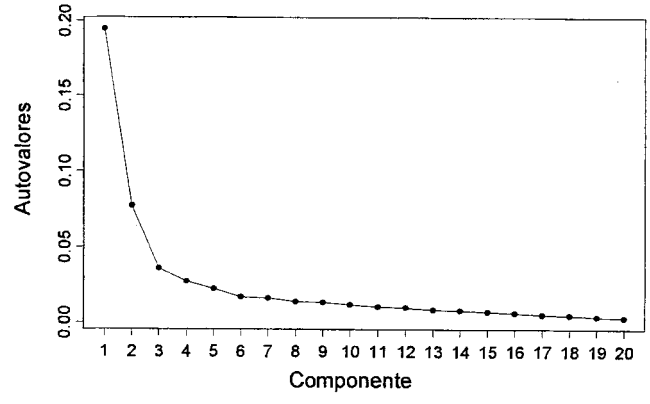
A aplicação dos métodos apresentados anteriormente está dividida em duas partes: na primeira é aplicada à análise fatorial em um conjunto de dados com os retornos de 20 ativos no período de 117 meses e os resultados são comparados com os resultados de um modelo com fatores macroeconômicos (utilizando o mesmo procedimento apresentado por Schor, Bonomo e Valls, 2000); na segunda parte é utilizada a análise assintótica de componentes principais e os modelos tipo "BARRA" para um conjunto de dados com os retornos de 136 ativos no período de 41 meses.

Os retornos dos ativos e as variáveis contábeis utilizadas neste estudo foram obtidos no Banco de Dados Economática, no Apêndice 2 são apresentados os gráficos dos retornos dos dois conjuntos de dados.

4.1. Parte 1

O primeiro passo da análise fatorial é a determinação do número de fatores a serem utilizados no modelo. Analisando o *scree-plot* (Figura 1) nota-se que a partir do sexto componente as diferenças entre os autovalores da matriz de covariâncias dos retornos são muito pequenas. Por esse critério existe uma indicação de que cinco componentes devem ser utilizados.

Figura 1 - *Scree-plot* – análise fatorial



A Tabela 1 apresenta os cinco primeiros autovalores e a média das variâncias dos retornos. Verifica-se que os quatro primeiros autovalores são maiores que a média das variâncias, logo, pelo critério de Kaizer, devem ser utilizados quatro fatores no modelo.

Tabela 1 - Autovalores da matriz de covariância dos retornos

λ_1	λ_2	λ_3	λ_4	λ_5	Média
0,194	0,077	0,036	0,027	0,023	0,025

Foram construídos os intervalos de confiança para os autovalores, utilizando a reamostragem *Bootstrap* (Efron, 1993 ou Lambert et al, 1990). Os intervalos de confiança dos seis primeiros autovalores são apresentados na Tabela 2, observa-se que o intervalo de confiança do sexto componente está abaixo da média. Por este resultado, os cinco primeiros componentes deveriam ser considerados no ajuste do modelo.

Tabela 2 - Intervalos de confiança dos autovalores da matriz de covariância dos retornos

Autovalor	Intervalo de Confiança ($\gamma = 95\%$)
λ_1	[0,115; 0,306]
λ_2	[0,021; 0,132]
λ_3	[0,020; 0,056]
λ_4	[0,019; 0,035]
λ_5	[0,017; 0,027]
λ_6	[0,015; 0,022]*

*** o intervalo está abaixo de 0,025 (média das variâncias dos retornos)**

Os critérios apresentados indicam que o modelo deve ter quatro ou cinco componentes. Pelas Tabelas 3 e 4, verifica-se que o modelo com quatro fatores explica 54,86% e o modelo com cinco fatores explica 57,53% da variabilidade total dos retornos, logo a diferença da porcentagem da variabilidade explicada pelos modelos é pequena. Decidiu-se utilizar quatro fatores no modelo.

Tabela 3 - Importância dos fatores - Modelo com quatro fatores

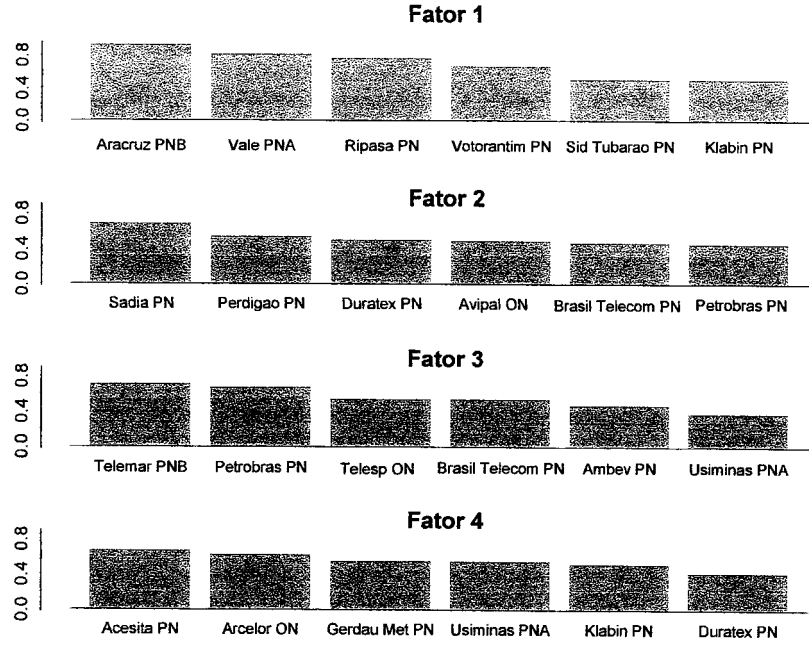
	Fator 1	Fator 2	Fator 3	Fator 4
Proporção individual (%)	17,69	12,57	12,31	12,29
Proporção acumulada (%)	17,69	30,26	42,56	54,86

Tabela 4 - Importância dos fatores - modelo com cinco fatores

	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
Proporção individual (%)	18,09	14,12	10,17	9,38	5,77
Proporção acumulada (%)	18,09	32,21	42,38	51,76	57,53

Na Figura 2, são apresentadas as cargas fatoriais dos retornos dos ativos, quanto maior a carga fatorial maior a importância do ativo na composição do fator. Observa-se que o primeiro fator encontrado representa a Vale do Rio Doce e as indústrias do setor de Papel e Celulose; o segundo fator representa basicamente as empresas do setor de Alimentos; o terceiro representa as empresas de Telecomunicações e a Petrobrás; e o quarto fator representa o setor de Siderurgia e Metalurgia.

Figura 2 - Cargas fatoriais – análise fatorial



Pela Tabela 5 verifica-se que os retornos dos ativos Aracruz PNB, Vale PNA, Petrobras PN e Sadia PN apresentam comunalidades elevadas, por outro lado Confab PN, Perdigão PN e Avipal ON apresentam baixas comunalidades.

Com a finalidade de comparar os resultados obtidos, utilizando o modelo de fatores macroeconômicos, além das comunalidades, a Tabela 5 apresenta os coeficientes de determinação do modelo, utilizando o mesmo procedimento apresentado por Schor, Bonomo e Valls (2002), nota-se que as comunalidades são maiores que os coeficientes de determinação para todos os ativos exceto para a Petrobras PN. Este resultado indica que, para estes ativos, os fatores estatísticos explicam maior parte da variabilidade dos retornos do que os fatores macroeconômicos.

Tabela 5 - Comunalidades e coeficientes de determinação dos modelos

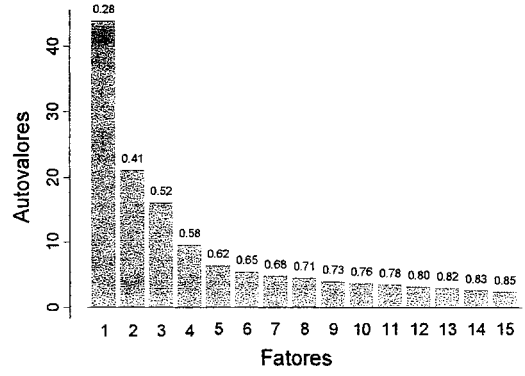
Ativo	Comunalidade (%)	R^2 (%)	Ativo	Comunalidade (%)	R^2 (%)
Acesita PN	59,68	25,91	**Ambev PN**	48,40	47,74
Aracruz PNB	85,91	77,70	**Arcelor ON**	40,50	9,79
Avipal ON	38,42	25,91	**Brasil Telecom PN**	58,43	56,31
Confab PN	29,48	25,37	**Duratex PN**	45,60	23,33
Gerdau Met PN	53,77	27,75	**Klabin PN**	56,77	28,37
Perdigão PN	34,24	21,47	**Petrobrás PN**	66,65	71,53
Ripasa PN	64,76	56,21	**Sadia PN**	62,49	35,59
Sid Tubarão PN	56,67	39,44	**Telemar PNB**	58,32	49,47
Telesp ON	45,29	42,58	**Usiminas PNA**	58,00	37,81
Vale PNA	72,15	56,40	**Votorantim PN**	61,64	41,74

4.2. Parte 2

Análise assintótica de componentes principais

Foi ajustado o modelo, utilizando o procedimento refinado de estimação dos fatores. A Figura 3 apresenta os autovalores dos 15 primeiros componentes e a proporção acumulada da variabilidade total dos dados explicado pelos componentes.

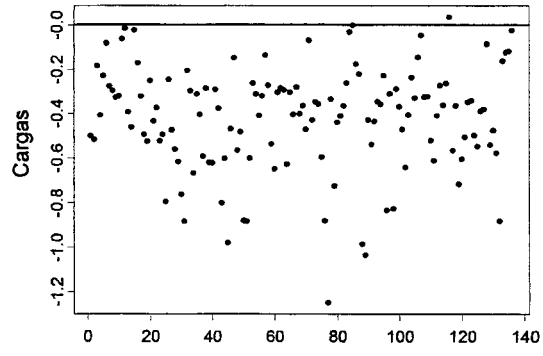
Figura 3 - *Scree-plot* – Análise assintótica de componentes principais



Foi utilizado o teste proposto por Connor e Korajczyk (1993) para verificar o número de fatores adequado para o modelo, esse teste indica que deve ser utilizado apenas um fator.

A Figura 4 apresenta as cargas dos ativos para o modelo com um único fator. Verifica-se que quase todas as cargas apresentam valores com mesmo sinal, logo este fator pode ser interpretado como um índice de mercado. Foi calculada a correlação entre as realizações do fator e o índice Bovespa no período e foi encontrado o valor -0,90, a correlação forte reforça a interpretação de que se trata de um índice de mercado.

Figura 4 - Cargas fatoriais – Análise assintótica de componentes principais



Como o número de ativos é elevado, são apresentadas na Tabela 6 algumas medidas, resumo dos coeficientes de determinação dos ajustes dos modelos. Verifica-se que, em média, o fator explica aproximadamente 32% da variabilidade dos retornos.

Tabela 6 - Algumas medidas resumo dos coeficientes de determinação

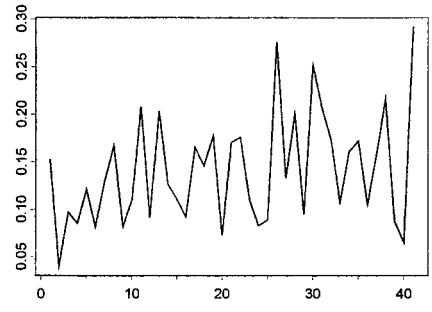
Número de ativos	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média
136	0,0374	0,1662	0,3212	0,4629	0,6812	0,3198

Modelo "BARRA"- Fatores dos tipos de indústrias

Nesse estudo, os tipos de indústrias correspondem à classificação das empresas disponibilizada pela Bovespa. As empresas da amostra selecionada estão divididas em dez tipos de indústrias: Bebidas e Alimentos, Energia Elétrica, Mineração, Papel e Celulose, Petróleo e Gás, Química, Siderurgia e Metalurgia, Telecomunicações, Têxtil e Veículos e Peças. Foram geradas então 11 variáveis de classificação, uma para cada tipo de indústria acima e a última para as indústrias do tipo "outros".

A Figura 5 apresenta os coeficientes de determinação dos modelos correspondentes aos 41 meses estudados, em média as variáveis de classificação explicam 14,13% da variabilidade dos retornos.

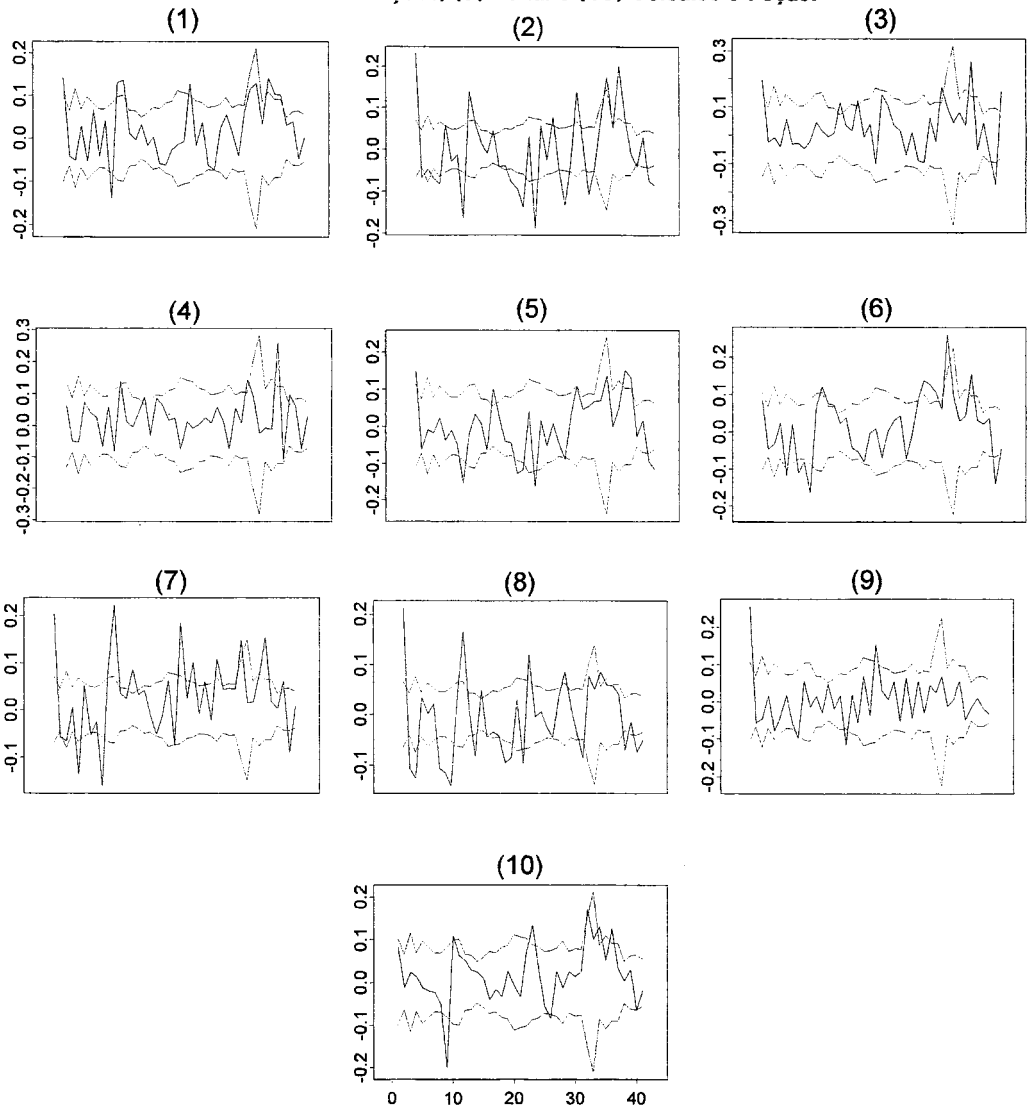
Figura 5 - Coeficientes de determinação



Na Figura 6 são apresentados os gráficos das realizações dos fatores obtidos a partir de (3) e os limites para os testes de significância[3]. Observa-se que as realizações dos fatores relacionados com as empresas de Telecomunicações (8), Energia Elétrica (2) e Siderurgia e Metalurgia (7) são significantes em um maior número de meses. Uma outra observação está relacionada com o fato de que as realizações dos fatores apresentam valores significantes positivo e negativo. Logo, considerando todo o período, não se pode afirmar que um determinado setor apresenta retornos médios maiores que os demais.

[3] Adotando um nível de significância de 5%, os limites foram calculados da seguinte forma
Limite inferior = 0 + $t_{0,025,N-2} \times$ (desvio padrão de),
Limite superior = 0 + $t_{0975,N-2} \times$ (desvio padrão de),
onde $t_{0,025,N-2}$ e $t_{0975,N-2}$ correspondem aos percentis de ordem 0,025 e 0,975 da distribuição t de student com $N-2$ graus de liberdade (nesse caso, N = 136).

Figura 6 - Fatores relacionados com os tipos de indústrias: (1) Bebidas e Alimentos, (2) Energia Elétrica, (3) Mineração, (4) Papel e Celulose, (5) Petróleo e Gás, (6) Química, (7) Siderurgia e Metalurgia, (8) Telecomunicações, (9) Têxtil e (10) Veículos e Peças.

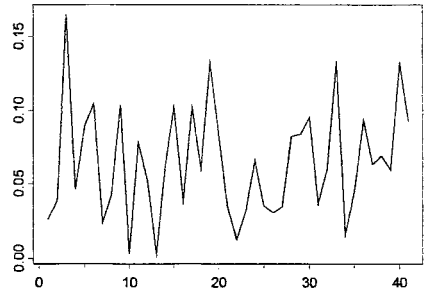


Modelo "BARRA"- Fatores contábeis

O mesmo modelo utilizado para a geração dos fatores dos tipos de indústrias foi utilizado para gerar os fatores relacionados com as variáveis: valor de mercado, valor de mercado/valor contábil, lucro/preço e alavancagem financeira. Nesse caso as cargas fatoriais β_k de (2) são substituídos pelos termos β_{kt} variantes no tempo, esses termos contêm os valores da variável k para todos os ativos no instante t.

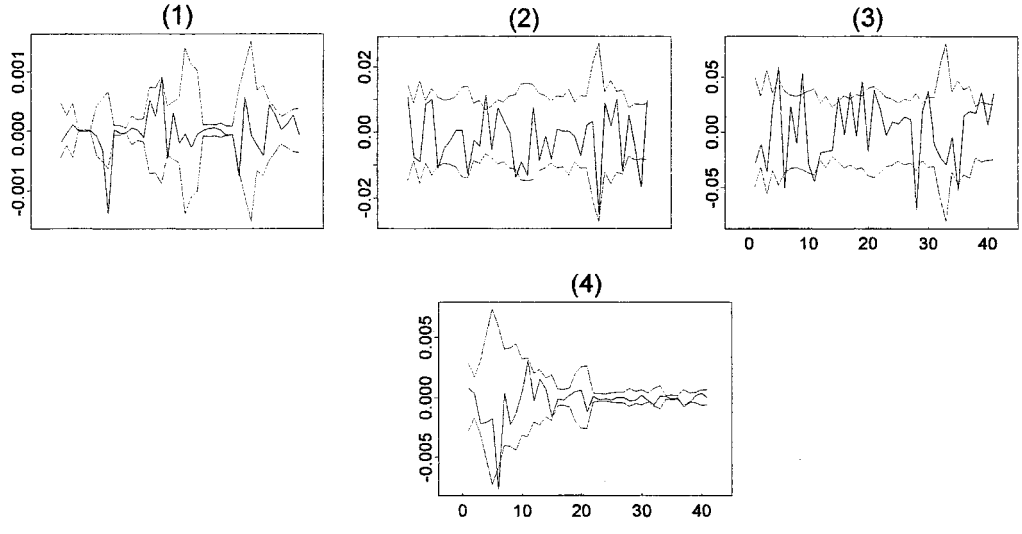
Na Figura 7, são apresentados os coeficientes de determinação do modelo para cada mês. Verifica-se que os coeficientes de determinação são baixos, na maior parte dos meses estão abaixo de 10%. Em média, 6,50% da variabilidade dos retornos é explicada pelas variáveis contábeis selecionadas.

Figura 7 - Coeficientes de determinação



Pela Figura 8 observa-se que as realizações dos fatores não são significantes na maior parte dos meses. Por esse resultado, conclui-se que as variáveis contábeis não são importantes para explicar os retornos das ações das empresas selecionadas.

Figura 8 - Fatores relacionados com as variáveis contábeis: (1) Lucro/Preço, (2) Valor de mercado, (3) Valor de mercado/Valor Contábil e (4) Alavancagem Financeira.



5. Conclusão

Os modelos de fatores estatísticos estudados foram os que apresentaram os melhores resultados. Comparando os resultados para o primeiro conjunto de dados observou-se que as comunalidades dos ativos no modelo fatorial estatístico são maiores que os coeficientes de determinação, obtidos no modelo macroeconômico, com apenas uma exceção para Petrobrás PN, logo, para estes ativos, os fatores estatísticos explicam maior parte da variabilidade dos retornos do que os fatores macroeconômicos.

Com relação ao segundo conjunto de dados, o modelo de fatores estatísticos proposto apresenta melhores resultados que o modelo fundamental, aproximadamente 32% da variabilidade dos ativos era explicada pelo fator estatístico, enquanto o modelo fundamental dos tipos de empresas explica apenas 14% da variabilidade e o modelo de fatores contábeis explica apenas 6,5% da variabilidade.

Vale ressaltar que esses modelos foram ajustados sob a suposição de que a estrutura de covariâncias é a mesma para todo período estudado, o que pode não se aplicar na prática. Em trabalhos futuros serão estudados modelos que consideram mudanças na estrutura de covariâncias, isto é, $\mathbf{Cov}(\varepsilon_{it}; \varepsilon_{jt}) = \sigma_{ijt}$ no lugar de σ_{ij} constante para todo t. Além disso, em grande parte dos resultados apresentados foi feita a suposição de normalidade dos dados, em trabalhos futuros pretende-se estudar o impacto da falta de normalidade nos resultados.

Referências bibliográficas

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

Connor, G. e Korajczyk, R.A. (1986). Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis. *Journal of Financial Economics*, 15, 373-394.

Connor, G. e Korajczyk, R.A. (1988). Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics*, 21, 255-289.

Connor, G. e Korajczyk, R.A. (1993). A Test for the Number of Factors in an Approximate Factor Model. *Journal of Finance*, 48(4), 1263-1292.

Efron, B. e Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Grinold, R.C. e Kahn, R.N. (2000). *Active Portfolio Theory and Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*, Second Edition. McGraw-Hill, New York.

Johnson, R.A. e Wichern, D.W. (1998). *Multivariate Statistical Analysis*. Prentice-Hall, Nova York.

Lambert, Z.V., Wildt, A.R. e Durand, R.M. (1990). Assessing Sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement*, 50, 30-49.

Ross S. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13, 341-360.

Schor, A. Bonomo, M. e Valls Pereira, P.L. (2002). Arbitrage Pricing Theory (APT) e Variáveis Macroeconômicas: um estudo empírico para o mercado acionário brasileiro. *Finanças Aplicada ao Brasil*, capítulo 4, 46-62, Fundação o Getúlio Vargas Editora, Rio de Janeiro.

Vinha, L. G. A. (2006). Modelos fatoriais para retorno de ativos. *Dissertação de Mestrado*, São Paulo, IME – USP.

Apêndice 1

Para demonstrar o resultado, o modelo fatorial é apresentado da seguinte forma

$$R^N = B^N F + E$$

onde

- R^N é a matriz $(N \times T)$ dos retornos observados dos N ativos para os T períodos;

- B^N é a matriz $(N \times K)$ das cargas dos fatores;

- F é a matriz $(K \times T)$ das realizações (não observadas) dos K fatores para os T períodos; e

- E^N é a matriz $(N \times T)$ dos erros.

A partir da formulação acima são definidas as matrizes

$$\begin{aligned} \Omega^T &= (1/N)\, R^{N\prime} R^N, \\ A^N &= (1/N)\, F' B^{N\prime} B^N F, \\ Y^N &= (1/N)\, (F' B^{N\prime} E^N + E^{N\prime} B^N F), \\ Z^N &= (1/N)\, E^{N\prime} E^N. \end{aligned} \tag{4}$$

A matriz Ω^T é observável enquanto as demais são não observáveis. Todas as matrizes são quadradas de dimensão T tal que $\Omega^T = A^N + Y^N + Z^N$. Considere o seguinte modelo de regressão

$$R_i = \alpha_i 1 + F'\beta_i + \varepsilon_i, \tag{5}$$

onde

- R_i é um vetor $(T \times 1)$ dos retornos observados do ativo i;

- α_i é o intercepto da equação;

- 1 é um vetor $(T \times 1)$ de uns;

- F é a matriz $(K \times T)$ das realizações (não observadas) dos fatores;

- β_i é o vetor $(K \times 1)$ dos betas dos fatores para o ativo i; e

- ε_i é o vetor $(T \times 1)$ dos resíduos.

Como as realizações dos fatores não são observáveis, esse modelo não pode ser ajustado. A proposta é mostrar que os autovetores da matriz Ω^T se aproximam desses fatores quando o número de ativos é grande. A seguir são apresentados alguns resultados importantes para a demonstração.

Em geral, denomina-se matriz de componentes principais a matriz $(k \times T)$ ortonormal dos k autovetores correspondentes aos k maiores autovalores de uma matriz W qualquer, simétrica e positiva semidefinida.

Define-se para a demonstração G^N e H^N as matrizes de componentes principais de Ω^T e A^N, respectivamente. As matrizes Λ_G e Λ_H são matrizes diagonais com os autovalores correspondentes. Pode-se demonstrar que H^N é uma transformação não-singular de F, ou seja, existe uma matriz não singular L^N tal que $L^N \mathbf{F} = H^N$.

Lema 1. Para qualquer N, existe uma matriz (k×k) não-singular L^N tal que $L^N F = H^N$.

Prova, ver Vinha (2006) ou Connor e Korajczyk (1986).

Supondo que a matriz H^N fosse observável e utilizando-a no lugar de F na equação (5), seriam obtidas estimativas do intercepto e dos erros $\alpha_{i,H}$ e $\varepsilon_{i,H}$. As estimativas dessa regressão hipotética são as mesmas estimativas que seriam encontradas utilizando a matriz F. Portanto, a substituição de H^N no lugar de F não tem efeito nessas estimativas, isso se deve ao fato que a matriz L^N desaparece com a utilização do método de mínimos quadrados.

Lema 2. Se H^N é uma transformação não-singular de F, a utilização de H^N no lugar de F não tem efeito nas estimativas de α_i e ε_i para o modelo (5).

Prova, ver Vinha (2006) ou Connor e Korajczyk (1986).

O próximo passo da demonstração é provar que a matriz G^N é aproximadamente igual a uma transformação não-singular de F, ou seja, G^N converge para H^N quando N tende ao infinito.

Pode-se observar que a matriz Y^N de (4) converge para uma matriz de zeros quando N cresce. A matriz Z^N representa a matriz de covariâncias dos resíduos, os termos da diagonal principal e os termos fora da diagonal de Z^N são dados por

$$(1/N)\sum_{i=1}^{N}\varepsilon_{it}^2 \;\; e \;\; (1/N)\sum_{i=1}^{N}\varepsilon_{it}\varepsilon_{is} \quad para \;\; s,t=1,...,T \;\; e \;\; s\neq t.$$

Sob a suposição que os termos ε_{it}^2 convergem para σ^2 (existe uma média das variâncias dos erros) e que os termos $\varepsilon_{it}\varepsilon_{is}$ convergem para zero (não existe correlação intertemporal entre os erros), logo a matriz Z^N tende a $\sigma^2 I_T$.

Como resultado tem-se que, quando N cresce, a matriz N se aproxima de $A^N + \sigma^2 I_T$. Além disso, os autovetores de uma matriz não se alteram com a adição de uma matriz escalar, logo a matriz G^N converge para H^N.

Uma vez que H^N é uma transformação singular de F, a matriz G^N é aproximadamente uma transformação singular de F. Portanto, quando N é grande, os autovetores da matriz dos retornos são aproximadamente equivalentes aos fatores da análise fatorial.

Apêndice 2

São apresentados a seguir os retornos dos ativos dos dois conjuntos de dados selecionados.

Figura A1 - Retornos dos ativos – Conjunto de dados 1

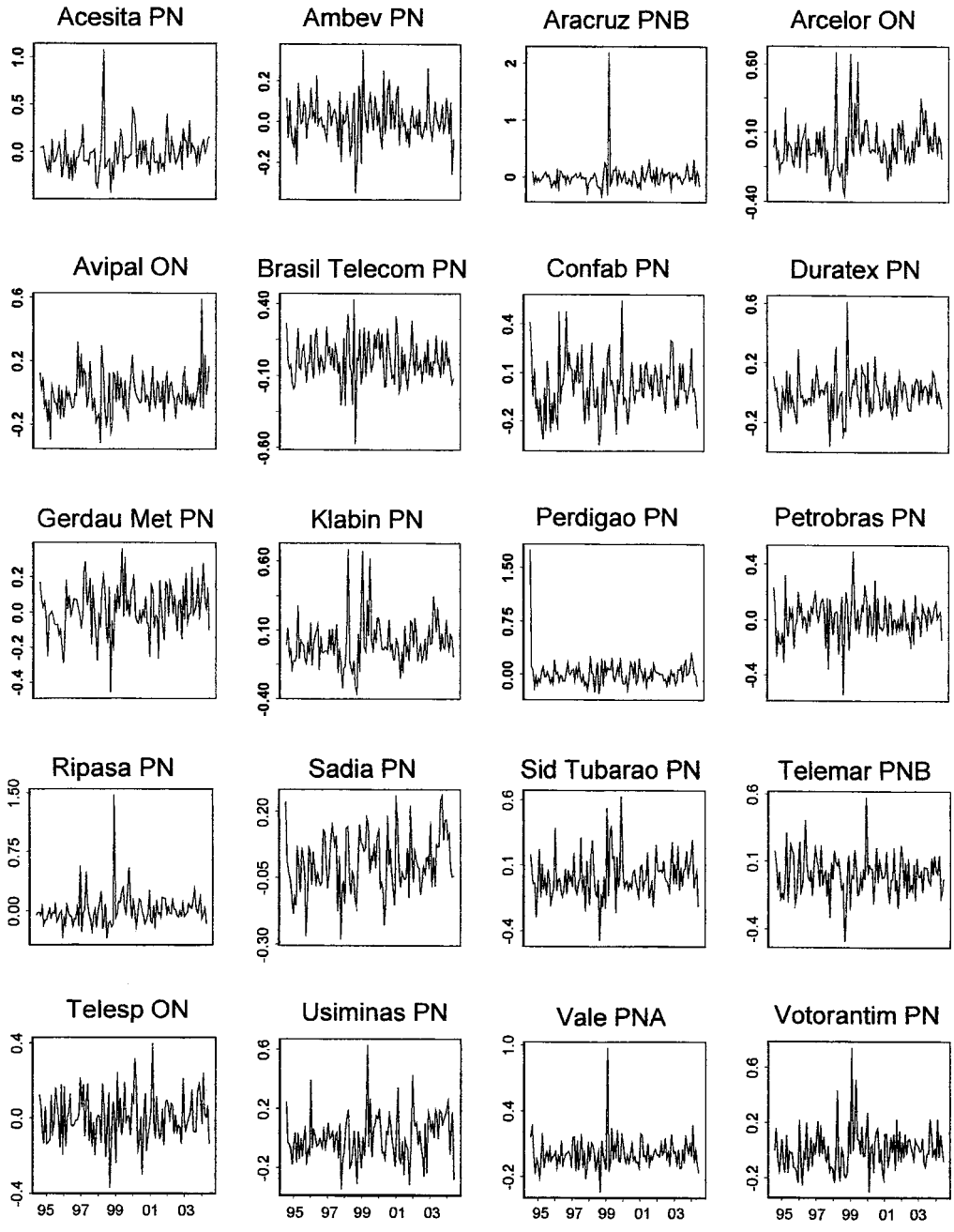


Figura A2 - Retornos dos ativos – Conjunto de dados 2

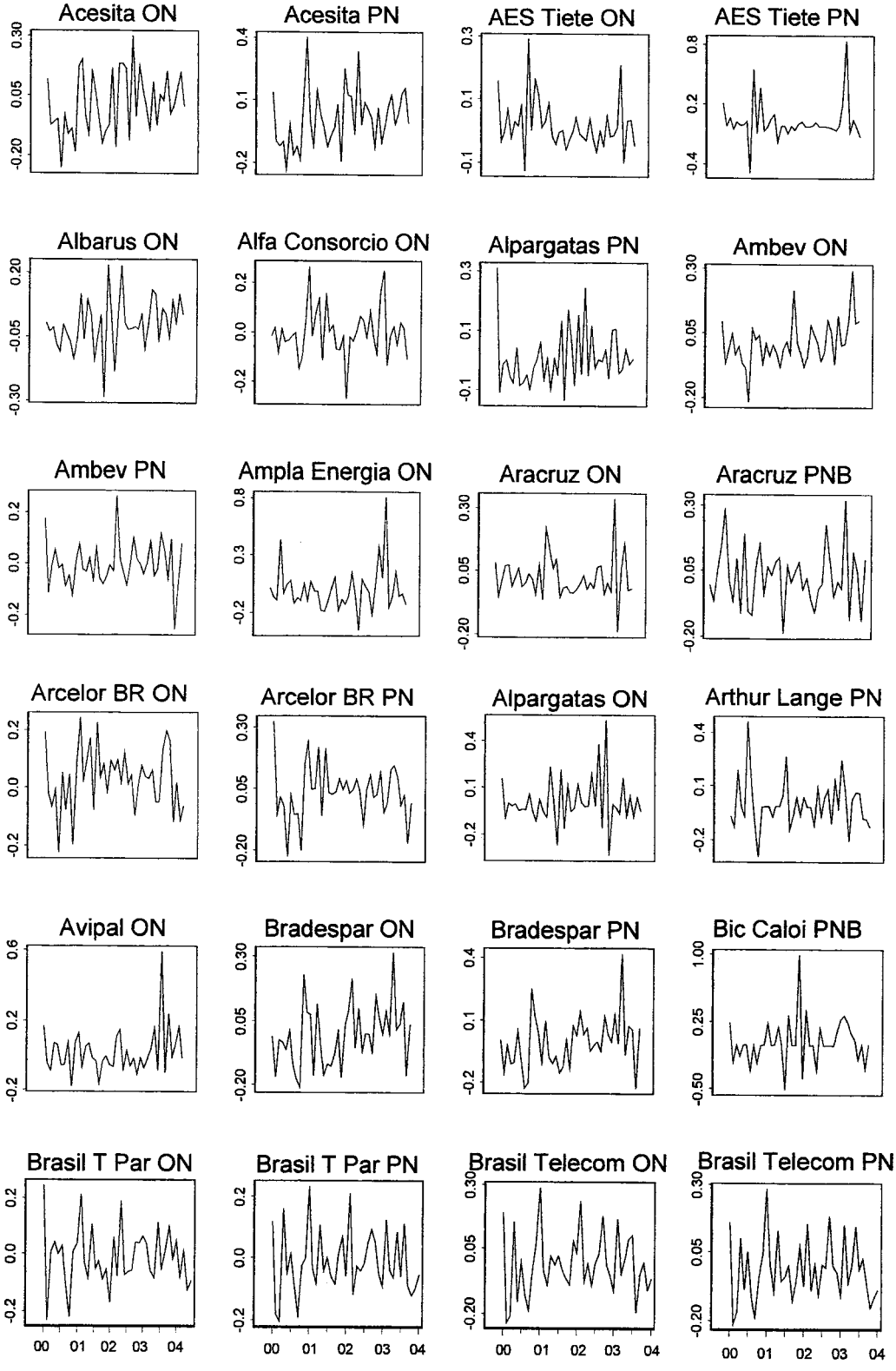


Figura A3 - Retornos dos ativos – Conjunto de dados 2

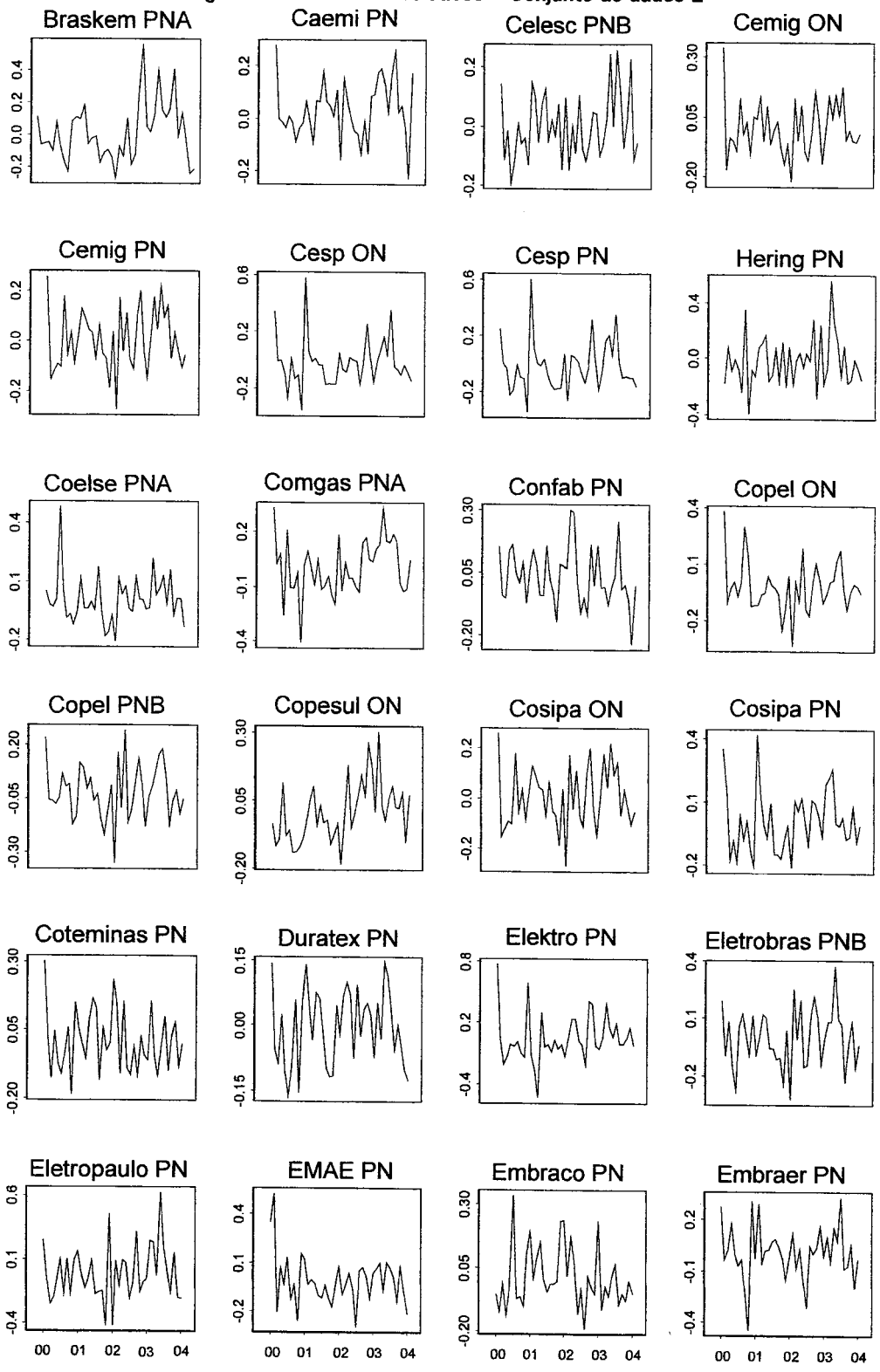


Figura A4 - Retornos dos ativos – Conjunto de dados 2

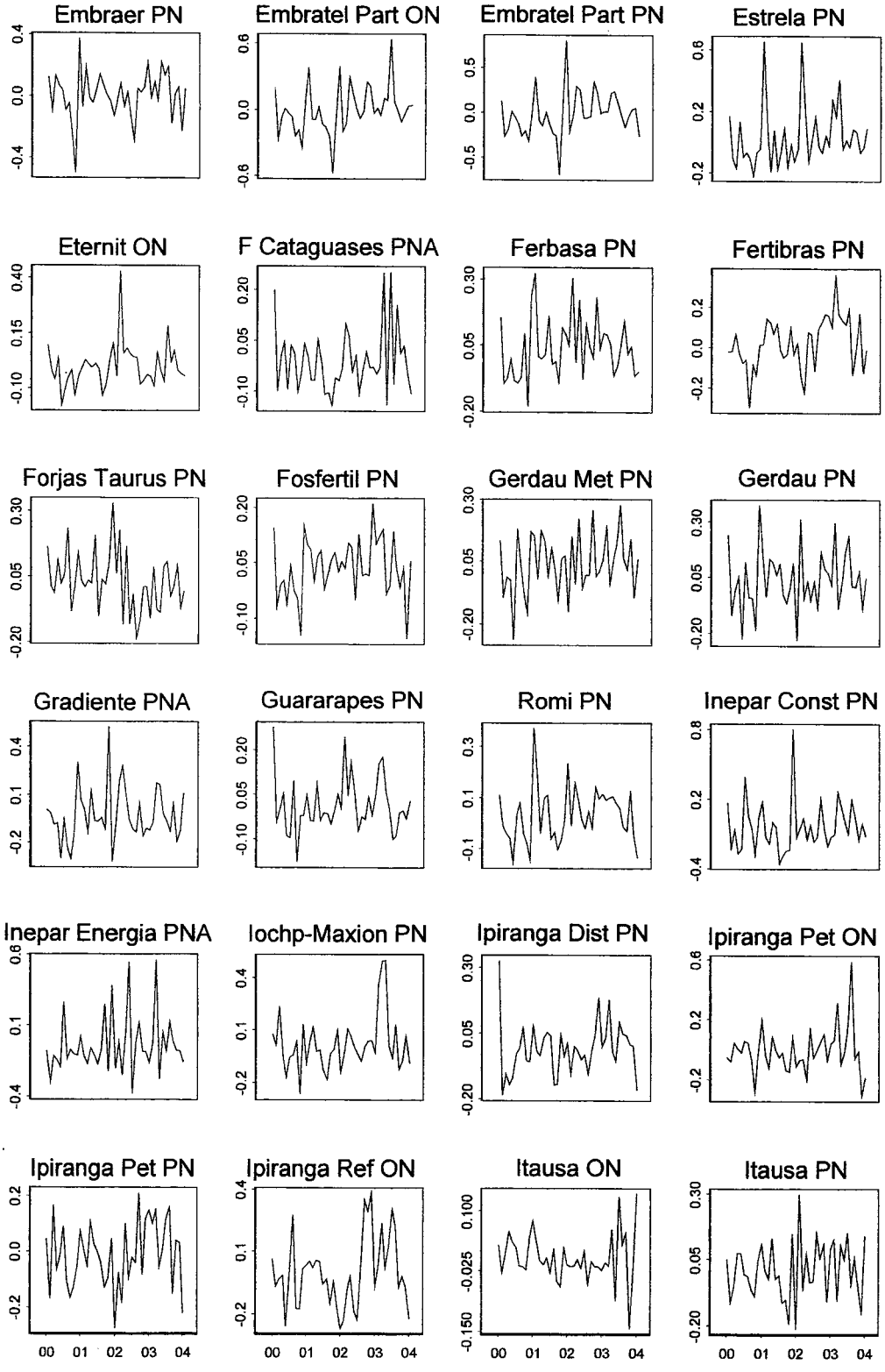


Figura A5 - Retornos dos ativos – Conjunto de dados 2

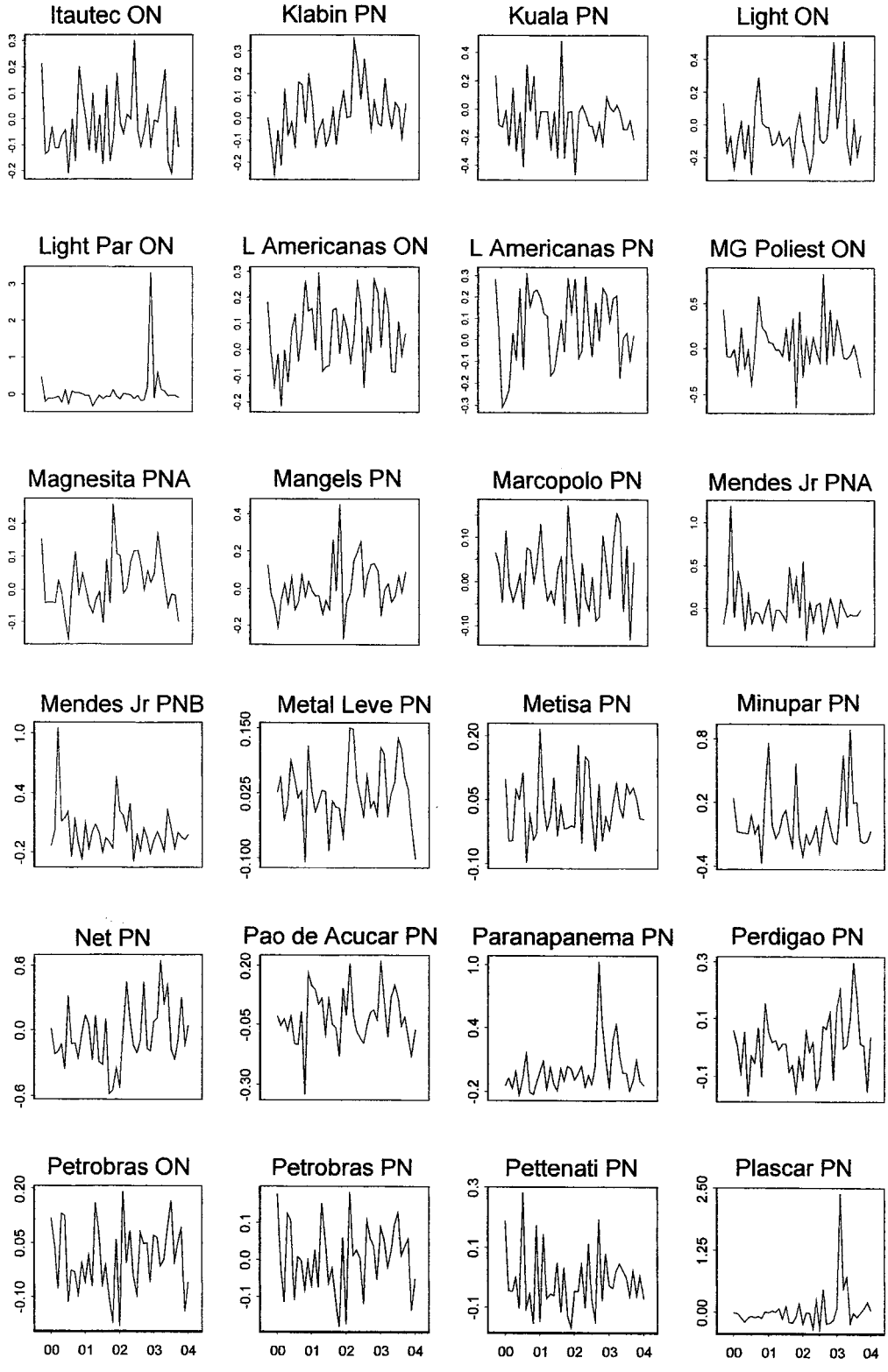


Figura A6 - Retornos dos ativos – Conjunto de dados 2

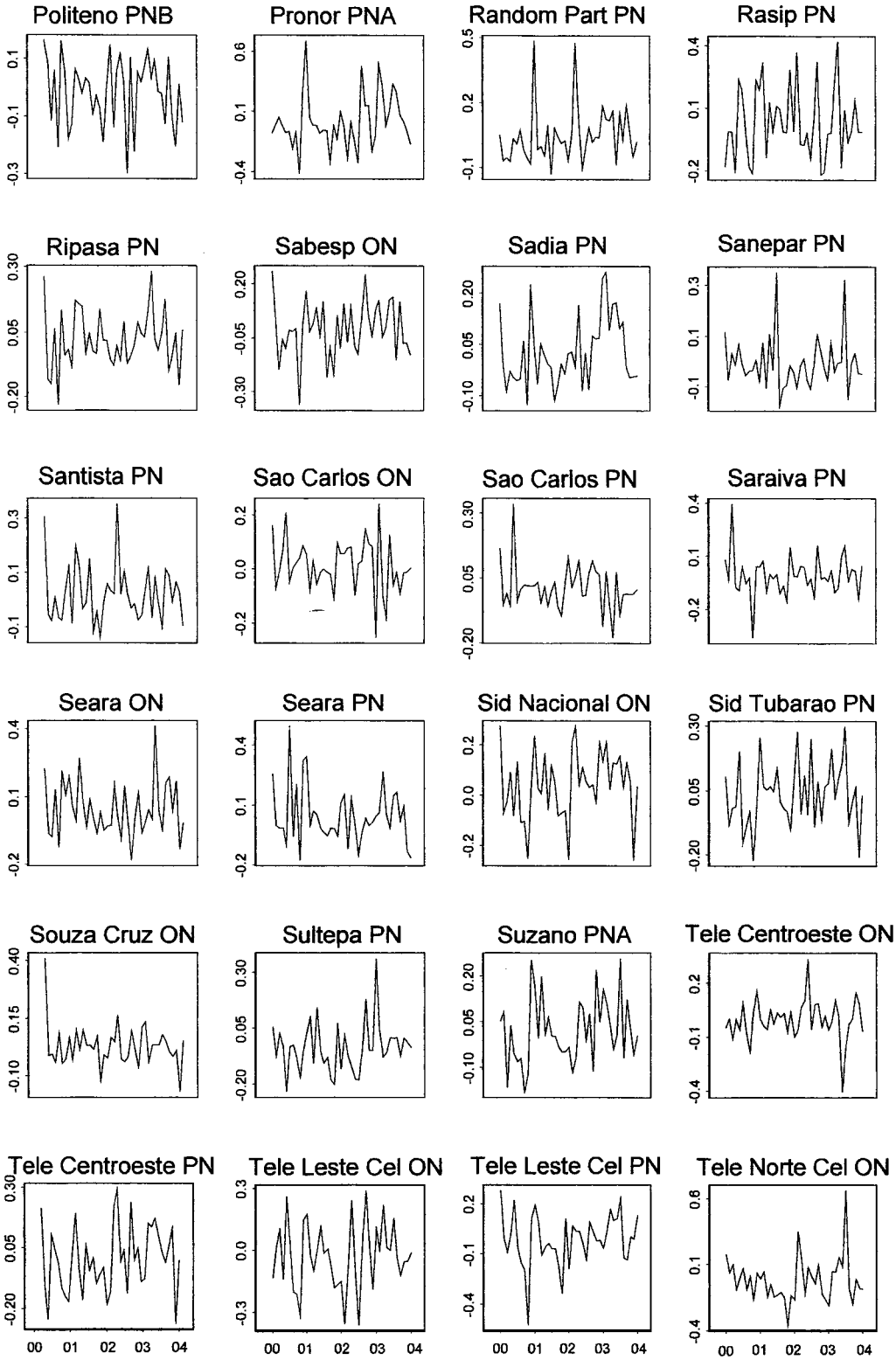
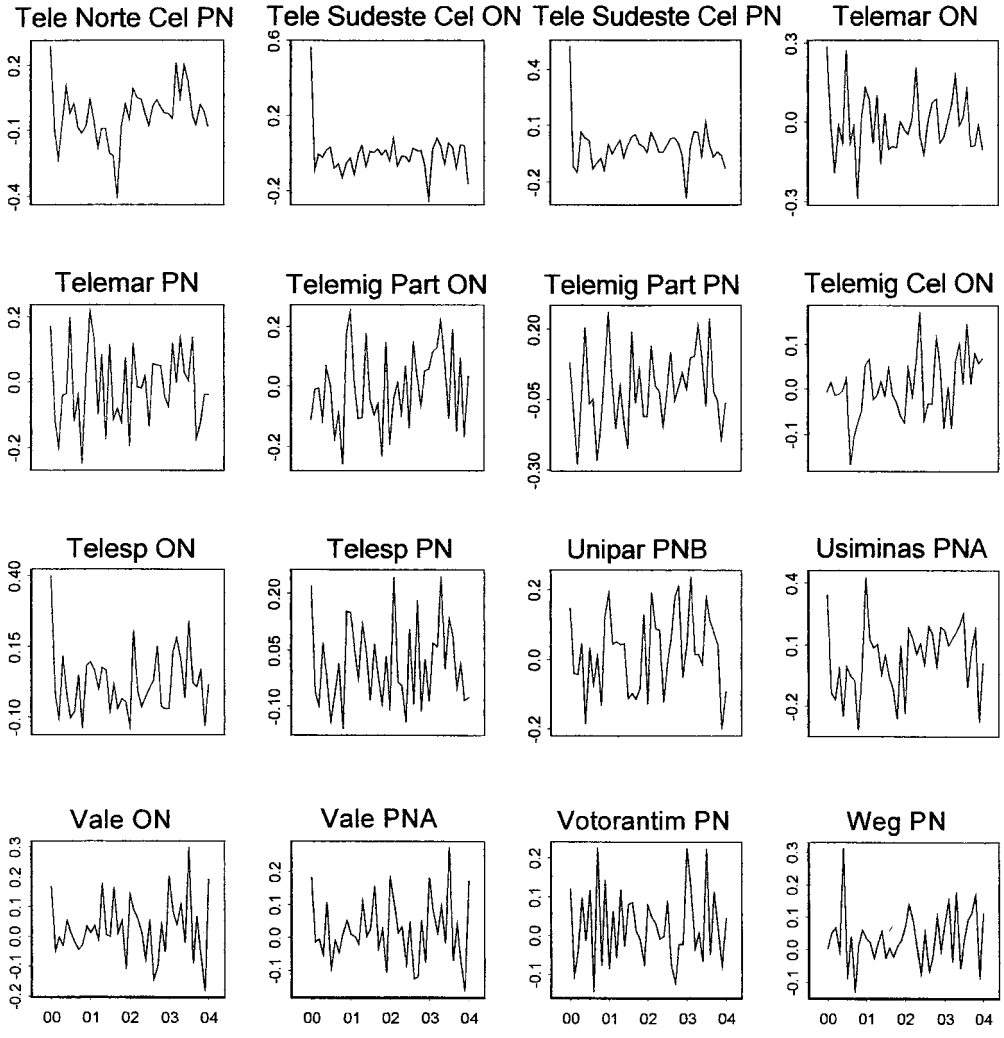


Figura A7 - Retornos dos ativos – Conjunto de dados 2



Abstract

The covariance structure among the asset returns represents a fundamental information for the investments portfolio formation. The responsible factors for the assets systematic risk can be identified by factor models, what make possible a robust estimation of the covariances among returns. The objective of the present study is to identify fundamental and statistical factors that influence the asset returns of the Brazilian market. The generation of the statistical factors and the study of the influence of these factors in the returns were made using traditional factorial analysis and asymptotic principal component analysis. Cross-section regressions models were used in the study of the fundamental factors importance. Among the studied models, the statistical factors models presented the better results.

Estimação a partir de amostras periódicas não sobrepostas: a experiência francesa

Andréa Diniz da Silva[*]
Jean-Michel Durr[**]

Resumo

A necessidade cada vez mais premente de produzir informações detalhadas em nível geográfico, e com maior freqüência que as produzidas a partir de censos qüinqüenais ou decenais; e ainda o aumento das dificuldades para garantir recursos financeiros para a realização dos censos nacionais, têm motivado a busca de alternativas metodológicas para o atendimento da demanda por informações, tradicionalmente atendida pelos Censos Demográficos. Uma das alternativas ao modelo corrente de enumeração completa da população é o uso de amostras periódicas não sobrepostas proposto pelo estatístico húngaro Leslie Kish. A base do modelo proposto por Kish é a acumulação de amostras ao longo do espaço e/ou do tempo. Dentre as experiências em curso, a francesa é a que se encontra em estágio mais adiantado na implementação da alternativa proposta por Kish, portanto, o presente artigo traz aspectos relacionados com plano amostral e estimação no Novo Censo Francês.

[*] Endereços para correspondências: Instituto Brasileiro de Geografia e Estatística – IBGE – *e-mail*: adiniz@ibge.gov.br.
[**] United Nations Statistics Division – *e-mail:* durr@un.org.

1. Introdução

Desde o final da década de 1970, o estatístico húngaro Leslie Kish vinha tecendo considerações sobre a utilização de métodos alternativos ao censo por enumeração completa da população de um país ou região, para a produção de informações detalhadas que descrevam suas características e estrutura. Na década de 1980, Kish (Kish; Verma, 1986) defendeu o uso de informações provenientes de registros administrativos e de pesquisas por amostragem como fontes complementares ao censo. Neste contexto, ressaltou que não se tratava da idéia de utilizar um dos métodos em substituição ao outro, mas de combinar os métodos. Mais tarde, Kish(Kish, 1988) fez uma crítica severa, reconhecendo que os censos por enumeração completa representam mais uma tradição ou força legal que propriamente a única alternativa metodológica para obtenção de informações detalhadas sobre a população. Contudo, enfatizou que não podia ser imediata a substituição do modelo tradicional e que a adoção da nova metodologia necessitava do desenvolvimento de diferentes técnicas para diferentes realidades.

A partir da experiência concreta de países escandinavos com o uso de registros administrativos para prover informações detalhadas sobre a população, em substituição ao modelo tradicional de censo, ainda na década de 1990, as Nações Unidas, no texto das Recomendações para os censos da rodada de 2000 na Região da Comissão Econômica Européia, deixam de considerar a enumeração completa da população a única maneira de se obter dados para um censo. As modalidades de coleta de dados mencionadas no documento são a tradicional, usando questionário, o uso de registros e outras fontes administrativas e a combinação de registros e outras fontes administrativas com pesquisas (enumeração completa ou amostra). Ainda durante esta década, a França e os Estados Unidos, realizaram uma série de estudos para a realização de pesquisas, utilizando amostras periódicas não sobrepostas para prover informações detalhadas sobre a população, baseando-se nos estudos desenvolvidos por Leslie Kish (Kish, 1979a, 1979b, 1981, 1983, 1990, 1997 e 1998; e Kish; Verma, 1983 e 1986). Já nos anos de 2000, o crescente número de países europeus em via de

implementar alguma modalidade alternativa à enumeração completa da população[1], trouxe a necessidade da ampliação do conceito de censo, o que foi feito no texto das Recomendações para os Censos da rodada de 2010, na Região da Comissão Econômica Européia. No documento, destaca-se que a definição de censo para a região passa a ser baseada no resultado produzido pela operação, mais que na metodologia utilizada. Neste sentido, considera-se censo a operação que produz, com intervalo regular, contagens da população de um país, para pequenos níveis geográficos, assim como um conjunto de características social e demográfica para o total da população[2].

As experiências já em curso destacam vantagens associadas a pelo menos três demandas mais emergentes em diversos países ao redor do mundo: redução de custo da operação, redução da participação dos informantes e a produção de informações com maior freqüência (Kamen, 2004. Nordholt, 2004 e Szenzenstein, 2004). O atendimento das demandas emergentes é possível em qualquer das modalidades alternativas, quer seja a baseada em registros administrativos, em amostras periódicas, ou na conjugação de ambas, contudo dentre as experiências apresentadas, apenas a França e os Estados Unidos apresentaram a pretensão de produzir informações detalhadas anualmente, ainda que os americanos não pretendam prescindir da enumeração completa da população a cada dez anos (Griffin, 2003). Holanda, Alemanha e Israel mencionaram apenas a intenção de substituir o censo no modelo tradicional, destacando somente as vantagens associadas à redução de custo e da participação do informante.

A necessidade da produção de informações detalhadas, com maior freqüência que as atualmente produzidas, é uma motivação contundente para que seja avaliada a utilização de modalidade alternativa ao censo tradicional. Atualmente, na maioria dos países, as informações para os níveis mais desagregados, para os mais diversos usos, são produzidas a cada dez anos quando da realização do Censo Demográfico. Durante o período intercensitário, as várias aplicações baseadas no tamanho ou nas características da população, são feitas com base em estimativas. Sendo assim, a distribuição de

---

[1] Dentre os 44 países investigados pela UNECE, 35 realizaram censo tradicional na rodada de 2000. Destes, somente 22 pretendiam utilizar a mesma metodologia na rodada de 2010. Os demais pretendiam utilizar alguma modalidade alternativa. Somente um país não informou qual modalidade usará em 2010.

[2] United Nations, Recommendations for the 2010 censuses of population and housing in the ECE Region, p. 8.

recursos financeiros públicos e a elaboração de políticas públicas nas áreas de educação, saúde e transporte, por exemplo, são feitas com base em estimativas que podem estar defasadas em até nove anos.

Segundo Kish (Kish; Verma, 1986), no entanto, a escolha entre censo por enumeração completa da população, censo com base em registros administrativos ou com base em pesquisas por amostra, envolve, entre outros aspectos, a definição de estratégias que considerem as situações nacionais específicas, relacionadas com a disponibilidade de fontes complementares, com a natureza da informação requerida e ainda com a periodicidade desejada. Neste contexto, ficou o desafio de desenvolver métodos de estimação que garantam o atendimento da demanda por informações, tradicionalmente atendida pela enumeração completa da população, a partir do modelo alternativo de censo.

No enfrentamento deste desafio, o *Institut National de la Statistique et des Études Économiques* – INSEE vem realizando estimativas, levando em consideração não somente os dados coletados a cada ano e a conjugação de dados coletados em anos consecutivos, mas também aqueles disponíveis em fontes administrativas. Considerando ser esta a experiência mais adiantada em sua implementação, pretende-se trazer no presente artigo aspectos relacionados com a estimação no novo censo francês.

2. O Novo Censo Francês

Desde 1801, a França vem realizando censo por enumeração completa da população. A pretensão de realização do censo a cada cinco anos se cumpriu até a Segunda Guerra Mundial. A partir de 1946, ocasião em que foi criado o *Institut National de la Statistique et des Études Économiques* – INSEE, a periodicidade do censo se tornou irregular, variando entre seis e nove anos. O aumento da periodicidade do censo foi resultante, entre outros fatores, da baixa popularidade da operação junto à população, que passou a questionar o alto investimento para a obtenção de informações cuja disponibilidade em um número cada vez maior de fontes em registros administrativos, era crescente (DURR, 2004).

Fatores associados à descentralização politico-administrativa em curso nos últimos 20 anos, que levou ao aumento da responsabilidade das municipalidades e à mudança do perfil da demanda que passou a estar mais voltada para informações estatísticas que pudessem subsidiar a elaboração de políticas públicas locais, portanto mais específicas, detalhadas e, principalmente, oportunas, contribuíram para que emergisse a necessidade de reformulação do modelo de censo até então realizado.

Para que fosse possível atender tanto às antigas demandas de diminuição do custo da operação censitária, quanto às novas de produção de informações mais detalhadas, tanto geograficamente quanto do ponto de vista temático, além de oportunas, foi adotado a partir de 2004, modelo de censo baseado em amostras periódicas não sobrepostas em substituição ao modelo de enumeração completa da população.

A partir do novo desenho do censo francês, as estimativas de população para todos os níveis geográficos, incluindo os mais desagregados como é o caso dos pequenos municípios, poderão ser obtidas anualmente a partir do quinto ano de coleta. Ao completar este primeiro ciclo, será possível acumular as cinco primeiras amostras. No ano seguinte, serão acumuladas as cinco amostras anteriores, ou seja, as amostras feitas entre o segundo e o sexto ano, e assim sucessivamente. Contudo, já a partir do primeiro ano (2004), foram obtidos os totais nacional, regional e para os 100 maiores municípios[3]. Para os 200 maiores municípios[4], os totais foram obtidos acumulando-se as amostras do primeiro e segundo anos (2004 e 2005). A partir do final do terceiro ano, pretende-se produzir estimativas para todos os municípios acima de 10 000 habitantes, acumulando-se as três primeiras amostras anuais.

Com o redesenho do plano de coleta, o custo estimado da operação total, prevista para ser realizada em cinco anos, é de pouco mais de 70% do custo estimado para a enumeração completa da população realizada em um único ano, portanto menor que o da operação tradicional. A redução do custo total é conseqüência da realização da coleta por amostragem nos municípios com mais de 10 000 habitantes, uma vez que

[3] População em torno de 50 000 habitantes ou mais.
[4] População em torno de 35 000 habitantes ou mais.

para os municípios menores, os quais são visitados uma vez a cada cinco anos, os custos se mantêm.

3. Amostra

Os quase 37 000 municípios franceses estão divididos em dois grupos: "pequenos e médios", que são os municípios com menos de 10 000 habitantes; e "grandes", aqueles com 10 000 ou mais habitantes. Das cerca de 60 milhões de pessoas que vivem na França, cerca de 30 milhões vivem em municípios pequeno e médio e as demais em grandes, portanto, cada grupo inclui aproximadamente metade da população.

No total, anualmente serão recenseadas cerca de 8,5 milhões de pessoas: 6 milhões em municípios pequeno e médio e 2,5 milhões em grandes municípios.

Os municípios pequeno e médio foram divididos em cinco grupos de rotação, balanceados[5] segundo as seguintes características: número de domicílios, número de domicílios coletivos, população dos departamentos[6], sexo e grupo de idade (menos de 20, 20-39, 40-59, 60-74 e 75 ou mais). Os dados utilizados para o balanceamento foram extraídos do Censo de 1999. A cada ano os municípios de um dos grupos de rotação serão recenseados, ou seja, todos os domicílios e pessoas destes municípios serão contados e caracterizados. Ao final de cinco anos todos os municípios pequenos e médios terão sido recenseados.

Todos os municípios grandes serão visitados anualmente, porém estes serão amostrados. Serão constituídos cinco grupos de rotação de endereços a partir de um cadastro de endereços *Répertoire d'immeubles localisés* - RIL, balanceados segundo os mesmos critérios utilizados para os pequenos e médios municípios. A cada ano será extraída amostra de 8% dos endereços pertencentes ao grupo do ano. Ao final de cinco anos serão pesquisados 40% dos endereços em municípios grandes.

Concluída a operação, ao fim de cinco anos, portanto, em 2008, terão sido visitados 70% dos endereços do país. Deste total, 50% se localizam em municípios com

---

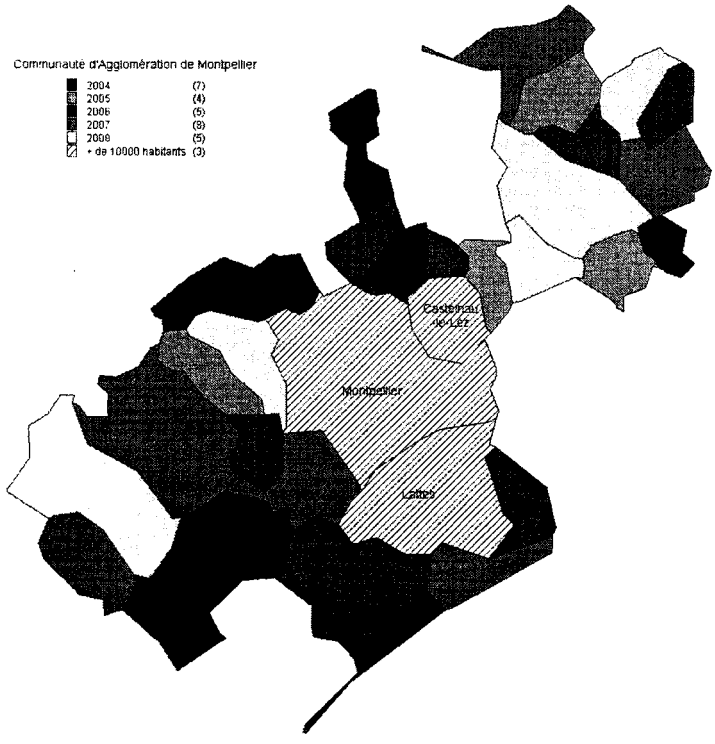
[5] A técnica do balanceamento foi desenvolvido por Deville e Tillé. Ver Deville, J. e Tillé, Y. 2000.

[6] A França é dividida segundo a seguinte hierarquia administrativa: regiões, departamentos, distritos, cantões e communes (municípios).

menos de 10 000 habitantes, nos quais o levantamento terá sido feito por enumeração completa da população municipal; e 20% nos municípios maiores, onde terá sido realizada amostra de endereços.

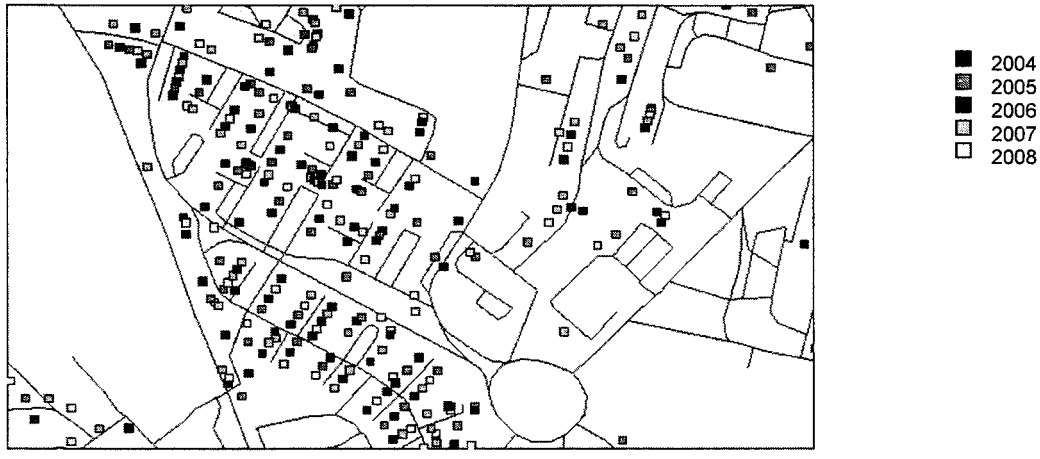
A Figura 1 mostra a estrutura de cobertura, nos primeiros cinco anos, para a área urbana de Montpellier, que tem quase 380 000 habitantes. A área é composta de três municípios com mais de 10 000 habitantes e 29 pequenos, portanto com tamanho populacional inferior. As áreas achuradas são os municípios grandes, portanto será amostrada todo ano. Nas demais áreas, cada cor representa um grupo de rotação, portanto um ano de coleta por enumeração completa da população.

Figura1- Estrutura do Novo Censo para a área urbana de Montpellier



Na Figura 2 é apresentada a distribuição das amostras para os cinco primeiros anos segundo os grupos de rotação.

Figura 2 - Cinco grupos de endereços



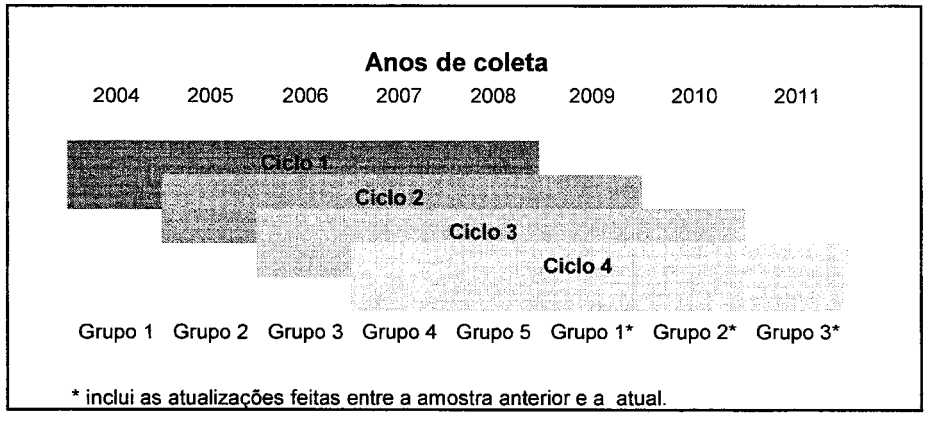
A partir do sexto ano, novas amostras serão extraídas dos grupos de rotação correspondentes a cada ano do ciclo de acumulação. Sendo assim, no sexto ano será extraída amostra do grupo correspondente ao primeiro ano, portanto grupo 1, no sétimo do grupo 2 e assim sucessivamente. Contudo, os grupos de rotação não são exatamente os mesmos uma vez que o cadastro de endereços é constantemente atualizado. Sendo assim, o grupo do sexto ano toma por base o grupo de rotação do primeiro ano, contudo incorporando as atualizações feitas nos cinco anos anteriores. Tais atualizações consistem basicamente na exclusão de endereços extintos e inclusão de novos endereços, mas considera a possibilidade de incorporação de endereços oriundos de municípios médios cuja população tenha crescido de modo a alcançar 10 000 habitantes no período entre a última e a próxima coleta no grupo.

Para efeito de acumulação das amostras, os ciclos têm duração de cinco anos. O primeiro ciclo de acumulação se iniciou em 2004 e estará completo em 2008. Portanto, os primeiros resultados gerais tomarão por base os dados coletados no período de 2004 a 2008. O segundo ciclo de acumulação tomará por base os dados coletados no período de 2005 a 2009. Desta forma, cada ciclo é formado tomando as informações dos cinco

anos imediatamente anteriores o que significa que a cada ano descarta-se o ano mais antigo e agrega-se os dados coletados no ano em curso. No ano em curso os grupos de rotação, tanto para os municípios pequeno e médio quanto para a extração da amostra nos municípios grandes, são os mesmos[7] utilizados no ano que está sendo descartado, de modo a evitar distorções quando da acumulação das amostras e agregação dos dados coletados nos municípios pequeno e médio.

A Figura 3 relaciona os ciclos de acumulação com o ano de coleta e o grupo de rotação. A figura mostra que a partir do segundo ciclo de acumulação os dados coletados no primeiro ano do ciclo anterior são substituídos por aqueles coletados no último ano ciclo corrente. Nota-se ainda que o grupo de rotação de ambos os anos é basicamente o mesmo. Este procedimento é repetido a cada ano.

Figura 3 - Ciclos de acumulação segundo o ano de coleta e o grupo de rotação



4. Estimação

A base da metodologia é coletar informações durante um ciclo qüinqüenal para produzir, a cada ano, dados significativos para o ano do meio do ciclo. Sendo assim, a cada ano Y são produzidos dados, a partir das informações coletadas nos anos Y-4, Y-3, Y-2, Y-1 e Y, os quais são significativos para o ano Y-2.

[7] O grupo base é o mesmo, contudo sofre atualizações a cada ano.

A partir do final dos primeiros cinco anos, portanto, de 2008, três tipos de resultados serão produzidos a cada ano Y:

- População oficial de cada município, independentemente do ano da última coleta;
- Dados para todos os níveis geográficos, baseados na combinação das informações coletadas ao longo dos cinco anos anteriores, contudo significativos para o ano Y-2; e
- Dados para a França e regiões, os quais serão baseados na pesquisa do ano Y.

4.1 População oficial para o ano Y-2

A estimação da população é feita a partir da acumulação das amostras dos cinco últimos anos e é válida para o ano do meio do ciclo, portanto Y-2. O método varia de acordo com o tamanho do município (Durr; Grosbras, 200-).

Para os municípios com menos de 10.000 habitantes, a última contagem da população terá sido feita em um dos anos entre Y-4 e Y. Considerando o ano da contagem, o total populacional será interpolado ou extrapolado para Y-2. A extrapolação usará a tendência de crescimento do total de domicílios, observada no "arquivo de taxa municipal".

A regra é a seguinte:

Se o ano do censo é Y-4, a estimação da população para Y-2 é feita por extrapolação, usando a tendência do número de domicílios, observada no "arquivo de taxa municipal", no período Y-4 a Y-2.

Se o ano do censo é Y-3, a estimação da população para Y-2 é feita por extrapolação, usando a tendência do número de domicílios, observada no "arquivo de taxa municipal", no período Y-3 a Y-2.

Se o ano do censo é Y-2, é utilizado o resultado obtido com a contagem do ano.

Se o ano do censo é Y-1, a estimação da população para Y-2 é feita por interpolação, usando a estimativa para Y-3 (baseada na tendência do número de domicílios entre Y-1 e Y-3) e o valor observado na contagem em Y-1.

Se o ano do censo é Y, a estimação da população para Y-2 é feita por interpolação, usando a estimativa para Y-3 (baseada na tendência do número de domicílios entre Y e Y-3) e o valor observado na contagem em Y.

Os municípios a partir de 10 000 habitantes, são amostrados a uma fração de 8% dos endereços ao ano. A amostra anual é composta de três tipos de endereços: novos, grandes e outros. Neste caso, a população média dos últimos cinco anos pode ser estimada, por médias móveis, da seguinte maneira:

$$Pop_{Y-4...Y} = \sum_{i=Y-4}^{Y} \left[Pop(GE_i) + Pop(NE_i) + w_i Pop(OE_i) \right],$$

onde GE_i representa os grandes endereços no ano i, NE_i os novos endereços, OE_i os outros endereços e W_i é o peso amostral para os outros endereços no ano i. Como os novos e os grandes endereços são visitados exaustivamente a cada ano, os pesos associados a estes endereços é 1.

A população em *Y-2* é então estimada por:

$$Pop_{Y-2} = Pop_{Y-4...Y} \frac{Ndom_{Y-2}}{Ndom_{Y-4...Y}},$$

onde $Ndom_{Y-2}$ é o número de domicílios no ano *Y-2*, dado pelo RIL[8], e $Ndom_{Y-4...Y}$ é o número médio de domicílios dos últimos cinco anos.

Sendo assim, a população total, em *Y-2*, é estimada pela soma das populações municipais, dada por:

$$Pop_{TOT_{y-2}} = \sum_{i=1}^{n} Pop_{y-2}(Mun_i^{<1000}) + \sum_{i=1}^{m} Pop_{y-2}(Mun_i^{\geq 10.000})$$

[8] Répertoire d'immeubles localisés ou RIL é o cadastro de endereços utilizado para a seleção das amostras. Como é atualizado continuamente, o cadastro fornece uma boa estimativa do crescimento do número de domicílios.

onde

n é o total de municípios com menos de 10 000 habitantes; e

m é o total de municípios com 10 000 habitantes ou mais.

4.2. População para o ano Y

Já a partir do primeiro ano, é possível obter resultados para o ano em curso, para níveis geográficos maiores. Para os níveis nacional e regional, assim como para os cerca de 100 maiores municípios[9], a população foi estimada com base na amostra do ano de 2004. Em 2005, estes totais foram atualizados com base nas amostras, dos anos de 2004 e 2005, as quais foram acumuladas, e ainda foram feitas estimativas para os cerca de 200 maiores municípios[10]. A partir do final do terceiro ano, pretende-se produzir estimativas também para os municípios acima de 10 000 habitantes, acumulando-se as três primeiras amostrais anuais. Contudo, a partir do momento em que o primeiro ciclo for completado, estes totais serão atualizados pelas estimativas baseadas nas amostras dos cinco anos antecedentes, as quais serão acumuladas.

5. Resultados detalhados

A cada ano, será constituído um arquivo contendo as informações coletadas ao longo dos últimos cinco anos. Este arquivo conterá registros para cada domicílio e para cada pessoa amostrada durante este período. Cada registro terá seu respectivo peso a partir do qual será feita a expansão de modo que o arquivo possa ser utilizado para qualquer tabulação, em qualquer nível de desagregação.

Os registros de pessoas terão o mesmo peso que os respectivos domicílios, uma vez que a amostra é feita em função dos domicílios.

O peso é dado por:

$$w_i = \frac{1}{(fa)} x \frac{Ndom_{Y-2}}{Ndom_{Y-4...Y}},$$

[9] População em torno de 50 000 habitantes.

[10] População em torno de 35 000 habitantes.

Onde a fa (fração amostral) é 1 para os grandes e para os novos domicílios e a fração real aplicada para os outros.

O coeficiente $\frac{Ndom_{Y-2}}{Ndom_{Y-4...Y}}$ estima aproximadamente a situação em Y-2.

O peso para os registros dos municípios com menos de 10 000 habitantes é um, mas também é possível obter um coeficiente para expansão a partir da população oficial estimada.

Para uma análise da precisão possível com os tamanhos amostrais utilizados, foram simuladas várias amostras de mesmo tamanho, utilizando-se os dados do Censo de 1999, a partir das quais foram calculadas as precisões associadas ao tamanho do valor estimado.

Tamanho	Precisão[11]
>50 000	<1%
20 000 - 50 000	1,5%
10 000 - 20 000	2,0%
6 000 - 10 000	2,5%
3 000 - 6 000	3,0%
2 000 - 3 000	3,5%
1 000 - 2 000	4,5%
500 - 1 000	6,0%
250 - 500	8,0%
<250	> 8,0%

Se considerada a precisão das estimativas obtidas através do censo tradicional nos períodos intercensitários, para a maioria dos usos, este método representa um considerável ganho(Durr, 2004).

[11] A precisão é dada pelo Coeficiente de Variação.

6. Utilização dos dados

A cada ano, mais de 4 milhões de domicílios e de 8 milhões de pessoas serão amostrados. Esta amostra é representativa para a França, para as regiões e para municípios com cerca 50 000 habitantes. Os resultados da pesquisa conduzida no início do ano serão publicados após o final do ano. Além disso, os resultados para os municípios com menos de 10 000 habitantes serão publicados ao mesmo tempo.

Um censo por enumeração completa produz um retrato instantâneo da população, contudo esta fotografia fica desatualizada à medida que se afasta do ano do censo. O censo baseado em amostras periódicas não produz um filme, mas um espécie de *slide.* Os resultados detalhados representam uma média sobre os últimos cinco anos e uma visão geral estimada da situação no início do ano.

Para a maioria das aplicações demográficas, com tendências moderadas, a medida obtida a partir da média dos últimos cinco anos é bastante apropriada. A vantagem está em ter o dado atualizado a cada ano, de modo que sempre que houver demanda dos usuários estas poderão ser atendidas com números recentes.

As informações para os níveis nacional e regional serão atualizadas a cada ano, sendo assim é possível observar os grandes fluxos migratórios interregionais e internacional. Além disso, a exemplo de outras, a área urbana de Montpellier (quase 380 000 habitantes) é composta por três municípios com 10 000 habitantes ou mais e 29 municípios menores. A cada ano, 45 000 pessoas, 12% da área urbana, serão amostradas. Esta amostra periódica manterá atualizadas as informações sobre o fluxo das pessoas que viajam diariamente a trabalho (movimento pendular).

Os dados econômicos requerem tratamento especial. Em pequenos municípios, as taxas de desemprego podem ser muito sensíveis à situação observada no ano em que o município foi visitado. A melhor forma de comparar as taxas municipais é utilizar pesquisas anuais: taxa de desemprego no município, dividida pela taxa de desemprego da região ou do país na mesma data. Assim, é possível comparar as taxas de desemprego em diferentes municípios.

Os municípios grandes são compostos por divisões administrativas com em torno de 2 000 habitantes, chamadas IRIS e os dados do censo são utilizados para a

elaboração de políticas públicas de educação, esporte, entre outras. Não obstante tenha havido preocupação com a precisão do novo dado, baseado em amostra, há que se considerar que quanto menor é a área, maior é o impacto das mudanças havidas em um curto período de tempo. Neste caso, ao contrário, o uso de amostras periódicas é talvez a melhor maneira de se obter uma medida mais aproximada da realidade.

7. Considerações finais

A experiência francesa é um exemplo do que pode ser feito para redesenhar um censo, tão logo os dogmas de exaustividade e simultaneidade sejam abandonados. Diferentes aspectos devem ser considerados. Primeiro, examinar a demanda por informações, considerando o nível geográfico e a periodicidade requeridos. Um censo tradicional pode fornecer informações precisas para qualquer nível geográfico, assim como para a combinação destes níveis, como bairros, por exemplo, com temas como trabalho ou outros, de acordo com as necessidades dos usuários. Quando se usa amostra ou acumulação de amostras, é necessário definir o nível para o qual se deseja obter informações precisas, e levar isso em consideração no desenho amostral. A periodicidade para a produção das informações também deve ser considerada no desenho amostral. Por exemplo, um ciclo de cinco anos garante informações defasadas em apenas dois anos, em média. Se comparado ao modelo tradicional de censo decenal, o ciclo de cinco anos representa um ganho de sete anos em atualização dos dados produzidos.

Considerando amostras anuais, como na França, dependendo do nível geográfico, pode-se obter amostras representativas a cada ano ou a cada ciclo de cinco anos. Estas amostras podem ser composta por enumeração completa da população, como nos pequenos municípios a cada cinco anos ou, amostras a serem utilizadas anualmente, se foram grandes o suficiente, como para regiões ou municípios maiores. Pode ainda ser feita combinação de mais de uma amostra anual ao longo do ciclo, dependendo do tamanho da área para a qual se deseja obter a estimativa.

O desenho amostral pode ser feito de diferentes maneiras. Se for possível

construir um cadastro de unidades, quer seja de endereços completos ou somente de prédios, isto pode ser muito útil. Caso contrário, as unidades amostradas podem ser pequenas áreas como setores censitários. Os setores podem ser completamente enumerados ou pode ser feita uma amostra de domicílios do setor, após terem sido listados. O custo de construir um cadastro de endereços deve ser comparado com o custo de aumentar o tamanho da amostra para obter a mesma precisão que seria obtida com a seleção direta dos domicílios cadastrados. Outras fontes potenciais devem ser examinadas. Tanto registros quanto outras fontes administrativas, como cadastros de impostos (IPTU e ITR), de fornecedores de energia elétrica ou outros serviços, podem fornecer informações sobre os endereços.

Também no Brasil, a necessidade de produção de informações detalhadas, com maior freqüência que as atualmente produzidas é uma motivação, contundente, para que seja avaliada a utilização de modalidade alternativa ao censo tradicional. Atualmente, as informações para os níveis mais desagregados, para os mais diversos usos, como elaboração de políticas públicas nas áreas de saúde, educação, transporte, entre outras, são produzidas a cada dez anos quando da realização do Censo Demográfico brasileiro. No período intercensitário, importantes aplicações, baseadas no tamanho da população, como a distribuição dos recursos do Fundo de Participação dos Municípios, são feitas tomando em conta estimativas que podem estar defasadas em até nove anos, uma vez que o censo brasileiro é realizado a cada dez anos. Considera-se crítica esta defasagem uma vez que os métodos utilizados para obtenção das estimativas não incorporam, de maneira precisa, fatores como a mobilidade da população, considerada alta no Brasil.

A necessidade de redução da participação do informante não é ainda a principal preocupação brasileira, assim como o é na Europa, contudo tem crescido a cada censo a incidência de pessoas não recenseadas (falhas de cobertura), principalmente em razão da mudança do padrão de comportamento associado, entre outros, ao crescimento da violência urbana e à necessidade de que todos os membros das famílias trabalhem, mais comum nos grandes centros urbanos. Sendo assim, há que se levar em consideração as vantagens associadas aos levantamentos de informação por amostra, usualmente com melhores taxas de cobertura.

Considerando a situação do país, onde as fontes de registros administrativos existentes não são suficientes para prover informações completas e de boa qualidade sobre a população, encontra-se, nisto mesmo, o primeiro obstáculo para a utilização da modalidade baseada em registros administrativos. Pensando na possibilidade de criação ou melhoria da qualidade dos registros já existentes, é inevitável que tenhamos que levar em conta, além do custo, as possibilidades concretas de fazê-lo. No caso do Brasil, já com mais de 180 milhões de habitantes, uma extensão territorial de 8 514 877 km^2, e grandes diversidades social, cultural e econômica, não é trivial manter, suficientemente, atualizados registros de nascimentos e de mortes, por exemplo. Além disso, o alto custo de uma operação desta natureza dificilmente justificaria a utilização desta forma como alternativa ao modelo tradicional de censo, ao menos no que concerne à redução de custo. Sendo assim, na atual realidade brasileira, a busca de alternativa metodológica que possa substituir o modelo tradicional de censo por enumeração completa da população, passa necessariamente por avaliar a possibilidade do uso de pesquisas periódicas por amostra, contudo o uso de tal metodologia suscita um sem número de aspectos a serem avaliados.

Dentre as principais questões a serem levadas em consideração, está o tipo de informação demandada e em que medida elas podem ser produzidas a partir de amostras periódicas não sobrepostas. Atualmente, são produzidas a partir do censo tradicional, além de características demográficas da população, informações referentes aos temas educação, trabalho e rendimento, fecundidade e mortalidade infantil, migração, entre outras. Para que a utilização de uma nova modalidade seja viável, é necessário que atenda pelo menos as principais demandas já existentes. Além disso, é desejável que seja capaz de atender novas demandas, uma vez que a produção de informações com maior freqüência, despertará o interesse de usuários não tradicionais.

Considerando-se o grande desafio que representa o estudo de uma nova modalidade de Censo Demográfico, o presente artigo mais que apresentar o modelo francês como alternativa ao censo brasileiro, pretende mostrar as possibilidades oferecidas pela experiência já em curso assim como suscitar a discussão em torno do tema. Contudo, muito trabalho ainda há para ser feito em relação ao tema.

A continuidade do debate pode se dar a partir dos estudos já realizados por Alexander(1997), Bertrand (200-), Desplanques(2003), Durr e Dumais(2002), Durr e Grosbras (200-), Jay (2004), Kish (1984, 1987, 1991 e 2001) e Rao (1986), entre outros autores.

Referências bibliográficas

ALEXANDER, C. H. Still Rolling: Leslie Kish's "Rolling Samples" and the American Community Survey. Proceedings of Statistics Canada Symposium 2001. Achieving Data Quality in a Statistical Agency: A Methodological Perspective. Canada. 2001.

ALEXANDER, C. H. A Continuous Measurement Alternative for the U.S. Census.

ALEXANDER, C. H. A Rolling Sample Survey for Yearly and Decennial Uses.

ALEXANDER, C. H. The American Community Survey: Design Issues and Initial Test Results. Proceedings of Symposium 97, New Directions in Surveys and Censuses. Ottawa. Canada. pp 187-192. 1997.

BERTRAND, Philippe e outros. Données produites par le recensement rénové de la population. Paris: INSEE, Programme de rénovation du recensement, [200-].

BERTRAND, Philippe e outros. Les plans de sondage du nouveau recensement. Paris: INSEE, Programme de rénovation du recensement, [200-].

DESPLANQUES, Guy. Le nouveau recensement français de la population. In: Conference of European Statisticians. Ohrid, the former Yugoslav Republic of Macedonia, 21 a 23 May 2003. Disponível em<http://www.unece.org/stats/.

DEVILLE, J.C., TILLÉ, Y. Echantillonnage équilibré par la méthode du cube et estimation de variance. Journées de Méthodologie. INSEE : Paris. 2000.

DURR, Jean-Michel e DUMAIS, Jean. Redesign of the French census of population. Ottawa: Statistics Canada. Vol 28, Nº 1, pp 43-49. 2002.

DURR, Jean-Michel e GROSBRAS, Jean-Marrie. La rénovation du recensement français: principes et méthode. Paris:INSEE, Programme de rénovation du recensement, [200-].

DURR, Jean-Michel. The French new rolling census. Working paper em UNECE Seminar on New Methods for Population Censuses. Genebra: 2004. Disponível em: http://www.unece.org/stats/documents/2004.11.censussem.htm.

GRIFFIN, D. H. The Role of the American Community Survey in Re-engineering the United States Census of Population and Housing. U.S. Census Bureau Draft version, 2003.

JAY, Preston & REIST, Burton H. Reengineering the Census of Population and Housing. Working paper em UNECE Seminar on New Methods for Population Censuses. Genebra: 2004. Disponível em: http://www.unece.org/stats/documents/2004.11.censussem.htm.

KAMEN, Charles S. The 2008 Israel Integrated Census of Population and Housing. Working paper em UNECE Seminar on New Methods for Population Censuses. Genebra: 2004. Disponível em: http://www.unece.org/stats/documents/2004.11.censussem.htm.

KISH, Leslie and VERMA, Vijay. Census plus Survey: combined uses and Designs. Bulletin of the International Statistical Institute. 1983.

KISH, Leslie e VERMA, Vijay. Complete Censuses and Samples. Journal of Official Statistics. Stockholm: Statistics Sweden. Vol 2, nº 4, pp 381-395. 1986.

KISH, Leslie. Samples and Censuses. International Statistical Review. 1979a.

KISH, Leslie. Rolling Samples instead of Censuses, Asian and Pacific Census Forum. 1979b.

KISH, Leslie. Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau, Washington, 1981.

KISH, Leslie. Data Collection for Details over Space and Time. Statistical Methods and the Improvement of Data Quality. New York: Academic. 1983

KISH, Leslie. Timing of Surveys for Public Policy. Australian Journal of Statistics. 1984.

KISH, Leslie. Statistical Research Design. John Wiley & Sons. New York. 1987.

KISH, Leslie. Rolling Samples and Censuses. Survey Methodology, 16. 1990.

KISH, Leslie. The Census of 2000 AD and Beyond. Review of Major Alternatives for the Census in the year 2000. Washington D.C.. 1991.

KISH, Leslie. Periodic and Rolling Samples and Censuses, in Statistics and Public Policy, 7. Clarendon Press, Oxford. 1997.

KISH, Leslie. Space/Time Variations and Rolling Samples. Journal of Official Statistics, 14, 1, 1998, pp. 31-46. 1998.

KISH, Leslie. Combining/Cumulating Population Surveys. Journal of Statistical Planning and Inference. 2001.

NORDHOLT, Eric Schulte. The Dutch virtual census 2001: A New Approach by Combining Different Sources. Working paper em UNECE Seminar on New Methods for Population Censuses. Genebra: 2004. Disponível em: http://www.unece.org/stats/documents/2004.11.censussem.htm.

RAO, J. N. K. Small Area Estimation. New York: John Willey & Sons Inc. 2003. Statistics. Stockholm: Statistics Sweden. . Vol 2, nº 4, pp 381-395. 1986.

SZENZENSTEIN, Johann. The new method of the next German population census. Working paper em UNECE Seminar on New Methods for Population Censuses. Genebra: 2004. Disponível em: http://www.unece.org/stats/documents/2004.11.censussem.htm.

UNITED NATIONS: Economic Commission for Europe and the Statistical Office of the European Communities. Recommendations for the 2000 censuses of population and housing in the ECE Region..
Disponível em:
http://www.unece.org/stats/documents/statistical_standards_&_studies/49.e.html#SIIA

UNITED NATIONS: Economic Commission for Europe and the Statistical Office of the European Communities. Recommendations for the 2010 censuses of population and housing in the ECE Region. Disponível em:
http://www.unece.org/stats/documents/census/2000/CensusRecommendations.html

UNITED STATES CENSUS BUREAU. American Community Survey Operations Plan. 2003. Disponível em http://www.census.gov/acs/.

Abstract

The most pressing necessity of producing detailed information at geographic level and more frequently than those from quinquennial or decennial censuses; as well as the increase of the difficulties to guarantee financial resources for the accomplishment of the national censuses, has motivated the search of alternative methods to supply the demand for information, traditionally complied by the Demographic Censuses. One of the alternatives to the current model of complete enumeration of the population is the use of non-overlapping periodic samples proposed by the Hungarian statistician Leslie Kish. The base of the model considered by Kish is the accumulation of samples throughout the space and/or time. Amongst the experiences in course, the French one is in the most advanced stage of the implantation of the modality proposed by Kish, thus, the present article brings aspects related to sampling plan and estimate in the New French Census.

Comparação de desempenho de modelos da família ARCH

Emerson Herbert Amorim[*]
Luis Aparecido Milan[**]

Resumo

Estimar e prever a volatilidade de índices de mercados de ações, produtos e moedas são de grande importância para as decisões dos gestores de risco envolvidos em aplicações financeiras. Neste trabalho apresentamos uma revisão de alguns modelos paramétricos da "família ARCH" (*Autoregressive Conditional Heterocedasticity*) e mostramos sua aplicação na estimação da volatilidade de ativos financeiros. Também apresentamos uma comparação de desempenho destes modelos. Como ilustração os métodos são aplicados à série do IBOVESPA desde o início do Plano Real (02/01/1992 a 24/03/2003). Os critérios de adequabilidade empregados demonstraram resultados satisfatórios para todos os modelos testados, com pequena superioridade do modelo GARCH-M (1,1).

[*] Endereços para correspondências: Departamento de Estatística Universidade Federal de São Carlos - Caixa Postal 676 - CEP: 13565-905 – São Carlos – SP – Brasil. – *e-mail*: ehamorim@hotmail.com.
[**] Departamento de Estatística Universidade Federal de São Carlos - Caixa Postal 676 CEP: 13565-905 – São Carlos – SP – Brasil - *e-mail*: dlam@power.ufscar.br.

1. Introdução

A dinâmica e agitação do mercado financeiro mundial têm exigido métodos de modelagem mais sofisticados, complexos e eficientes com o intuito de melhor descrever as tendências e características dos ativos financeiros.

No mercado financeiro, a volatilidade indica se o preço de um ativo está variando muito ou pouco. Esta é a medida da incerteza quanto às variações de preço. Em épocas de alta variabilidade dos preços, são maiores as chances de grandes perdas ou lucros nos investimentos, isto é, são as épocas de maior risco. Já quando a volatilidade é baixa, o risco é menor.

As principais características dos ativos financeiros, como a média aproximadamente zero, alta curtose, efeito de assimetria e não estacionariedade na média, podem ser reconhecidas através de uma simples análise gráfica. A não estacionariedade na média é a principal característica. Para tratar esse tipo de dados utilizamos transformações.

A transformação mais utilizada é a taxa de retorno y_t definida por

$$y_t = \ln\left(\frac{P_t}{P_{t-1}}\right), \tag{1.1}$$

onde P_t é o preço de um ativo no tempo t.

O efeito de alavancagem (*leverage effect*) é outra característica importante e pode ser definido pela presença de regimes de baixa e alta variabilidade ou volatilidade no retorno em contraste com as tendências da série de preços, ou seja, quando o nível de preços apresenta movimentos com certa constância de alta, a série de retorno segue com baixa volatilidade; já quando a série de preço apresenta-se em tendência de queda, a série de retorno segue com alta considerável na volatilidade.

A presença de regimes de baixa e alta volatilidade na série de retorno ao longo do tempo é a característica de maior interesse para um gestor de risco no planejamento de

suas regras de decisões. Como os tradicionais modelos lineares Gaussianos de séries temporais não conseguem incorporar com eficiência tais características dos ativos financeiros, temos por objetivo neste trabalho apresentar alguns modelos auto-regressivos heterocedásticos capazes de prever a volatilidade das séries de retorno facilitando as decisões de investimentos dos gestores de risco, assim como medir a eficiência desses modelos no ajuste de uma série financeira brasileira, a série de índices IBOVESPA (02/01/1992 a 24/03/2003).

Neste trabalho apresentamos os tradicionais modelos da "família ARCH" (*Autoregressive Conditional Heterocedasticity*) como uma solução viável para modelagem do comportamento da volatilidade em séries de índices de preços de ações como o IBOVESPA.

A seção 2 apresenta as especificações dos modelos da "família ARCH" na seqüência de sua evolução ao longo do tempo. A seção 3 apresenta em detalhes o processo de estimação dos parâmetros de interesse assim como algumas distribuições utilizadas em tal processo. A seção 4 descreve o processo de seleção dos modelos propostos bem como o processo de previsão da volatilidade. A seção 5 apresenta o ajuste desses modelos à série de índices IBOVESPA e a comparação dos mesmos. As demais seções descrevem as conclusões obtidas neste trabalho, bem como as referências utilizadas para tanto.

2. Modelos auto-regressivos heterocedásticos

Os modelos auto-regressivos heterocedásticos apresentados nessa seção são de grande interesse para as áreas de economia e finanças. Isto se deve à viabilidade desses modelos quanto a sua aplicação e interpretação, e à capacidade de relacionar volatilidade, risco e retorno.

Esse relacionamento é possível pela flexibilidade dos modelos na modelagem da média e da variância da série y_t.

O primeiro modelo denotado pela sigla ARCH, que vem do seu nome em inglês *Autoregressive Conditional Heterocedasticity*, foi desenvolvido por Engle em 1982. Na

seqüência temos os modelos GARCH desenvolvidos por Bollerslev (1986). A seguir viera os modelos GARCH-M, desenvolvidos por Engle et al (1987) e os modelos GARCH-L desenvolvidos por Runkle et al (1993).

Descreveremos, a seguir, cada um destes modelos.

2.1. ARCH (m)

Dada uma série de dados $Y_t = \{y_t\}, \quad t = 1,2,\ldots,T$ o modelo linear univariado ARCH (m), proposto por Engle (1982), pode ser descrito como

$$y_t = x_t'\delta + z_t \tag{2.1}$$

$$z_t \sqrt{h_t}\varepsilon_t \tag{2.2}$$

$$h_t = E[z_t^2 \left| z_{t-1}^2,\ldots,z_{t-m}^2 \right.] = \alpha_0 + \sum_{j=1}^{m} \alpha_j z_{t-j}^2 \tag{2.3}$$

onde,

$x_t'\delta$ é a média do processo y_t, sendo x_t um vetor de variáveis explicativas e δ um vetor de parâmetros;

$\{\varepsilon_t\}$ é uma seqüência de variáveis aleatórias independentes e identicamente distribuídas, iid, com média zero e variância unitária; e

$\{z_t\} > 0$ é uma seqüência correlacionada, onde $z_t^2 \left| z_{t-1}^2,\ldots,z_{t-m}^2 \right.$ segue uma distribuição $P(0,h_t)$. A distribuição $P(o)$ pode ser escolhida entre várias distribuições conhecidas, entre elas destacam-se o uso das distribuições normal e t-Student. A escolha da distribuição mais adequada para $P(\cdot)$ depende do tipo de aplicação e dos dados disponíveis.

Neste modelo a variância condicional h_t, dada em 2.3, é função dos valores passados ao quadrado do processo z_t. Para assegurar que h_t seja estritamente positiva, temos a restrição que $\alpha_j > 0$ para $j = 0,1,\ldots,m$.

2.2. GARCH (r, m)

Para descrever corretamente o comportamento da variância condicional h_t, é necessário em algumas séries acrescentar vários valores passados de z_t^2, o que normalmente infringe as condições pressupostas para os parâmetros α_j, $j = 0,1,\ldots,m$. O modelo GARCH (r, m), introduzido por Bollerslev (1986), sugere uma estrutura com mais flexibilidade, pois incrementa à variância condicional h_t os valores passados de h_{t-r}, o que torna o modelo mais parcimônico.

O modelo linear univariado GARCH (r, m) é definido 2.1, 2.2 e sua variância condicional é dada por

$$h_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j z_{t-j}^2 + \sum_{i=1}^{r} \beta_i h_{t-i} \,. \tag{2.4}$$

Neste modelo, h_t é função dos valores dos erros passados ao quadrado e dos valores passados da variância condicional.

A seguinte restrição deve ser aplicada sobre os parâmetros,

$$\sum_{j=1}^{m} \alpha_j + \sum_{i=1}^{r} \beta_i < 1$$

para garantir que (i) a variância incondicional de z_t seja finita e positiva, (ii) a variância condicional h_t seja positiva e também para garantir a estacionariedade do processo GARCH (r, m), ver o teorema 1 em Bollerslev (1986).

Temos também que

(i) $z_t \sim P(0,\sigma_z^2)$, onde $\sigma_z^2 = \dfrac{\alpha_0}{1-\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{r}\beta_i}$

e

(ii) $z_t^2 \left| z_{t-1}^2,\ldots,z_{t-m}^2 \right. \sim P(0,h_t)$

No caso de violação da restrição sobre os parâmetros, Engle e Bollerslev (1986) propõem o modelo *Integrated* GARCH, denotado por IGARCH (r, m), onde a variância incondicional de z_t é infinita.

2.3. GARCH-M (r, m)

Em algumas séries heterocedásticas, principalmente nas séries financeiras, a média do processo y_t sofre mudanças de escala durante o tempo.

Para estes casos, o modelo univariado GARCH-M (r, m) é introduzido por Engle et al (1987). Este modelo é uma extensão do modelo GARCH e introduz na equação da média a variância condicional com um peso γ com a finalidade de captar mudanças de escala na média do processo.

A equação da média do modelo GARCH-M é dada por

$$y_t = x_t'\delta + \gamma h_t + z_t . \qquad (2.5)$$

Sendo z_t dado por 2.2 e a variância condicional h_t é dada por (2.4), cada qual seguindo, respectivamente, suas restrições.

2.4. GARCH-L (r, m)

O modelo GARCH-L (r, m) proposto em Runkle et al (1993), também, é uma extensão do modelo GARCH. Este modelo acrescenta na variância condicional h_t um parâmetro λ que tem por objetivo capturar o efeito de assimetria sobre z_t.

O modelo GARCH-L (r, m) é definido por 2.1. A variância condicional é descrita da forma,

$$h_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j z_{t-j}^2 + \sum_{i=1}^{r} \beta_i h_{t-i} + \lambda z_{t-1}^2 I(z_{t-1}), \quad (2.6)$$
$$(\geq 0)$$

e

$$\left\{ \begin{matrix} I(z_{t-1}) = 1, & se & z_{t-1} \geq 0 \\ I(z_{t-1}) = 0, & se & z_{t-1} < 0 \end{matrix} \right\}.$$

O parâmetro λ varia no intervalo (0, 1) e capta o efeito da assimetria. Portanto, se $\lambda = 0$ não existe o efeito de assimetria. As restrições sobre os demais parâmetros da variância condicional h_t são as mesmas descritas para a variância condicional do modelo GARCH.

Além do modelo GARCH-L, existe o modelo *Exponential* GARCH proposto em Nelson (1991) que também capta o efeito de assimetria sobre z_t, mas sem a necessidade de qualquer restrição sobre os parâmetros da variância condicional de z_t, ver detalhes em Nelson (1991).

Existe uma variabilidade de outros modelos propostos para modelagem da variância condicional h_t dentro dos modelos denominados "família ARCH". Entre eles, destacam-se os modelos com mudança de regimes markovianos na variância do processo, modelos estes mais recentes que foram introduzidos em Hamilton e Susmel (1994) e Dueker (1997).

3. Processo de Estimação dos Parâmetros

No processo de estimação dos parâmetros envolvidos nos modelos da "família ARCH" nos deparamos com alguns métodos, dentre eles: estimação via método dos

momentos, máxima verossimilhança e quasi-máxima verossimilhança, ver Hamilton (1994), cap. 21, pp. 661-665.

O método de estimação via maximização da função de verossimilhança é o mais utilizado. Nesse método encontramos na literatura o uso da distribuição normal para y_t, no primeiro trabalho, Engle (1982). Em seguida, Bollerslev (1987) propõe o uso da distribuição t-Student.

Nesta seção, descrevemos em detalhes o processo de estimação via maximização da função de verossimilhança apenas para o modelo GARCH com distribuição normal, visto que os demais modelos são simplificações ou complementações do mesmo e seguem o mesmo procedimento com pequenas variações.

Denotando Y_t como um vetor de observações obtidas ao longo do tempo t, temos que se $\varepsilon_t \sim N(0,1)$, então a distribuição condicional de y_t é normal com média $x_t'\delta$ e variância condicional h_t.

A densidade condicional de y_t é denotada por

$$f(y_t|x_t, Y_{t-1}) = \frac{1}{\sqrt{2\pi h_t}} \exp\left(-\frac{(y_t - x_t'\delta)^2}{2h_t}\right). \tag{3.1}$$

No modelo GARCH, os parâmetros desconhecidos que temos interesse em estimar formam o seguinte vetor $\underset{\sim}{\theta}$,

$$\underset{\sim}{\theta} = (\delta', \alpha_0, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_r) = (\delta', W').$$

A função de verossimilhança para o vetor $\underset{\sim}{\theta}$ dado z_t, é dada por

$$L(\underset{\sim}{\theta}|z_t) = \prod_{t=d+1}^{T} \left(\frac{1}{h_t}\right)^{\frac{1}{2}} \exp\left(-\frac{z_t^2}{2h_t}\right), \tag{3.2}$$

onde $d = \max\{r, m\}$.

A média do logaritmo da função de verossimilhança é

$$L(\underset{\sim}{\theta}|z_t) = \left(\frac{1}{T}\right)\sum_{t=d+1}^{T} l_t(\underset{\sim}{\theta}|z_t), \tag{3.3}$$

onde $l_t(\underset{\sim}{\theta}|z_t)$ indica o logaritmo da função da verossimilhança na t-ésima observação, ou seja,

$$l_t(\underset{\sim}{\theta}|z_t) = -\frac{1}{2}\log(h_t) - \frac{z_t^2}{2h_t}. \tag{3.4}$$

As derivadas de primeira ordem, com respeito aos parâmetros da variância condicional h_t, são da forma,

$$\begin{aligned} \frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial w} &= -\frac{1}{2h_t}\frac{\partial h_t}{\partial w} + \frac{z_t^2}{2h_t^2}\frac{\partial h_t}{\partial w} \\ \frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial w} &= \frac{1}{2h_t}\left(\frac{\partial h_t}{\partial w}\right)\left(\frac{z_t^2}{h_t} - 1\right) \end{aligned}, \tag{3.5}$$

onde w é o vetor de parâmetros de h_t.

As derivadas de segunda ordem, com respeito aos parâmetros da variância condicional h_t, são da forma,

$$\begin{aligned} \frac{\partial^2 l_t(\underset{\sim}{\theta}|z_t)}{\partial w \partial w'} &= \frac{\partial h_t}{\partial w'}\left[\frac{1}{2h_t}\left(\frac{\partial h_t}{\partial w}\right)\left(\frac{z_t^2}{h_t} - 1\right)\right], \\ \frac{\partial^2 l_t(\underset{\sim}{\theta}|z_t)}{\partial w \partial w'} &= -\frac{1}{2h_t^2}\left(\frac{\partial h_t}{\partial w}\right)\left(\frac{\partial h_t}{\partial w'}\right)\left(\frac{z_t^2}{h_t} - 1\right) \\ &\quad + \frac{1}{2h_t}\left(\frac{\partial h_t}{\partial w \partial w'}\right)\left(\frac{z_t^2}{h_t} - 1\right) + \frac{1}{2h_t}\left(\frac{\partial h_t}{\partial w'}\right)\left(-\frac{z_t^2}{h_t}\frac{\partial h_t}{\partial w}\right). \end{aligned} \tag{3.6}$$

Com o resultado acima, é possível calcular a matriz de informação de Fisher, obtida através da esperança condicional do termo 3.6, ou seja, a matriz I é obtida da seguinte forma

$$I = -E\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_t(\underset{\sim}{\theta}|z_t)}{\partial w \partial w'}\right].$$

A esperança condicional dos dois primeiros termos em 3.6 é zero, pois

$$E\left[\frac{z_t^2}{h_t} - 1\right] = \frac{1}{h_t}E[z_t^2|z_{t-1},\ldots] - 1 = \frac{h_t}{h_t} - 1 = 0,$$

logo a matriz de informação de Fisher é consistentemente estimada pelo último termo de 3.6, que envolve apenas a primeira derivada de h_t, portanto

$$I = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2h_t}\left(\frac{\partial h_t}{\partial w'}\right)\left(\frac{\partial h_t}{\partial w}\right). \tag{3.7}$$

No processo de estimação do modelo GARCH, a única diferença em relação ao processo de estimação do modelo ARCH é quanto à inclusão de valores passados da variância condicional h_t, ou seja, o processo de estimação passa a ser condicionado nas estimativas de h_t.

Portanto, temos que a derivada $\left[\frac{\partial h_t}{\partial w}\right]$ no modelo GARCH é denotado por

$$\frac{\partial h_t}{\partial w} = z_t + \sum_{i=1}^{r}\beta_i\frac{\partial h_{t-i}}{\partial w}. \tag{3.8}$$

Para inicializar o processo recursivo descrito em (3.8), Bollerslev (1986) sugere que para $t \leq 0$, h_t e z_t^2 podem ser obtidos por $\frac{\sum_{t=1}^{T} z_t^2}{T}$, ou seja, pela variância do processo z_t.

As derivadas de primeira ordem, com respeito aos parâmetros da média, são da forma,

$$\frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial \delta} = z_t x_t h_t^{-1} + \frac{1}{2} h_t \frac{\partial h_t}{\partial \delta}\left(\frac{z_t^2}{h_t} - 1\right), \tag{3.9}$$

e as derivadas de segunda ordem, com respeito aos parâmetros da média, são da forma,

$$\begin{aligned}\frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial \delta \partial \delta'} = &-h_t^{-1} x_t x_t' - \frac{1}{2} h_t^{-2} \frac{\partial h_t}{\partial \delta}\frac{\partial h_t}{\partial \delta'}\left(\frac{z_t^2}{h_t}\right) \\ &-2h_t^{-2} z_t x_t \frac{\partial h_t}{\partial \delta} + \left(\frac{z_t^2}{h_t} - 1\right)\frac{\partial h_t}{\partial \delta'}\left[\frac{1}{2} h_t^{-1}\frac{\partial h_t}{\partial \delta}\right],\end{aligned} \tag{3.10}$$

onde

$$\frac{\partial h_t}{\partial \delta} = -2\sum_{j=1}^{m} \alpha_j x_{t-1} z_{t-j} + \sum_{i=1}^{r} \beta_i \frac{\partial h_{t-i}}{\partial \delta}. \tag{3.11}$$

Para encontrarmos as estimativas de máxima verossimilhança dos parâmetros incluídos no vetor $\underset{\sim}{\theta}$, um método iterativo foi desenvolvido por Berndt et al (1974) e presentado por Bollerslev (1986), é apresentado para encontrar as estimativas de $\underset{\sim}{\theta}$. O algoritmo BHHH pode ser descrito pela relação

$$\theta^{(k+1)} = \theta^{(k)} + \lambda_{(k)} \left(\sum_{t=1}^{T} \frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial \theta} \frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial \theta'} \right)^{-1} \left(\sum_{t=1}^{T} \frac{\partial l_t(\underset{\sim}{\theta}|z_t)}{\partial \theta} \right), \tag{3.12}$$

onde $\lambda_{(k)}$ é escolhido de forma a maximizar a função de verossimilhança e é calculado usando-se o método da razão áurea, ver Luenberger (1984).

Para os estimadores de máxima verossimilhança do vetor $\underset{\sim}{\hat{\theta}}$ podemos assumir distribuição assintótica com distribuição limite, dada por

$$\sqrt{T}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N(0, I^{-1}), \tag{3.13}$$

onde I é a matriz de informação é dada por 3.7.

O processo descrito, até aqui, utiliza-se do fato de z_t assumir distribuição normal, entretanto nem sempre essa suposição é aceita. Bollerslev (1987) propõe a utilização da distribuição t-Student para z_t com ν graus de liberdade e parâmetro de escala M_t, neste caso a densidade condicional de z_t é dada por,

$$f(z_t|M_t) = \frac{\Gamma(\frac{\nu+1}{2})}{(\pi\nu)^{\frac{1}{2}}\Gamma(\frac{\nu}{2})} M_t^{-\frac{1}{2}} \left[1 + \frac{z_t^2}{M_t \nu}\right]^{-\frac{(\nu+1)}{2}}, \tag{3.14}$$

onde $\Gamma(\cdot)$ é a função gama.

Para $\nu > 2$ temos que $E(z_t) = 0$ e $E(z_t^2) = M_t \frac{\nu}{(\nu-2)}$.

Para se obter uma distribuição t-Student com média ν e variância condicional h_t, Bollerslev (1987) propõe o fator de escala M_t,

$$M_t = h_t \frac{(\nu-2)}{\nu}. \tag{3.15}$$

Com isso, a densidade 3.14 é denotada por,

$$f(z_t|h_t)=\frac{\Gamma(\frac{v+1}{2})}{\pi^{\frac{1}{2}}\Gamma(\frac{1}{2})}(v-2)^{-\frac{1}{2}}h_t^{-\frac{1}{2}}\left[1+\frac{z_t^2}{h_t(v-2)}\right]^{-\frac{(v+1)}{2}}. \tag{3.16}$$

O logaritmo da função de verossimilhança condicionado nas $d=\max\{r,m\}$ primeiras observações é,

$$\partial l_t(\underset{\sim}{\theta}|z_t)=\sum_{t=d+1}^{T}\log f\left(y_t|x_t,Y_{t-1},\underset{\sim}{\theta}\right) \tag{3.17}$$

$$\begin{aligned}\partial l_t(\underset{\sim}{\theta}|z_t)=T\log\left\{\frac{\Gamma\left(\frac{(v+1)}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{v}{2}\right)}(v-2)^{-\frac{1}{2}}\right\}-\left(\frac{1}{2}\right)\sum_{t=1}^{T}\log(h_t)\\ -\left[\frac{v+1}{2}\right]\sum_{t=d+1}^{T}\log\left[1+\frac{(y_t-x_t'\delta)^2}{h_t(v-2)}\right].\end{aligned} \tag{3.18}$$

O logaritmo da função de verossimilhança descrito em 3.18 é maximizado numericamente de acordo com o método descrito em 3.12 para o vetor $\underset{\sim}{\theta}$ de parâmetros,

$$\underset{\sim}{\theta}=(\delta',\alpha_0,\alpha_1,\ldots,\alpha_m,\beta_1,\ldots\beta_r,v),$$

se $v>2$ graus de liberdade.

O processo de estimação dos demais modelos GARCH-M e GARCH-L seguem o mesmo procedimento do modelo GARCH apresentado nesta seção. Entretanto, em ambos os modelos temos um parâmetro a mais de interesse.

No modelo GARCH-M, temos o parâmetro γ que captura a influência das oscilações da variância ao longo do tempo, na média do processo y_t. No modelo

GARCH-L, temos o parâmetro λ que captura o efeito de assimetria na variância condicional h_t.

O processo de estimação incluindo esses novos parâmetros podem ser vistos em detalhes em Engle et al (1987) e Runkle et al (1993), respectivamente.

4. Seleção de Modelos de Previsão de h_t

No processo de seleção do modelo mais adequado para descrever uma série financeira heterocedástica alguns fatores devem ser levados em conta, entre eles estão a escolha do modelo em relação à ordem (r, m) dos parâmetros da variância condicional h_t, a qualidade do ajuste do modelo e a eficiência na previsão de h_t.

Em relação à escolha da ordem dos parâmetros da variância condicional h_t, Bollerslev (1986) apresenta uma representação equivalente dos modelos GARCH (r, m), dada por,

$$z_t^2 = \alpha_0 + \sum_{j=1}^{m} \alpha_j z_{t-j}^2 + \sum_{i=1}^{r} \beta_i z_{t-i}^2 - \sum_{i=1}^{r} \beta_i v_{t-i} + v_t \qquad (4.1)$$

$$v_t = z_t^2 - h_t \qquad (4.2)$$

onde v_t é serialmente não correlacionado e com média zero. Neste caso, o processo z_t^2 pode ser analisado como um processo ARMA de ordem $p = \max\{r, m\}$ e $q = r$, respectivamente.

Quanto à qualidade do ajuste dos modelos em análise, podemos usar alguns critérios de seleção introduzidos por Box e Jenkins (1976), dentre eles o valor do logaritmo da função de máxima verossimilhança, onde o maior valor indica o modelo mais plausível para descrever as características da série y_t, os critérios AIC (Akaike, 1974) e BIC (Shwarz, 1978) onde esses dois critérios qualificam o modelo pelo "critério

de parcimônia", escolhendo o melhor modelo pela relação entre o ajuste mais adequado e o menor número de parâmetros possíveis.

Os critérios AIC e BIC são dados por,

$$AIC = -\frac{2}{T} l(\underset{\sim}{\theta}) + \frac{2k}{T} \tag{4.3}$$

e

$$BIC = -\frac{2}{T} l(\underset{\sim}{\theta}) + \frac{k \ln(T)}{T}, \tag{4.4}$$

onde T é o número total de observações da série y_t, $l(\underset{\sim}{\theta})$ é o valor do logaritmo da função de máxima verossimilhança e $k = d + 1$, onde d é o número de parâmetros do modelo ajustado.

A estatística $Q(\cdot)$ de Ljung-Box (detalhes em Box e Jenkins, 1976) analisa os resíduos para verificar se existe uma estrutura de autocorrelação residual, e o teste ARCH-LM (detalhes em Bollerslev, 1986) indica se os resíduos exibem alguma estrutura ARCH adicional.

A eficiência do modelo quanto à previsão de h_t é sem dúvida um dos fatores mais importantes que devemos observar. Nos modelos da "família ARCH", a previsão m passos à frente da variância condicional h_t é dada por,

$$E[h_{t+m}] = E[z_{t+m}^2 | z_{t+m-1}, z_{t+m-2}, \ldots]. \tag{4.5}$$

No modelo GARCH (1, 1), por exemplo, temos que,

$$E[h_{t+m}] = E[\alpha_0 + \alpha_1 z_{t+m-1}^2 + \beta_1 h_{t+m-1}] \tag{4.6}$$

$$E[h_{t+m}] = \alpha_0 + \alpha_1 E[z_{t+m-1}^2 | z_{t+m-2}] + \beta_1 E[h_{t+m-1}]$$
$$E[h_{t+m}] = \alpha_0 + \alpha_1 h_{t+m-1} + \beta_1 h_{t+m-1}$$
$$E[h_{t+m}] = \alpha_0 + \eta h_{t+m-1}$$

onde $\eta = (\alpha_1 + \beta_1)$. O valor de η é conhecido como o valor da persistência na equação da variância condicional. Altos valores de η prejudicam a previsão dos modelos ajustados, pois η está diretamente relacionado com o peso dado ao erro cometido no instante $t-1$.

Devido ao parâmetro λ (responsável pela captura do efeito de assimetria) o valor da persistência no modelo GARCH-L é denotado por,

$$E[h_{t+m}] = E\left[\alpha_0 + \alpha_1 z_{t+m-1}^2 + \beta_1 h_{t+m-1} + \lambda z_{t+m-1}^2 I(z_{t+m-1})_{(\geq 0)}\right] \tag{4.7}$$

$$E[h_{t+m}] = \alpha_0 + \alpha_1 E[z_{t+m-1}^2 | z_{t+m-1}] + \beta_1 E[h_{t+m-1}] + \frac{\lambda}{2} E[z_{t+m-1}^2 | z_{t+m-1}]$$

$$E[h_{t+m}] = \alpha_0 + \alpha_1 h_{t+m-1} + \beta_1 h_{t+m-1} \frac{\lambda}{2} h_{t+m-1}$$

$$E[h_{t+m}] = \alpha_0 + \eta h_{t+m-1}$$

onde $\eta = (\alpha_1 + \beta_1 + \frac{\lambda}{2})$.

Outro índice que está diretamente ligado com o valor de η da persistência é a meia vida, que é dado por

$$MV = 1 - \frac{\log(2)}{\log(\eta)}. \tag{4.8}$$

O valor de MV fornece uma idéia do tempo que o modelo ajustado leva para estabilizar-se ou recuperar-se após ocorrências de picos na série z_t. Quanto maior o valor de η maior o valor de MV.

Alguns autores já analisaram os altos valores de η e MV em algumas séries. Segundo Dielbold (1986), alta persistência pode ser provocada pela não observação da mudança de regime na variância não condicional ao longo da série em análise.

Os modelos com mudança de regime SWARCH introduzido por Hamilton e Susmel (1994) e SWGARCH introduzido por Dueker (1997) analisam a mudança de regime na variância do processo. Esses modelos não serão abordados nesse trabalho, mas são boas alternativas para tais situações.

Seguindo o processo de análise da acurácia preditiva dos modelos da família ARCH, Hamilton e Susmel (1994) propõem quatro funções básicas de perda no contexto de previsão m passos à frente. Essas funções comparam os valores de h_t e z_t^2, são elas:

1. o erro médio quadrático, ou MSE, de Mean Square Error;

$$MSE = \sum_{i=1}^{L} \frac{(z_{T+i}^2 - h_{t|T+i})^2}{L}; \tag{4.9}$$

2. o erro absoluto médio, ou MAE, de Mean Absolute Error;

$$MAE = \sum_{i=1}^{L} \left| \frac{(z_{T+i}^2 - h_{t|T+i})}{L} \right|; \tag{4.10}$$

3. o erro quadrático médio dos logaritmos denotado por $[LE]^2$

$$[LE]^2 = \sum_{i=1}^{L} \frac{[\ln(z_{T+i}^2) - \ln(h_{t|T+i})]^2}{L}; \text{ e} \tag{4.11}$$

4. o erro absoluto médio dos logaritmos denotado por $|LE|$

$$|LE| = \sum_{i=1}^{L} \frac{\left|\ln(z_{T+i}^2) - \ln(h_{t|T+i})\right|}{L}. \tag{4.12}$$

As quatro funções indicam o quão distantes estão as estimativas da variância condicional da variância realizada ou observada.

O valor dessas quatro funções de perda pode ser calculado em uma amostra separada de tamanho L, conhecida como amostra de previsão, ou seja, divide-se a amostra total em duas partes. A primeira e maior parte é conhecida como amostra de ajuste do modelo e a segunda amostra de previsão é onde calcularemos os valores respectivos a cada função de perda dada por 4.9 a 4.12.

Além do cálculo das funções de perda, podemos verificar a freqüência dos intervalos de confiança estimados para y_t que cobrem o valor observado de y_t. Os intervalos com 95% de confiança assumindo distribuição normal ou t-Student são dados por,

$$IC(95\%) = \hat{\mu}_t \pm 1{,}96\sqrt{\hat{h}_t} \tag{4.13}$$

e

$$IC(95\%) = \hat{\mu}_t \pm \frac{t_{\hat{v}}\left(1-\frac{0{,}05}{2}\right)}{\sqrt{\frac{\hat{v}}{\hat{h}(\hat{v}-2)}}}, \tag{4.14}$$

respectivamente.

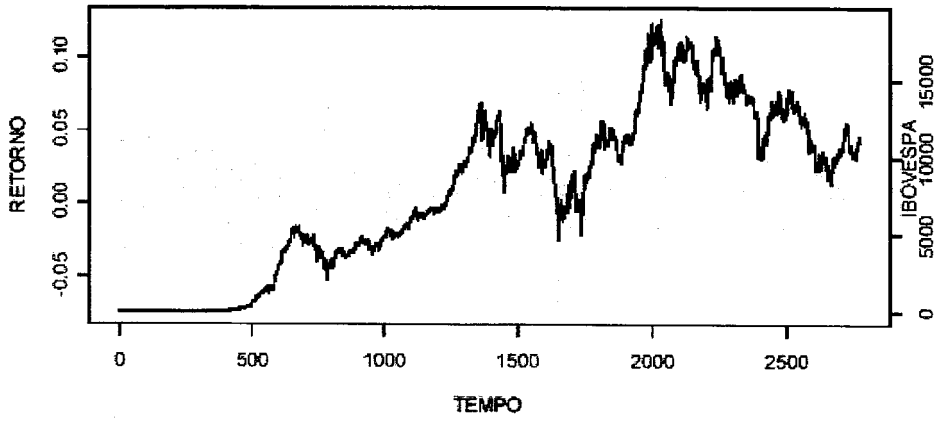
5. Aplicação à série IBOVESPA

A série IBOVESPA abrangendo o período de 02/01/1992 a 24/03/2003, tendo no total 2 771 observações, será analisada neste trabalho para verificar e comparar a eficiência dos modelos auto-regressivos heterocedásticos. Uma análise da série IBOVESPA utilizando modelos da família GARCH pode ser encontrada em Sobrinho (2001).

O IBOVESPA é sem dúvidas o indicador mais importante do mercado de ações brasileiro, pois ele retrata o comportamento das principais ações negociadas na BOVESPA. O índice é formado a partir de uma aplicação imaginária, em Reais, em uma quantidade teórica de ações e tem como finalidade básica servir como indicador médio do comportamento do mercado. Como as ações que fazem parte dessa carteira têm

grande representatividade, podemos dizer que se a maioria delas estiver subindo, o mercado, medido pelo Índice IBOVESPA, está em alta, e se estiver caindo, está em baixa.

Figura 1 - IBOVESPA e retorno



A Figura 1 apresenta a série do IBOVESPA assim como a série de retorno. Verificamos neste gráfico o efeito de alavancagem presente, determinado pelos regimes de baixa e alta volatilidade da série de retorno decorrente do crescimento ou queda da série de índices IBOVESPA.

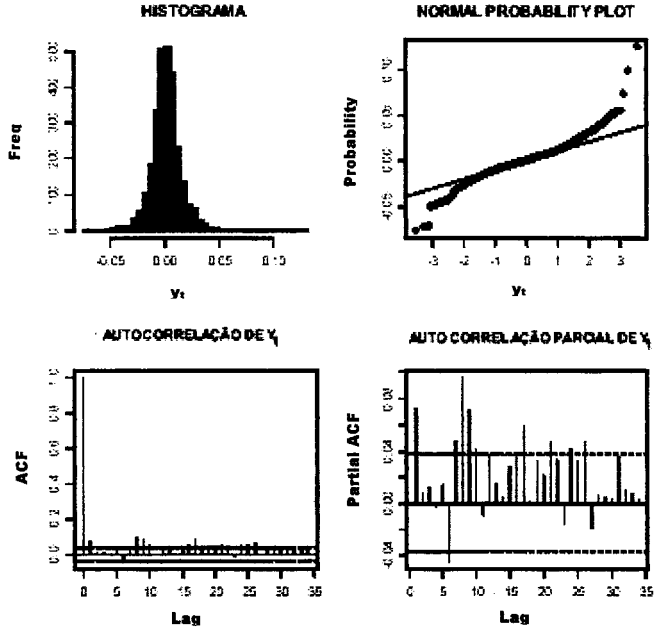
A Tabela 1 apresenta estatísticas descritivas da série de retorno, de onde verificamos que a média é aproximadamente zero e a curtose é relativamente alta.

Tabela 1 - Análise descritiva da série de retorno

Estatísticas	Valor
Média	0,001521
Mediana	0,001237
Máximo	0,125153
Mínimo	-0,481610
Desvio Padrão	0,013610
Curtose	5,598351

A Figura 2 apresenta o histograma e o gráfico probabilístico normal do retorno, desses dois gráficos identificamos que a série de retorno não é normal e possui caldas relativamente pesadas. Os gráficos de autocorrelação e autocorrelação parcial mostram baixa correlação indicando que a série de retorno pode ser considerada um passeio aleatório. Diante desse fato, ajustaremos os modelos da família ARCH com média constante.

Figura 2 - Análise da distribuição de Y_t



A série z_t, gerada a partir da média do processo descrito na Tabela 1, foi analisada como um processo ARMA de acordo com o procedimento descrito em 4.1. Esse procedimento aponta para o ajuste de modelos com ordem $m = r = 1$.

A Tabela 2 apresenta as estimativas dos parâmetros dos modelos GARCH (1,1), GARCH-L (1,1) e GARCH-M (1,1), assumindo distribuição normal e t-Student. O ajuste

foi feito com 2 500 observações, sendo que as 271 observações restantes foram utilizadas para verificação de previsão.

Tabela 2 - Parâmetros estimados

Parâmetros	GARCH(1,1)–N	GARCH(1,1)–t	GARCH-M (1,1)–N	GARCH-M (1,1)–t	GARCH-L (1,1)–N	GARCH-L (1,1)–t
(média)	0,001074	0,001105	0,000163	0,000246	0,000891	0,00099
ARCH-M	—	—	8,944281	8,666405	—	—
ARCH0	2,3728E-06	1,8432E-06	2,4393E-06	1,8603E-06	2,1058E-06	1,7148E-06
ARCH1	0,14331	0,129214	0,14121	0,12788	0,102831	0,098696
GARCH1	0,852046	0,867415	0,852874	0,868192	0,859051	0,867383
ARCH-L	—	—	—	—	0,073901	0,067423
GI	—	9,608948	—	9,62822	—	10,028881

Todos os modelos ajustados e apresentados na Tabela 2 foram ajustados no programa Ox 3.3 e todos os parâmetros obtidos são significativos, evidenciando que a série de retorno apresenta mudança de escala na média do processo ao longo do tempo e efeito de assimetria na variância condicional.

As estatísticas $Q(\cdot)$ de Ljung-Box e o teste ARCH-LM, obtidas no ajuste de cada modelo, mostram que existe uma estrutura de correlação nos resíduos e que a estrutura da variância ajustada pelos modelos dados na Tabela 2 estão bem especificados.

A Tabela 3 apresenta as estatísticas de adequabilidade, citadas na seção 4, dos modelos onde verificamos que os modelos ajustados assumindo distribuição t-Student se mostram mais adequados pelos critérios AIC e BIC e pelo valor do logaritmo da função de máxima verossimilhança. Entretanto, os modelos ajustados com distribuição normal apresentam menor persistência e mais rápida recuperação após picos na série y_t.

Tabela 3 - Estatísticas de adequabilidade dos modelos

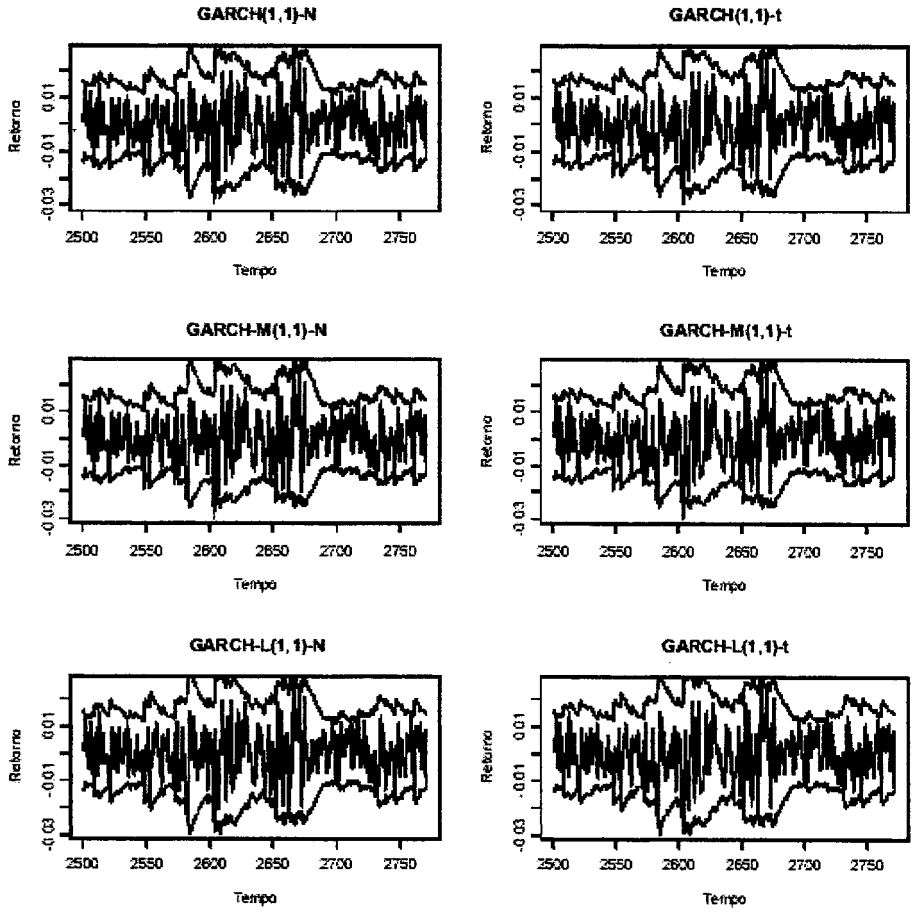
Estatísticas	GARCH(1,1)-N	GARCH(1,1)-t	GARCH-M(1,1)-N
Log MV	7580,934000	7614,318000	7594,785000
Persistência	0,995356	0,996629	0,994084
Meia-Vida	149,909656	206,273872	117,817921
AIC	-6,887213	-6,916653	-6,898895
BIC	-6,884162	-6,912992	-6,895234
MSE	0,000000	0,000000	0,000000
MAE	0,000082	0,000082	0,000081
$[LE]^2$	1,648414	1,628105	1,456677
$[LE]$	0,838143	0,834657	0,818132
Freq. no IC(95%)	0,944649	0,944649	0,933579

Estatísticas	GARCH(1,1)-N	GARCH(1,1)-t	GARCH-M(1,1)-N
Log MV	7627,517000	7590,745000	7620,931000
Persistência	0,996080	0,998833	0,999791
Meia-Vida	177,476460	594,355441	3309,231842
AIC	-6,927743	-6,895223	-6,921755
BIC	-6,923471	-6,891562	-6,917484
MSE	0,000000	0,000000	0,000000
MAE	0,000081	0,000085	0,000085
$[LE]^2$	1,546627	1,369887	1,506332
$[LE]$	0,825193	0,916657	0,831756
Freq. no IC(95%)	0,940959	0,937269	0,948340

As funções de perdas evidenciam a superioridade dos modelos ajustados assumindo distribuição t-Student e o modelo GARCH-M (1,1), com distribuição t-Student, apresenta pequena superioridade aos demais modelos, mostrando ser o modelo mais adequado para descrever as características da série de retorno.

A Figura 3 apresenta os limites inferiores e superiores estimados por cada modelo descrito na Tabela 2. Esses gráficos comprovam eficiência parecida em quase todos os modelos com pequena superioridade do modelo GARCH-M (1,1).

Figura 3 - Limites estimados



6. Conclusões

Os modelos ajustados conseguem descrever com boa eficiência a série de retorno do IBOVESPA. No ajuste desses modelos, constatamos que a série de retorno não tem distribuição normal e a isso se deve o fato da melhor adequabilidade dos modelos ajustados assumindo distribuição t-Student. Constatamos, também, que existe efeito de assimetria na variância condicional e existe mudança de escala na média do processo ao longo do tempo.

Comparando os resultados dos modelos, identificamos o modelo GARCH-M, que teve pequena superioridade aos demais, como o modelo que melhor se adequa aos dados e às características da série.

Os altos valores obtidos na persistência indicam possíveis extensões para esse trabalho que incluem entre outras opções a aplicação e comparação dos modelos de mudanças de regime.

Referências bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723.

Berndt, E. K.; Hall, B. e Hausman, J. A.(1974). Estimation and Inference in Nonlinear Structural Models. Annals of Economic and Social Measurement, 3, 653–665.

Box, G. E. P. e Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden Day.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics, 31, 307–327.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rate of return, Review of Economics and Statistics, 69, 542–547.

Diebold, X. F. (1986). Modeling the persistence of conditional variances: A comment. Econometric Reviews, 5, 51–56.

Dueker, M. J. (1997). Markov Switching in GARCH processes mean-reverting stock-market, Journal of Business & Economic Statistics, 15, 26–34.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation, Econometrica, 50, 987–1007.

Engle, R. F. e Bollerslev, T. (1986). Modeling the persistence of conditional variances, Econometric Reviews, 78, 111–125.

Engle, R. F.; Lilien, D. e Robins, R. (1987). Estimating time varying risk premia in the term structure: the ARCH-M model, Econometrica, 55, 391–407.

Hamilton, J. D. e Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime, Journal of Econometrics, 64, 307–333.

Hamilton, J. D. (1994). Time series analysis, Princeton University Press.

Luenberger, D. G. (1984). Linear and Nonlinear Programming. Addison–Wesley, Reading, Mass.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach, Econometrica, 59, 347–370.

De Almeida, N. M. C. G. (2000). Modelos de mudança de regime: uma aplicação em finanças empíricas, Dissertação de Mestrado em Economia da FEA–USP, Orientador: Pedro L. Valls Pereira.

Runkle, D. E.; Glosten, L. R. e Jagannathan, R. (1993). On the relation between the expected value and the volatility of normal excess return on stocks, Journal of Finance, 48, 1779–1801.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461–464.

Sobrinho, N. F. S. (2001). Extração da volatilidade do IBOVESPA, Resenha BM&F–nº 144. Disponível em http://www2.bmf.com.br/cimConteudo/W_ArtigosPeriodicos/004.144.pdf

Abstract

Estimating and forecasting the volatility of stock markets index, commodities and currencies are of great importance for decisions of managers involved in financial applications. We present here a review of a few parametric models of the "ARCH family" (Autoregressive Conditional Heterocedasticity) generally used for estimation of volatility. We also present a comparison of performance of these models. As illustration apply them to the IBOVESPA series after the "Plano Real" (from 02/01/1992 to 24/03/2003). The criteria used indicate satisfactory results for all models, with a small advantage for the GARCH-M (1,1) model.

Agradecimento

Os autores agradecem à CAPES pela bolsa de mestrado do primeiro autor.

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística Aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará, também, artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográficos e econométricos. Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores seja brasileiro e nem resida no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo 'duplo cego', isto é, os artigos são avaliados sem identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

O processo editorial da RBEs é eletrônico. Os artigos devem ser submetidos via *e-mail* para: helem.ortega@ibge.gov.br

Após a submissão, o autor correspondente receberá um código para acompanhar o processo de avaliação do artigo. Caso não receba um aviso com este número no prazo de uma semana, fazer contato com a secretaria da revista no endereço:

Revista Brasileira de Estatística
IBGE – Diretoria de Pesquisas - Coordenação de Métodos e Qualidade
Av. República do Chile, nº 500, 10º andar
Centro, Rio de Janeiro – RJ
CEP: 20031-170
Tel.: 55 21 2142-0472
55 21 2142-4549
Fax: 55 21 2142-4802

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais entregues para publicação devem obedecer às normas seguintes.

1. Originais processados pelo editor de texto *Word for Windows* são preferidos. Entretanto, serão aceitos também, originais processados em LaTeX desde que sejam encaminhados acompanhados de versões em pdf, conforme descrito no item 3 a seguir;

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, também, devem figurar nesta página;

3. No caso da submissão não ser em *Word for Windows*, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, LaTeX). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página, como descrito no item 2, e outro contendo apenas o título, sem identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;

4. A segunda página do original deve conter resumos em português e inglês (*abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser digitado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com, no máximo, 150 palavras;

5. O artigo deve ser dividido em seções, numeradas progressivamente, com títulos conciso e apropriado. Todas as seções e subseções devem ser numeradas e receber título apropriado;

6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

7. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;

8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções de trabalho; e

9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos.