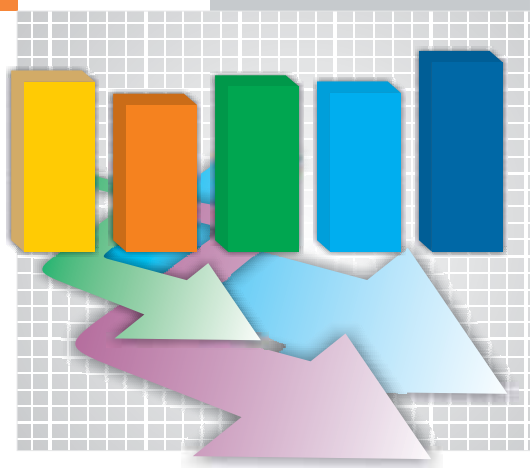
REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175



volume 74
número 239
julho / dezembro 2013

IBGE
Instituto Brasileiro de Geografia e Estatística

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 74 número 239 julho/dezembro 2013

ISSN 0034-7175

R. Bras. Estat., Rio de Janeiro, v. 74, n. 239, p. 1-133, jul./dez. 2013

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 através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.

Os originais para publicação deverão ser submetidos para o site http://rbes.submitcentral.com.br/login.php

Os artigos submetidos à 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.

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.

Gerência de Biblioteca e Acervos Especiais CDU 31(05)
RJ-IBGE/88-05 (rev.2009) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota da Editora **5**

Artigos

Estimação por máxima verossimilhança dos parâmetros da distribuição Birnbaum-Saunders usando C, Ox e R **7**

Helton Saulo
Manoel Santos Neto
Jeremias Leão
Josimar Mendes Vasconcelos

Uso de estatística vetorial para análise de lineamentos da Bacia do Rio Três Forquilhas, Nordeste do Rio Grande do Sul **21**

Andrea Valli Nummer
Carlos Alberto da Fonseca Pires
Adelir José Strieder

Um método baseado em combinação de soluções com coassociação para o problema de agrupamento automático **43**

Gustavo Silva Semaan
Cláudio de Carvalho Torres
José André de Moura Brito
Luiz Satoru Ochi

Combinação em série e em paralelo de modelos de Redes Neurais e Regressão Logística – Um estudo de caso em Cros-Selling..... **69**

Sabrina Zanatta Grebin
Lisiane Priscila Roldão Selau

Planejamento de amostras domiciliares no Brasil explorando a malha setorial do Censo Demográfico 2010 ..... **101**

Sâmela Batista Arantes
Pedro Luis do Nascimento Silva

Nota da Editora

O número 239 da RBEs publica cinco artigos em que são abordados estimação de parâmetros, Análise de Agrupamentos, Regressão Logística, Redes Neurais e cálculo de tamanhos de amostras. O artigo de Helton Saulo, Manoel Santos Neto, Jeremias Leão e Josimar Mendes Vasconcelos apresenta uma comparação por simulação do desempenho dos estimadores de máxima verossimilhança dos parâmetros da distribuição Birnbaum-Saunders, obtidos pelas linguagens de programação C, Ox e R. O artigo de autoria de Carlos Alberto da Fonseca Pires, Andrea Valli Nummer e Adelir José Strieder faz uma análise da orientação e do padrão dos lineamentos (expressão topográfica) na Bacia do Rio Três Forquilhas, RS, em que o diagrama de rosetas é usado para formar grupos. O artigo de Gustavo Silva Semaan, Cláudio de Carvalho Torres, José André de Moura Brito e Luiz Satoru Ochi propõe um método de combinação de soluções baseado na Matriz de Coassociação para o Problema de Agrupamento Automático. O artigo de Sabrina Zanatta Grebin e Lisiane Priscila Roldão Selau mostra uma comparação do desempenho da Regressão Logística e de Redes Neurais e de duas formas de combinação das duas na solução de problema de venda cruzada de uma instituição financeira com grande banco de dados. Finalmente, o artigo de autoria de Sâmela Batista Arantes e Pedro Luis do Nascimento Silva apresenta uma proposta de cálculo de tamanhos amostrais usando o Efeito do Plano Amostral em amostras complexas.

Agradeço a colaboração dos Editores Executivos Pedro Luis do Nascimento Silva (ENCE/IBGE) e Mário de Castro Andrade Filho (ICMC-USP), do Editor de Estatísticas Oficiais José André de Moura Brito (ENCE/IBGE) e do Editor de Metodologias Fernando Antonio da Silva Moura (UFRJ). Agradeço também aos Editores Associados, aos autores, IBGE, ABE, aos revisores, que anonimamente contribuíram para mais este número da Revista Brasileira de Estatística e a Marilene Pereira Piau Câmara pela editoração da revista.

Desejo a todos que tenham uma excelente leitura.

Saudações cordiais

Lúcia Pereira Barroso
Editora Responsável

Estimação por máxima verossimilhança dos parâmetros da distribuição Birnbaum-Saunders usando C, Ox e R

Helton Saulo [1]
Manoel Santos-Neto [2]
Jeremias Leão [3]
Josimar Vasconcelos[4]

Resumo

Nesse artigo realizamos uma avaliação numérica usando estimação de máxima verossimilhança para os dois parâmetros da distribuição Birnbaum-Saunders [Birnbaum, Z.W., Saunders, S.C., 1969a. A new family of life distributions. J. Appl. Probab. 6, 319-327]. Em particular, comparamos as linguagens de programação C, Ox e R através de simulações de Monte Carlo. Os resultados numéricos indicam que a linguagem C é computacionalmente menos onerosa e que as três linguagens são semelhantes em termos de estimação do modelo.

Palavras-chave: Distribuição Birnbaum-Saunders; BFGS; Linguagens C, Ox e R; Simulação de Monte Carlo.

[1] Instituto de Matemática e Estatística, Universidade Federal de Goiás, Goiânia, Go, Brasil
[2] Departamento de Estatística, Universidade Federal de Campina Grande, Campina Grande, PB, Brasil
[3] Departamento de Informática e Estatística, Universidade Federal do Piauí, Teresina, PI, Brasil.
[4] Departamento de Matemática, Universidade Federal do Piauí, Picos, PI, Brasil

1. Introdução

A distribuição Birnbaum-Saunders (*BS*) vem sendo bastante estudada nesta última década. Esta distribuição é unimodal, possui assimetria positiva, o suporte está definido nos reais não-negativos e é indexada por dois parâmetros que controlam a forma e escala da distribuição, respectivamente. Para maiores detalhes desta distribuição; ver Birnbaum & Saunders (1969 a,b) e Johnson et al. (1995, pags. 651-663). O interesse pela distribuição Birnbaum-Saunders se deve aos seus argumentos teóricos físicos, suas propriedades atrativas e sua relação com a distribuição normal. A distribuição *BS* teve sua origem na engenharia, no entanto, tem sido aplicada em outras áreas como negócios, meio ambiente e medicina, ver por exemplo, Leiva et al. (2007, 2008, 2009, 2011), Barros et al. (2008), Bhatti (2010), Vilca et al. (2011) e Paula et al. (2012).

Diferentes aspectos relacionados à estimação dos parâmetros da distribuição *BS* têm sido estudados. Em Birnbaum & Saunders (1969b) são apresentados os estimadores de máxima verossimilhança (MV) para os parâmetros desta distribuição. Bhattacharyya & Fries (1982) observam que o fato da distribuição *BS* não pertencer a família exponencial dificulta o desenvolvimento de procedimentos de inferência para os seus parâmetros. Engelhardt et al. (1981), Ahmad (1988), Achcar (1993), Chang & Tang (1994) e Dupuis & Mills (1998) discutem outros estimadores para os referidos parâmetros. No entanto, em todos os casos, não é possível obter expressões explícitas para estes estimadores sendo necessário a utilização de procedimentos numéricos. Podemos, também, encontrar estudos computacionais dos estimadores de máxima verossimilhança dos parâmetros da distribuição *BS* em Cysneiros et al. (2008), Lemonte et al. (2007), Lemonte et al. (2008) e Santos-Neto et al. (2012).

O objetivo desse artigo é estudar o comportamento das estimativas de MV, dos parâmetros que indexam a distribuição *BS*, quando obtidas por diferentes linguagens de programação. Neste artigo, foram utilizadas as linguagens de programação C, Ox e R. A obtenção de estimativas de MV de forma eficiente e prática é de grande importância principalmente pela grande aplicabilidade dessa distribuição. Os estimadores de MV da distribuição *BS* são encontrados usando o método de otimização não-linear Broyden-Fletcher-Goldfarb-Shanno (BFGS). Detalhes sobre esse método podem ser encontrados em Nocedal & Wright (2006).

A linguagem C foi desenvolvida por Dennis Ritchie entre 1969 e 1973 e pertence a uma família de linguagens que, dentre outras características, possuí geração de código eficiente, confiabilidade, regularidade, simplicidade e facilidade de uso; ver Kernighan & Ritchie (1978). A linguagem Ox foi criada por Jurgen Doornik em 1994. Ela é bastante flexível e foi desenvolvida com base na linguagem C; ver Cribari-Neto & Zarkos (2003) e Doornik (2001). Por fim, a linguagem R foi criada por Ross Ihaka e Robert Gentleman e caracteriza-se pela flexibilidade oferecida por algumas linguagens compiladas (e.g., C e C++) e a praticidade dos tradicionais pacotes estatísticos; ver Cribari-Neto & Zarkos (1999) and Crawley (2007).

O artigo está organizado da seguinte forma. A Seção 2 introduz a distribuição Birnbaum-Saunders e algumas de suas propriedades. Na Seção 3 apresentamos como são obtidos os estimadores de máxima verossimilhança através do método de otimização não-linear BFGS. Na Seção 4 estudamos a performance dos estimadores de MV através de simulações de Monte Carlo nas linguagens de programação C, Ox e R. Finalmente, na Seção 5 discutimos algumas das conclusões deste artigo.

2. Distribuição Birnbaum-Saunders

Seja $M = \{f_T(t;\theta) : \theta \in \Theta\}$, um modelo estatístico, em que T é uma variável aleatória que pertence ao espaço amostral T e $f_T(t;\theta)$ é a função densidade de probabilidade (f.d.p) parametrizada por θ, em relação a alguma medida dominante comum P em T. Podemos considerar, $\theta = (\theta_1 \cdots, \theta_p)^T \in \Theta \subset \Re^p$. Por exemplo, o modelo *BS* é uma família de distribuições de probabilidade possuindo a seguinte f.d.p

$$f_T(t;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\alpha^2}\left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right) \frac{[t+\beta]}{2\alpha\sqrt{\beta t^3}}, \qquad (1)$$

em que o espaço amostral $T = (0,\infty) \bigcap \Re^1$, com a medida de Lebesgue $dP = dt$ e $\theta = (\alpha, \beta)^T$ é um vetor bi-dimensional. Em particular, α é um parâmetro de forma, ao passo que β representa um parâmetro de escala. Note que o espaço paramétrico Θ é um semi-plano,

$$\Theta = \{(\alpha,\beta): \alpha > 0, \beta > 0\}$$

Desta maneira, o conjunto M é composto por todas as distribuições BS, e cada distribuição é especificada pelo vetor $\theta = (\alpha, \beta)^{\mathrm{T}}$. Adicionalmente, temos que a distribuição BS possui as seguintes propriedades: $(\mathrm{P1})\ \mathrm{c\,T} \sim \mathrm{BS}(\alpha, \mathrm{c}\,\beta)$, com $\mathrm{c} > 0$, e $(\mathrm{P2})\ 1/\mathrm{T} \sim \mathrm{BS}(\alpha, 1/\beta)$. O *k*-ésimo momento de T é

$$\mathrm{E}[\mathrm{T}^{\mathrm{k}}] = \beta^{\mathrm{k}} \sum_{\mathrm{j}=0}^{\mathrm{k}} \binom{2\mathrm{k}}{2\mathrm{i}} \sum_{\mathrm{i}=0}^{\mathrm{j}} \binom{\mathrm{j}}{\mathrm{i}} \frac{(2\mathrm{k} - 2\mathrm{j} + 2\mathrm{i})!}{2^{\mathrm{k}-\mathrm{j}+\mathrm{i}}(\mathrm{k} - \mathrm{j} + \mathrm{i})!} \left[\frac{\mathrm{a}}{2}\right]^{2\mathrm{k}-2\mathrm{j}+2\mathrm{i}}$$

e desta forma $\mathrm{E}[\mathrm{T}^{\mathrm{K}}] = \beta\left[1 + \frac{1}{2}\alpha^2\right]$ e $\mathrm{Var}[\mathrm{T}] = [\beta\alpha]^2\left[1 + \frac{5}{4}\alpha^2\right]$. A função quantílica de T é $\mathrm{Q_{BS}}(\mathrm{p}) = [\beta/4]\left[\alpha \mathrm{Q_N}(\mathrm{p}) + \sqrt{\mathrm{a}^2 \mathrm{Q_N}(\mathrm{p})^2 + 4}\right]^2$, em que $\mathrm{Q_N}(\mathrm{p})$ é o p-ésimo quantil de $\mathrm{Z} \sim \mathrm{N}(0,1)$ em que $\mathrm{N}(0,1)$ denota a distribuição normal padrão. Note que $\mathrm{Q_{BS}}(0.5) = \beta$, isto é, β é a mediana da distribuição *BS*.

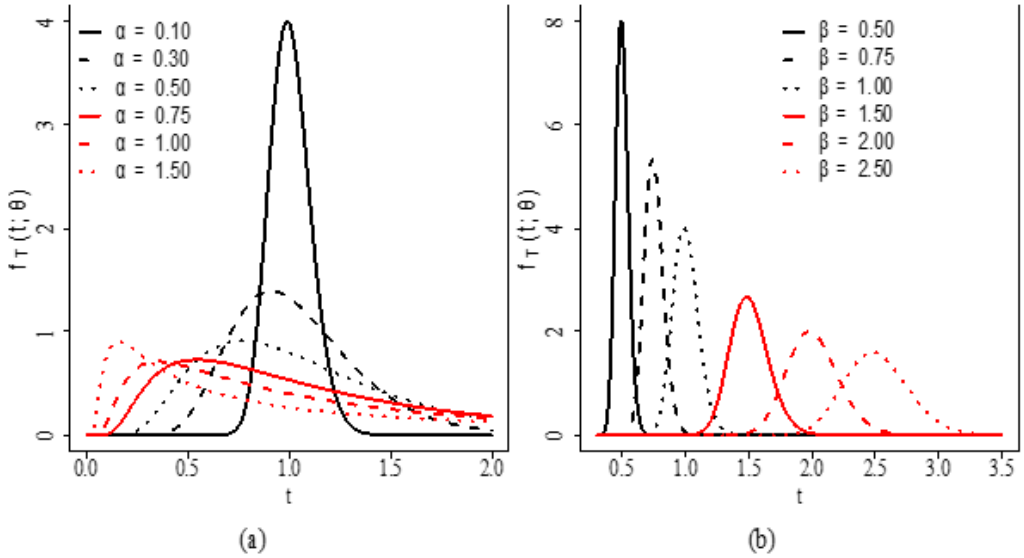
Na Figura 1(a) é possível observamos que o incremento no valor do parâmetro α faz com que o grau de assimetria da distribuição BS aumente, além disso, temos o aumento na variabilidade e no grau de achatamento da distribuição. Já na Figura 1(b) temos que o incremento no valor do parâmetro β faz com que a densidade se desloque no mesmo sentido de β e aumente a variabilidade da distribuição. Estes resultados reforçam o fato dos parâmetros α e β controlarem a forma e escala da distribuição *BS*, respectivamente.

3. Estimação de máxima verossimilhança usando BFGS

Seja $\{t_i : i = 1, \cdots, n\}$ numa amostra aleatória de uma distribuição BS(α, β). Estão a função de log-verossimilhança $\ell(\theta; t)$ desse modelo pode ser representada por

$$\ell(\theta; t) \;\; \alpha - n\log(\alpha) - \frac{n}{2}\log(\beta) + \sum_{i=1}^{n}\log(t_i + \beta) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}\left[\frac{t_i}{\beta} + \frac{\beta}{t_i} - 2\right] \qquad (2)$$

Figura 1: Gráficos da f.d.p da *BS* para $\beta = 1.0(a)$ e $\alpha = 0.1\,(b)$



Os elementos do vetor escore, $U\;(\theta) = \left(U_\alpha(\theta), U_\beta(\theta)\right)^T$, são:

$$U_\alpha(\theta) \;=\; \frac{\partial\ell(\alpha,\beta)}{\partial\alpha} = -\frac{n}{\alpha}\left[1 + \frac{2}{\alpha^2}\right] + \frac{1}{\alpha^3\beta}\sum_{i=1}^{n} t_i + \frac{\beta}{\alpha^3}\sum_{i=1}^{n}\frac{1}{t_i}$$

e

$$U_\beta(\theta) \;=\; \frac{\partial\ell(\alpha,\beta)}{\partial\beta} = -\frac{n}{2\beta} + \sum_{i=1}^{n}\frac{1}{t_i + \beta} + \frac{1}{2\alpha^2\beta^2}\sum_{i=1}^{n} t_i - \frac{1}{2\alpha^2}\sum_{i=1}^{n}\frac{1}{t_i}$$

Birnbaum & Saunders (1969b) mostraram que o estimador de máxima verossimilhança (MV) de α é :

$$\hat{\alpha} = \left[\frac{s}{\hat{\beta}} + \frac{\hat{\beta}}{r} - 2 \right]^{1/2} \tag{3}$$

em que $r = n^{-1} \sum_{i=1}^{n} t_i$ e $s = \left[n^{-1} \sum_{i=1}^{n} t_i^{-1} \right]^{-1}$. Note que para Obter $\hat{\beta}$, precisamos resolver uma equação não-linear em β, ou seja:

$$\beta^2 - \beta \left[2r + K(\beta) \right] + r \left[s + K(\beta) \right] = 0 , \tag{4}$$

em que $K(\xi) = \left(\frac{1}{n} \sum_{i=1}^{n} \left[\xi + t_i \right]^{-1} \right)^{-1}$, para $\xi > 0$. Birnbaum & Saunders (1969b) apresentaram dois métodos para encontrar $\hat{\beta}$. O primeiro funciona bem quando $\alpha < 0.5$, no entanto, não quando $\alpha > 2$. O segundo método também não funciona bem para alguns valores de α. Neste artigo, encontramos os estimadores de MV de α e β pela maximização da função de log-verossimilhança usando o método de otimização não-linear BFGS com derivadas analíticas. Este método é considerado como o algoritmo de otimização não-linear mais confiável; ver (Mittelhammer et al., 2000, p. 199). Os estimadores de momentos modificados $\alpha_0 = \left\{ 2 \left(\left[\frac{s}{r} \right]^{1/2} - 1 \right) \right\}^{1/2}$ e $\beta_0 = [rs]^{1/2}$, definidos em Ng et al. (2003), são usados como chutes iniciais. O algoritmo BFGS usa o mesmo princípio do método de Newton-Raphson, exceto por utilizar uma sequência de matrizes simétricas e positivas definidas $B^{(k)}$ em alternativa à matriz $-H^{-1}$, de maneira que

$$\lim_{k \to \infty} B^{(k)} = -H^{-1} \tag{5}$$

Em geral, a matriz identidade de mesma ordem $B^{(k)}$ é tomada como a matriz inicial, pois ela é positiva definida e simétrica, conduzindo assim a aproximações $B^{(k)}$ positivas definidas e simétricas. A Forma recursiva para tais matrizes é dada por

$$B^{(k+1)} = B^{(k)} - \frac{B^{(k)} g^{(k)} \left(g^{(k)}\right)^T B^{(k)}}{\left(g^{(k)}\right)^T B^{(k)} g^{(k)}} + \frac{h^{(K)} \left(h^{(k)}\right)^T}{\left(h^{(k)}\right)^T g^{(k)}}, \quad K = 0, 1, \ldots, \tag{6}$$

Em que $g^{(k)} = \theta^{(k+1)} - \theta^{(k)}$ e $h^{(k)} = U\left(\theta^{(k+1)}\right) - U\left(\theta^{(k)}\right)$. Desse modo, o máximo é obtido por

$$\theta^{(k+1)} = \theta^{(k)} - \lambda^{(k)} B^{(k)} U\left(\theta^{(k)}\right), \quad K = 0, 1, \ldots, \tag{7}$$

Em que $\lambda^{(k)}$ é um escalar determinado por um procedimento de busca linear de $\theta^{(k)}$ na direção $-B^{(k)}U\left(\theta^{(k)}\right)$

4. Simulações de Monte Carlo

Nesta seção, apresentamos a avaliação numérica dos estimadores de MV dos parâmetros do modelo BS usando o método BFGS nas linguagens de programação C, Ox e R. Se T é uma variável aleatória com distribuição BS, então temos que

$$X = \frac{1}{2}\left[\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right] \tag{8}$$

é uma distribuição $N\left(0, \frac{\alpha^2}{4}\right)$. Logo, a partir da equação (8) podemos escrever T em função da distribuição normal

$$T = \beta\left\{1 + 2X^2 + 2X\left(1 + X^2\right)^{1/2}\right\} \tag{9}$$

Assim, temos que os números pseudo-aleatórios da variável aleatória T podem ser obtidos a partir de números pseudo-aleatórios da distribuição normal da forma como é dado em (9). Consideramos os seguintes tamanhos amostrais $n = 10, 30, 60 \text{ e } 100$, e os

valores assumidos para o parâmetro de forma foram $\alpha = 0.10, 0.50, 0.75 \text{ e } 1.00$. O parâmetro de escala, sem perda de generalidade, é fixado em 1.0, e admitimos 10000 réplicas de Monte Carlo em todos os experimentos.

As versões dos programas (em plataforma Linux Ubuntu 12.04.1) usadas para implementar e realizar as simulações foram: R 2.13.1, usando o pacote gbs, desenvolvido por Barros et al. (2009), disponível em www.R-project.org; Ox 6.20, disponível em www.doornik.com; e gcc 4.6 para C, usando a biblioteca GSL a qual fornece diversas rotinas matemáticas e estatísticas e pode ser obtida em www.gnu.org/software/gsl/. Os números pseudo-aleatórios foram obtidos através do gerador de George-Marsaglia.

A Tabela 1 mostra que os valores das estimativas para o parâmetro α é menor que o valor fixado como verdadeiro em todos os cenários, ou seja, em geral temos que os verdadeiros valores do parâmetro α são subestimados. Notem que o viés de estimação em valor absoluto aumenta à medida que aumentamos o valor de α, contudo para um mesmo valor de α à medida que aumentamos o tamanho amostral observa-se uma redução de viés (vide Tabela 3). Por exemplo, quando $n = 10$ e , os vieses absolutos relativos para α obtidos por C, Ox e R, eram 0.0730, 0.0771 e 0.0760, no entanto, quando $\alpha = 1.00$, as medidas foram 0.0809, 0.0842 e 0.0842, respectivamente. Esse comportamento é semelhante, também, para as medidas de dispersão apresentadas nas Tabelas 2 e 4.

Note também que, para $n = 10$, as estimativas dos vieses (em valor absoluto) para os parâmetros $\alpha \text{ e } \beta$ obtidos a partir da linguagem C são menores do que os obtidos com as linguagens Ox e R. É interessante notar que comparando-se a variabilidade das estimativas dos parâmetros em relação as linguagens C, Ox e R nota-se que, em geral, temos níveis de variabilidades no mesmo patamar para as três linguagens. Na Tabela 5 mostramos o tempo de execução em segundos das simulações de Monte Carlo. Notamos que os tempos de execução obtidos quando simulamos no R são muito maiores do que as outras linguagens, e que a linguagem C sobressai-se sobre as demais.

Tabela 1: Estimativas de MV para os diferentes valores de α, baseadas em simulações de Monte Carlo $(\beta = 1.00)$

		Estimativas de α				Estimativas de β			
α		$n = 10$	$n = 30$	$n = 60$	$n = 100$	$n = 10$	$n = 30$	$n = 60$	$n = 100$
0.10	C	0.0222	0.0128	0.0090	0.0070	0.0314	0.0185	0.0128	0.0100
	Ox	0.0221	0.0129	0.0091	0.0071	0.0314	0.0183	0.0129	0.0100
	R	0.0221	0.0129	0.0092	0.0071	0.0317	0.0183	0.0129	0.0100
0.50	C	0.1106	0.0638	0.0451	0.0352	0.1551	0.0901	0.0624	0.0486
	Ox	0.1094	0.0646	0.0454	0.0349	0.1542	0.0889	0.0626	0.0484
	R	0.1104	0.0645	0.0459	0.0354	0.1567	0.0895	0.0625	0.0485
0.75	C	0.1657	0.0957	0.0676	0.0528	0.2286	0.1309	0.0903	0.0703
	Ox	0.1676	0.0963	0.0686	0.0532	0.2280	0.1292	0.0908	0.0700
	R	0.1655	0.0967	0.0688	0.0532	0.2314	0.1301	0.0905	0.0700
1.00	C	0.2209	0.1277	0.0902	0.0704	0.2976	0.1670	0.1145	0.0891
	Ox	0.2196	0.1279	0.0901	0.0708	0.3064	0.1627	0.1146	0.0890
	R	0.2208	0.1289	0.0918	0.0709	0.3021	0.1664	0.1151	0.0888

Tabela 2: Estimativas dos desvios-padrão, baseadas nas simulações de Monte Carlo $(\beta = 1.00)$

		Estimativas de α				Estimativas de β			
α		$n = 10$	$n = 30$	$n = 60$	$n = 100$	$n = 10$	$n = 30$	$n = 60$	$n = 100$
0.10	C	0.0222	0.0128	0.0090	0.0070	0.0314	0.0185	0.0128	0.0100
	Ox	0.0221	0.0129	0.0091	0.0071	0.0314	0.0183	0.0129	0.0100
	R	0.0221	0.0129	0.0092	0.0071	0.0317	0.0183	0.0129	0.0100
0.50	C	0.1106	0.0638	0.0451	0.0352	0.1551	0.0901	0.0624	0.0486
	Ox	0.1094	0.0646	0.0454	0.0349	0.1542	0.0889	0.0626	0.0484
	R	0.1104	0.0645	0.0459	0.0354	0.1567	0.0895	0.0625	0.0485
0.75	C	0.1657	0.0957	0.0676	0.0528	0.2286	0.1309	0.0903	0.0703
	Ox	0.1676	0.0963	0.0686	0.0532	0.2280	0.1292	0.0908	0.0700
	R	0.1655	0.0967	0.0688	0.0532	0.2314	0.1301	0.0905	0.0700
1.00	C	0.2209	0.1277	0.0902	0.0704	0.2976	0.1670	0.1145	0.0891
	Ox	0.2196	0.1279	0.0901	0.0708	0.3064	0.1627	0.1146	0.0890
	R	0.2208	0.1289	0.0918	0.0709	0.3021	0.1664	0.1151	0.0888

Tabela 3: Estimativas dos viés relativo, baseadas nas simulações de Monte Carlo $(\beta = 1.00)$

		Estimativas de α				Estimativas de β			
α		$n = 10$	$n = 30$	$n = 60$	$n = 100$	$n = 10$	$n = 30$	$n = 60$	$n = 100$
0.10	C	-0.0730	-0.0269	-0.0126	-0.0076	0.0003	0.0004	0.0001	0.0001
	Ox	-0.0771	-0.0275	-0.0122	-0.0079	0.0004	0.0003	0.0000	0.0000
	R	-0.0760	-0.0230	-0.0120	-0.0070	0.0006	0.0004	0.0002	-0.0001
0.50	C	-0.0755	-0.0278	-0.0131	-0.0079	0.0107	0.0049	0.0020	0.0013
	Ox	-0.0803	-0.0273	-0.0141	-0.0100	0.0100	0.0050	0.0019	0.0010
	R	-0.0788	-0.0242	-0.0124	-0.0072	0.0125	0.0052	0.0024	0.0007
0.75	C	-0.0781	-0.0286	-0.0135	-0.0082	0.0231	0.0098	0.0040	0.0026
	Ox	-0.0826	-0.0280	-0.0127	-0.0082	0.0234	0.0070	0.0035	0.0039
	R	-0.0815	-0.0251	-0.0128	-0.0075	0.0274	0.0109	0.0054	0.0023
1.00	C	-0.0809	-0.0296	-0.0139	-0.0085	0.0382	0.0153	0.0063	0.0041
	Ox	-0.0842	-0.0283	-0.0146	-0.0086	0.0459	0.0117	0.0056	0.0026
	R	-0.0842	-0.0260	-0.0133	-0.0078	0.0468	0.0184	0.0095	0.0049

Tabela 4: Estimativas do $\sqrt{\text{EQM}}$ **, baseadas nas simulações de Monte Carlo** $(\beta = 1.00)$

		Estimativas de α				Estimativas de β			
α		$n = 10$	$n = 30$	$n = 60$	$n = 100$	$n = 10$	$n = 30$	$n = 60$	$n = 100$
0.10	C	0.0233	0.0130	0.0091	0.0071	0.0314	0.0185	0.0128	0.0100
	Ox	0.0234	0.0132	0.0092	0.0071	0.0314	0.0183	0.0129	0.0100
	R	0.0224	0.0141	0.0100	0.0100	0.0316	0.0173	0.0141	0.0100
0.50	C	0.1168	0.0653	0.0456	0.0354	0.1555	0.0902	0.0624	0.0486
	Ox	0.1166	0.0660	0.0460	0.0353	0.1545	0.0890	0.0626	0.0484
	R	0.1170	0.0656	0.0458	0.0361	0.1572	0.0894	0.0624	0.0480
0.75	C	0.1757	0.0981	0.0684	0.0531	0.2298	0.1313	0.0904	0.0703
	Ox	0.1787	0.0985	0.0693	0.0536	0.2292	0.1294	0.0909	0.0701
	R	0.1764	0.0985	0.0693	0.0539	0.2330	0.1304	0.0906	0.0700
1.00	C	0.2353	0.1311	0.0913	0.0709	0.3000	0.1677	0.1146	0.0892
	Ox	0.2352	0.1310	0.0913	0.0714	0.3098	0.1631	0.1148	0.0890
	R	0.2362	0.1315	0.0927	0.0714	0.3058	0.1673	0.1153	0.0889

Tabela 5: Estimativas de execução (em segundos) das simulações de Monte Carlo $(\beta = 1.00)$

		Estimativas de α			
α		$n = 10$	$n = 30$	$n = 60$	$n = 100$
	C	0.26	0.90	2.13	4.13
0.10	Ox	1.28	2.46	4.25	7.64
	R	44.58	48.49	52.25	57.25
	C	0.24	0.83	2.01	3.96
0.50	Ox	2.34	3.90	6.24	9.42
	R	55.34	60.55	65.67	71.59
	C	0.18	0.67	1.65	3.30
0.75	Ox	2.40	3.84	6.03	9.06
	R	58.27	66.14	71.50	80.67
	C	0.16	0.56	1.35	2.72
1.00	Ox	2.38	3.74	5.99	8.70
	R	56.75	61.82	66.71	72.74

5. Considerações Finais

Neste artigo, foi realizada uma avaliação numérica dos estimadores de máxima verossimilhança dos parâmetros da distribuição BS através das linguagens C, Ox e R via simulações de Monte Carlos. Para avaliarmos o desempenho dos estimadores de MV foram utilizadas as seguintes quantidades: média, desvios-padrão, viés relativo e a raiz quadrada do erro quadrático médio $\sqrt{\text{EQM}}$. Notamos que as estimativas de MV dos parâmetros da distribuição BS obtidas pelas linguagens C, Ox e R tiveram comportamento muito parecidos. Sendo assim irrelevante a escolha da linguagem quanto à qualidade da estimação. Além disso, observamos que a linguagem C é computacionalmente mais rápida, independentemente do tamanho amostral, do que as demais linguagens aqui utilizadas, comprovando que ela é bastante eficiente. No entanto, uma desvantagem da linguagem C com relação as linguagens Ox e R é o número muito pequeno de ferramentas estatísticas disponíveis, o que torna o trabalho do pesquisador mais árduo no momento da programação. Desta forma concluímos que a escolha da linguagem a ser utilizada dependerá da habilidade computacional do usuário e da complexidade do estudo a ser realizado.

Referências bibliográficas

Achcar, J. A. (1993) Inference for the Birnbaum-Saunders fatigue life model using Bayesian methods. Comp. Stat. Data Anal., 15, 367–380.

Ahmad, L. A. (1988) Jackknife estimation for a family of life distributions. J. Stat. Comp.Simul., 29, 211–223.

Barros, M., Paula, G.A., Leiva, V. (2008) A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. Lifetime Data Anal., 14, 316–332.

Barros, M., Paula, G.A., Leiva, V. (2009). An R implementation for generalized Birnbaum- Saunders distributions. Comp. Stat. Data Anal,. 53, 1511–1528.

Bhatti, C.R. (2010) The Birnbaum-Saunders autoregressive conditional duration model. Math. Comp. Simul., 80, 2062–2078.

Bhattacharyya, G. K., Fries, A. (1982) Fatigue failure models: Birnbaum-Saunders versus inverse Gaussian. IEEE Trans. Rel., 31, 439–440.

Birnbaum, Z.W., Saunders, S.C. (1969a) A new family of life distributions. J. Appl. Prob., 6, 319–327.

Birnbaum, Z.W., Saunders, S.C. (1969b) Estimation for a family of life distributions with applications to fatigue. J. Appl. Prob., 6, 328–347.

Chang, D.S., Tang, L.C. (1994) Graphical analysis for Birnbaum-Saunders distribution. Microelect. Rel., 34, 17–22.

Crawley, M.J. (2007) The R Book. Wiley, West Sussex.

Cribari-Neto, F., Zarkos, S.G. (1999) R: yet another econometric programming environment. J. Appl. Econ., 14, 319–329.

Cribari-Neto, F., Zarkos, S.G. (2003) Econometric and statistical computing using Ox. Comput. Econ., 21, 277–295.

Cysneiros, A.H.M.A., Cribari-Neto, F., Araujo, C.G.J. (2008) On Birnbaum-Saunders infe- rence. Comp. Stat. Data Anal., 52, 4939–4950.

Doornik, J.A. (2001). Ox: an Object-Oriented Matrix Language. fourth ed. Timberlake Consultants Press, London, Oxford. http://www.doornik.com.

Dupuis, D.J., Mills, J.E. (1998) Robust estimation of the Birnbaum-Saunders distribution. IEEE Trans. Rel., 47, 88–95.

Engelhardt, M., Bain, L.J., Wright, F.T. (1981) Inferences on the parameters of the Birnbaum- Saunders fatigue life distribution based on maximum likelihood estimation. Technometrics, 23, 251–256.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1995) Continuous Univariate Distributions: Vol. 2. Wiley, New York.

Kernighan, B.W., Ritchie, D.M. (1978). The C Programming Language, first ed. Prentice-Hall, Englewood Cliffs, New Jersey.

Leiva, V., Barros, M., Paula, G.A., Galea, M. (2007) Influence diagnostics in log Birnbaum- Saunders regression models with censored data. Comp. Stat. Data Anal., 51, 5694–5707.

Leiva, V., Barros, M., Paula, G.A., Sanhueza, A. (2008) Generalized Birnbaum-Saunders distributions applied to air pollutant concentration. Environmetrics, 19, 235–249.

Leiva, V., Sanhueza, A., Angulo, J.M. (2009) A length-biased version of the Birnbaum- Saunders distribution with application in water quality. Stoch. Environ. Res. Risk Assess., 23, 299–307.

Leiva, V., Athayde, E., Azevedo, C., Marchant, C. (2011) Modeling wind energy flux by a Birnbaum-Saunders distribution with unknown shift parameter. J. Appl. Stat., 38, 2819–2838.

Lemonte, A., Cribari-Neto, F., Vasconcellos, K.L.P. (2007) Improved statistical inference for the two-parameter Birnbaum-Saunders distribution. Comp. Stat. Data Anal., 51, 4656–4681.

Lemonte, A., Simas, A., Cribari-Neto, F. (2008) Bootstrap-based improved estimators for the two-parameter Birnbaum-Saunders distribution. J. Stat. Comp. Simul., 78, 37–49.

Mittelhammer, R.C., Judge, G.G., Miller, D.J. (2000). Econometric Foundations. Cambridge University Press, New York.

Ng, H.K., Kundu D., Balakrishnan N. (2003). Modified moment estimation for the two- parameter Birnbaum-Saunders distribution. Comp. Stat. Data Anal, 43, 283-298.

Nocedal, J., Wright, S.J. (2006). Numerical Optimization. Springer-Verlag, New York.

Paula, G.A., Leiva, V., Barros, M., Liu, S. (2012) Robust statistical modeling using the Birnbaum-Saunders-t distribution applied to insurance. Appl. Stoch. Model Bus. Ind., 28, 16–34.

Santos-Neto, M., Cysneiros, F.J.A. Leiva, V., Ahmed, S.E. (2012) On new parameterizations of the Birnbaum-Saunders distribution. Pak. J. Statist., 28, 1–26.

Vilca, F., Santana, L., Leiva, V., Balakrishnan, N. (2011) Estimation of extreme percentiles in Birnbaum-Saunders distributions. Comp. Stat. Data Anal., 55, 1665–1678.

Abstract

We perform a numerical evaluation using maximum likelihood estimation for the two-parameter Birnbaum-Saunders distribution [Birnbaum, Z.W., Saunders, S.C., 1969a. A new family of life distributions. J. Appl. Probab. 6, 319-327]. In particular, we compare the C, Ox and R programming languages through Monte Carlo simulations. The numerical results in- dicate that the C language is computationally less costly and that the three languages are similar in terms of model estimation.

Keywords: Birnbaum-Saunders distribution; BFGS; C, Ox and R languages; Monte Carlo simulations.

Uso de estatística vetorial para análise de lineamentos da Bacia do Rio Três Forquilhas, Nordeste do Rio Grande do Sul

Andrea Valli Nummer [1]
Carlos Alberto F. Pires [2]
Adelir José Strieder [3]

Resumo

A bacia do Rio Três Forquilhas localiza-se no nordeste do Estado do Rio Grande do Sul e abrange os municípios de Terra de Areia, Itati e parte de São Francisco de Paula. Ao longo do trecho do rio Três Forquilhas, encontra-se parte da rodovia RS 486, conhecida como Rota do Sol. A geologia da região é composta de rochas vulcânicas da Fm. Serra Geral e arenitos da Formação Botucatu. Este trabalho apresenta uma análise da orientação e do padrão dos lineamentos que ocorrem na Bacia do Rio Três Forquilhas, cujo estudo foi realizado por meio de estatística vetorial e geométrica que poderá servir como subsídio, juntamente com o mapeamento estrutural, a análise geomorfológica e das litologias, para prever o comportamento das águas subterrâneas nesta região. Na análise que se fez, considerou-se o conceito de lineamento de Strieder & Amaro (1997), sendo identificados os lineamentos Tipo 2 (dois) correspondentes à zonas de fraturas. O comprimento e a atitude dos lineamentos foram armazenados num banco de dados do Software Excel. Com os dados relativos à direção dos lineamentos, foi construído um diagrama de rosetas para separar as principais Famílias de estruturas. Com os dados relativos aos comprimentos, fez-se uma avaliação estatísitca vetorial dos lineamentos conforme (Curray,1956 e Pincus,1956, citados por Cunha, 1996), que teve como objetivo caracterizar e delimitar cada um dos conjuntos azimutais que ocorrem na área. Para a área da bacia, foram extraídos 862 lineamentos separados em seis Famílias diferentes: F1:355°-20°; F2:21°-55°; F3:56°-85°; F4:86°-120°; F5: 121°-150°; F6: 151°-155° sendo que, nas Famílias 3, 4 e 2, tem-se o maior número de descontinuidades. Os histogramas e os diagramas de dispersão elaborados para cada Família demonstram a existência de falhas principais (maior comprimento) e subsidiárias (menores). A análise geométrica confirma uma padrão de ramificações constituído de fraturas principais (mais longas) conectadas a fraturas subsidiárias de menor comprimento. Com base nas avaliações realizadas e considerando o campo tensional definido por Reginato (2003) para rochas nas proximidades, é possível estabelecer que as fraturas subsidiárias encontradas na Bacia do Rio Três Forquilhas são transtrativas (têm componente de tração) e que, se forem abertas e não estiverem preenchidas por veios e diques, têm maior capacidade de percolação de água que as demais.

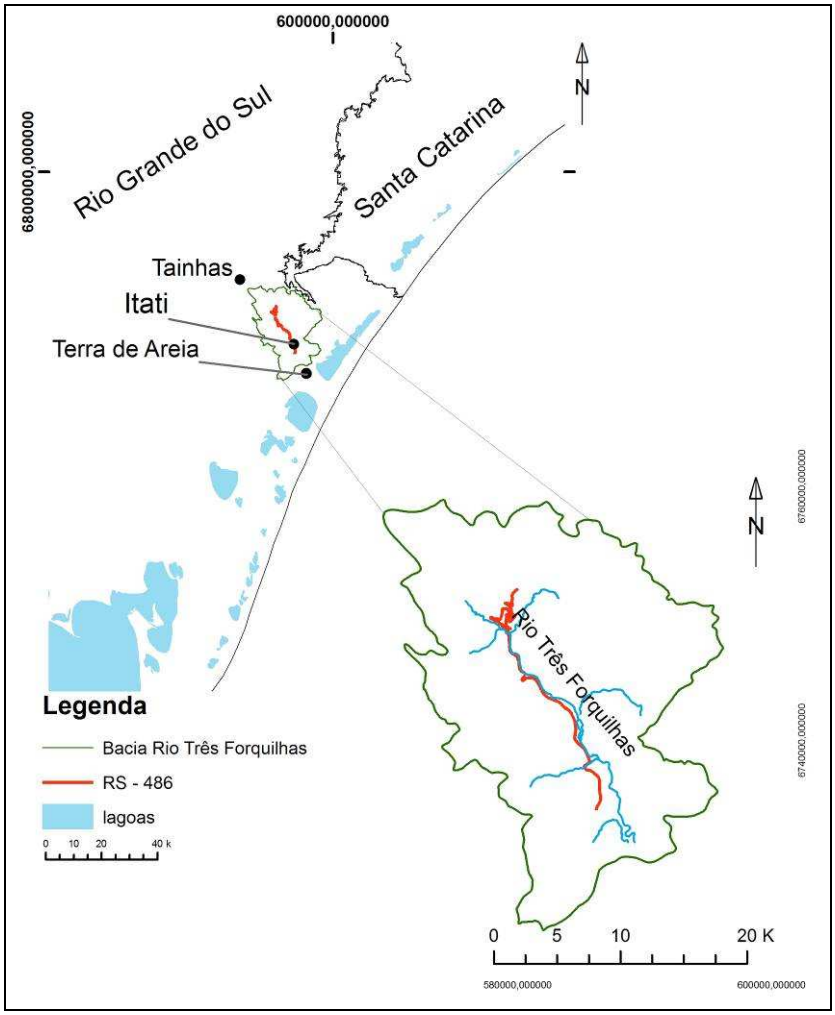
[1] Departamento de Geociências - Universidade Federal de Santa Maria – RS, e mail: a.nummer@gmail.com
[2] Departamento de Geociências - Universidade Federal de Santa Maria – RS, e mail: calpires@terra.com.br
[3] Centro de Des. Tecnológico – Universidade Federal de Pelotas-RS, e-mail: delirstrieder@uol.com.br

1. Introdução

A bacia do Rio Três Forquilhas localiza-se no nordeste do Estado do Rio Grande do Sul e abrange os municípios de Terra de Areia, Itati e o distrito de Tainhas, pertencente ao município de São Francisco de Paula (Figura 1). Ao longo do trecho do rio Três Forquilhas, encontra-se parte da rodovia RS 486, conhecida como Rota do Sol.

A rodovia RS 486 liga o município de Terra de Areia, na altura do km 46 da BR 101, ao município de Estrela, nas proximidades do Km 350 da BR 386. A Rota do Sol foi construída para facilitar a integração da Serra Gaúcha com o Litoral Norte, ligando cidades importantes como Caxias do Sul, Bento Gonçalves a BR 101 e permitir o encurtamento das rotas de transporte de cargas pesadas, provenientes dos complexos industriais localizados na serra.

Figura 1. Localização da Bacia do rio Três Forquilhas.



O trecho da RS 486 situado entre o distrito de Tainhas e o município de Terra de Areia (BR-101), no vale do rio Três Forquilhas, atravessa um pacote de rochas vulcânicas ácidas e básicas da Fm. Serra Geral e arenitos da Formação Botucatu, apresentando, ademais, inúmeros processos superficiais de movimentos de massa. Segundo Nummer (2003), os movimentos de massa mais comuns, que ocorrem na região, são os rastejos, ligados a coberturas coluvionares, deslizamentos planares, quedas e tombamentos de blocos relacionados às estruturas tectônicas e atectônicas que afetam as rochas vulcânicas.

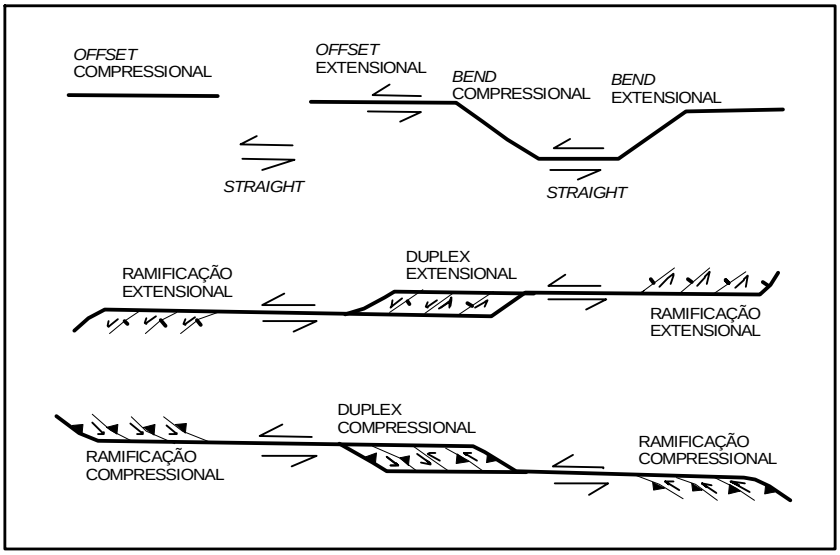
Um maciço rochoso pode ser submetido a diversos tipos de tensões (compressivas ou distensivas), em diferentes direções. O tipo de tensão atuante sobre o maciço determinará a orientação das estruturas deformacionais, e a diferença de magnitude entre as direções de tensão implicará no padrão geométrico destas fraturas. Para Ramsay e Huber (1987), fraturas são descontinuidades disjuntivas presentes nas rochas que podem ser designadas de juntas ou falhas. A distinção entre elas dá-se pela magnitude do deslocamento relativo dos blocos isolados: as falhas são fraturas que mostram um deslocamento superior a 5mm, enquanto, nas juntas, o deslocamento dos blocos é inferior a 5 mm.

Segundo Hobbs et al. (1976), se a rocha for submetida a uma pequena magnitude de tensão e a baixa pressão confinante, haverá a formação de pequenas fraturas denominadas de trativas que são paralelas à direção de compressão. À medida que aumentam as tensões aplicadas e/ou a temperatura, as rochas podem desenvolver um par conjugado de fraturas denominado de fraturas de cisalhamento.

Para Woodcock e Fischer (1986), em zonas de cisalhamento transcorrente desenvolvem-se os chamados duplexes que são arranjos de falhas imbricadas. (Figura 2). Em conformidade com os autores, straights são segmentos lineares subparalelos ao vetor de deslocamento regional, e bends são segmentos oblíquos a esse vetor. Os deslocamentos podem ser transferidos entre duas falhas por meio de um offset. Os duplexes compressionais podem se formar em bends ou offsets comprimidos e, de forma análoga, a abertura de espaços vazios correspondem aos duplexes extensionais. Os duplexes são normalmente limitados por duas zonas de falhas principais contínuas e, entre estas zonas ocorrem, falhas en échelon que completam sua estrutura. Nas

extremidades das falhas principais, formam-se as ramificações pinadas, com a mesma orientação e sentido de movimento das falhas *en échelon* que compõem os duplexes.

Figura 2. Esquema da geração de duplexes e ramificações, extensionais e compressionais, em um sistema de zona de cisalhamento transcorrente levógiro (modificado de Woodcock e Fischer, 1986).



Os estudos relacionados à análise estrutural de uma área devem determinar a orientação das estruturas deformacionais, seu padrão geométrico bem como o tipo de tensão atuante sobre o maciço, auxiliando, desse modo, na compreensão dos processos de movimentos de massa que ocorrem nos taludes rochosos, assim como do comportamento da água subterrânea nestes locais.

Desta forma, este trabalho apresenta uma análise dos lineamentos que ocorrem na Bacia do Rio Três Forquilhas, em que está inserida a rodovia RS 486. A análise em questão foi realizada por meio de estatística vetorial e geométrica que poderá servir como subsídio, juntamente com o mapeamento estrutural, a análise geomorfológica e das litologias, para prever o comportamento das águas subterrâneas nesta região, visto que os aquíferos predominantes são do tipo fraturados.

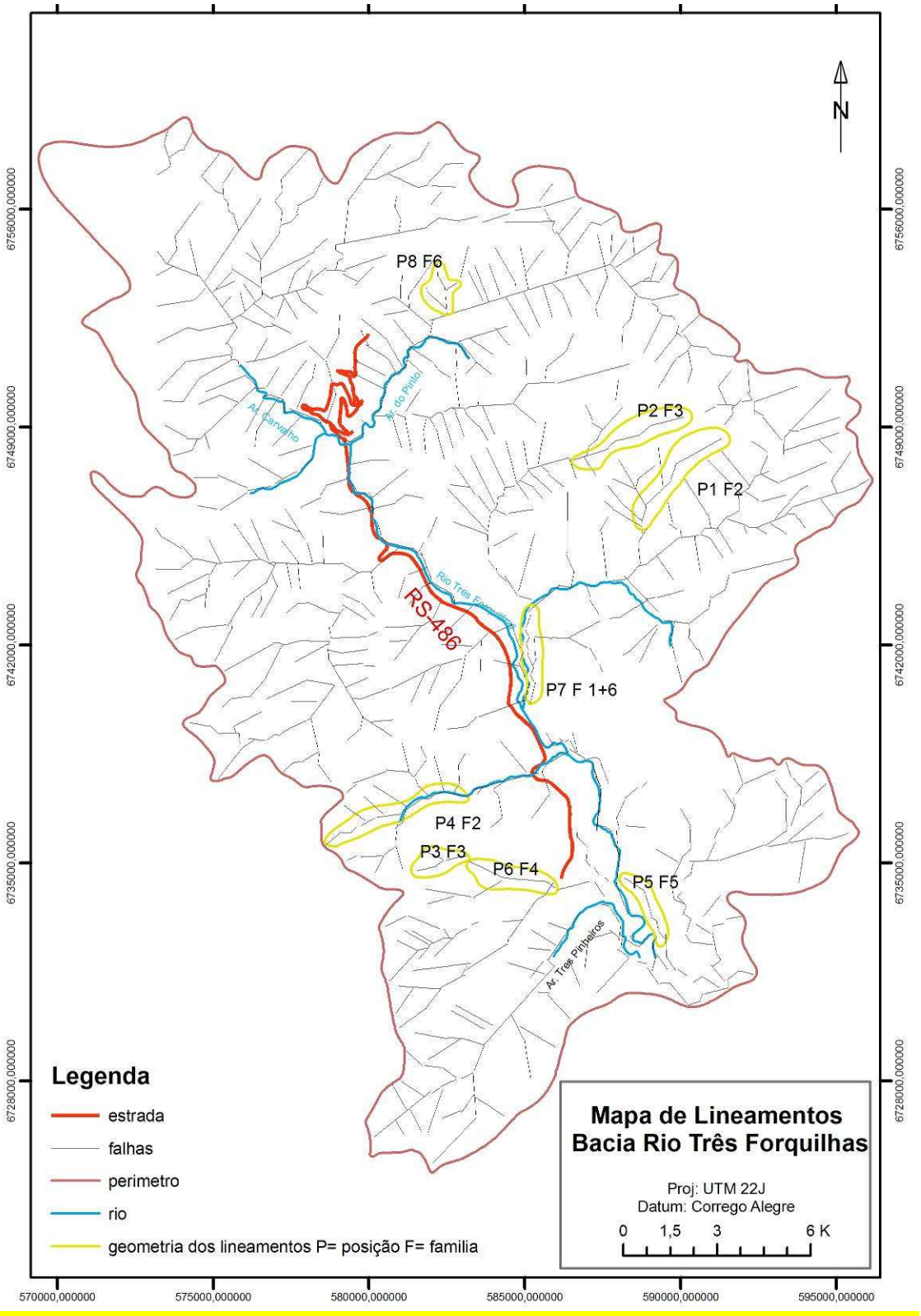
2. Desenvolvimento

A identificação e a caracterização das estruturas da Bacia do Rio Três Forquilhas deu-se por meio de análise de fotografias aéreas na escala 1: 110.000, de onde foram extraídos os lineamentos e, posteriormente, operou-se a execução de trabalhos de campo.

Neste trabalho, considerou-se o conceito de lineamento de Strieder & Amaro (1997). Para os autores, os lineamentos representam uma expressão topográfica, na superfície do terreno, de estruturas tridimensionais presentes na crosta terrestre que podem ser altos topográficos (positivos) e/ou baixos topográficos (negativos).

Para a área da bacia do Três Forquilhas, foram identificados os lineamentos Tipo 2 (dois) que, segundo os autores, correspondem a zonas de fraturas, isto é, são segmentos retilíneos de drenagem que representam uma descontinuidade da rocha, onde houve percolação de água superficial e, portanto, maior propensão à alteração intempérica. Eles desenvolvem padrões regionais de fraturas com trends que se se interceptam em diversos locais (Figura 3).

Figura 3. Mapa de lineamentos da Bacia do rio Três Forquilhas e principais padrões geométricos encontrados.



Para a extração dos lineamentos, utilizou-se um par de fotografias aéreas na escala 1:110.000 (Fx: 98, fotos nº 122487 e nº 122488, de 1975), em função da possibilidade de observação em três dimensões do relevo da região. O estereopar foi analisado em estereoscópio de espelho marca WILD e os lineamentos Tipo 2 (dois) foram delimitados pela extensão dos segmentos de drenagem e registrados em um overlay. Após a identificação, os lineamentos foram transferidos para um mapa topográfico base, em escala 1:50.000 do Serviço Geográfico do Exército (SGE), considerando a interpretação morfológica do terreno.

O comprimento e a atitude dos lineamentos foram obtidos com uma bússola do tipo Silva e armazenados num banco de dados do *Software Excel*. Com os dados relativos à direção dos lineamentos, foi construído, com o auxílio do *Software StereoNet for Windows*, um diagrama de rosetas para separar as principais famílias de estruturas. De posse dos dados relativos ao comprimento dos lineamentos, fez-se uma avaliação estatística vetorial dos lineamentos (Curray,1956 e Pincus,1956, *apud* 1996), que teve como objetivo caracterizar e delimitar cada um dos conjuntos azimutais que ocorrem na área.

A partir do histograma circular de frequência (diagrama de rosetas), os lineamentos foram divididos em famílias, cuja análise deu-se individualmente. Em continuidade, foi construído um banco de dados com o comprimento e o azimute (geográfico) de cada lineamento. Para cada família de descontinuidades, criou-se uma planilha de cálculo, utilizando-se o *software Excel*, seguindo os seguintes procedimentos:

a) Os dados de comprimento e azimute de cada lineamento foram transcritos para suas respectivas planilhas

b) Para a Família 1 (um), como só foram feitas leituras de 0º à 180º(geográfico), foi necessário rebater os ângulos de 175º à 180º para 355º à 360º, empregando-se a seguinte relação:

$$\text{Azimute lido rebatido} = \text{Azimute lido} + 180^\circ \qquad (2.1)$$

c) Como o *software* trabalha com coordenadas trigonométricas, foi necessário transformar os ângulos lidos de azimutes geográficos em trigonométricos (em cada planilha). Para isso, as relações utilizadas foram:

Azimute Geográfico (lido)	Azimute Trigonométrico (cálculos)
De 0° à 90°	= 90° - Azimute lido
De 90° à 180°	= 180° - Azimute lido + 270°

d) Os azimutes corrigidos foram dobrados para eliminar o erro da resultante, erroneamente tendenciosa para leste (Krumbein, 1939, *apud* Cunha, 1996)

e) Calculou-se a componente leste-oeste do vetor correspondente a cada lineamento da seguinte forma:

Comprimento lido do Alinhamento x Seno do Azimute dobrado (2.2)

f) Calculou-se a componente norte-sul do vetor correspondente a cada lineamento da seguinte forma:

Comprimento lido do Alinhamento x Co-seno do Azimute dobrado (2.3)

g) Calculou-se o azimute médio (θm) pelo arco tangente da divisão entre o somatório das componentes leste-oeste e o somatório das componentes norte-sul dos vetores correspondentes de cada Família.

$$\theta m = \frac{1}{2}\operatorname{arctg}\frac{\sum C \times \operatorname{sen} 2\theta_i}{\sum C \times \cos 2\theta_i} + q \qquad (2.4)$$

Os valores de θm obtidos, no Excel, são ângulos que variam entre 0º e 90º ou entre 0º e -90º (em coordenadas trigonométricas), pois o software fornece, como resposta de arco tangente, valores de ângulos nestes intervalos. A componente q (Moustafa, 1992) é um valor angular que adapta as propriedades geométricas às representações segundo azimutes geográficos, quando estas não se encontrarem no quadrante correto indicado pelo diagrama de rosetas

h) Calculou-se a magnitude do vetor resultante (R), para cada uma das planilhas, a partir do somatório das componentes leste-oeste e norte-sul dos vetores, através da fórmula:

$$R = \sqrt{(\sum C \times \cos 2\theta_i)^2 + (\sum C \times sen2\theta_i)^2} \qquad (2.5)$$

i) Calculou-se o comprimento médio (Cm) dos alimentos dividindo-se R pelo número de amostras (n) de cada uma das planilhas, através da fórmula:

$$Cm = \frac{R}{n} \qquad (2.6)$$

j) Para avaliar o grau de dispersão dos dados, calculou-se a significância do vetor (Sv) para cada família. A significância do vetor é dada pela divisão do valor R pela soma dos comprimentos de todos os alinhamentos lidos daquela família. A Sv varia entre 0 (zero) e 1 (um), sendo que valores próximos de 1 (um) indicam pequena dispersão e valores próximos de 0 (zero) mostram alta dispersão.

$$Sv = \frac{R}{\sum C} \qquad (2.7)$$

Além disso, foram construídos gráficos de dispersão correlacionando azimute geográfico verus comprimento, a fim de identificar as direções das fraturas principais (maior comprimento) e das fraturas subsidiárias (menor comprimento) dentro de cada família.

2.1 Análise dos Dados

Na área da bacia, foram extraídos 862 lineamentos do tipo 2 (dois) e, com os dados relativos a suas direções, foi elaborado um histograma circular de frequência (diagrama de rosetas) que permitiu identificar seis famílias de descontinuidades diferentes, conforme Figura 4:

a) Família 1: distribuída em um *range* de 25 graus entre os azimutes geográficos de 355° a 20° com 112 medidas;

b) Família 2: distribuída em um range de 34 graus entre os azimutes geográficos de 21° a 55° com 162 medidas;

c) Família 3: é a família que apresenta maior frequência, distribui-se em um range de 29 graus entre os azimutes geográficos de 56° a 85° com 170 medidas;

d) Família 4: distribuída em um range de 34 graus entre os azimutes geográficos de 86° a 120° com 168 medidas;

e) Família 5: distribuída em um range de 29 graus entre os azimutes geográficos de 121° a 150° com 135 medidas;

f) Família 6: distribuída em um range de 24 graus entre os azimutes geográficos de 151° a 175° com 115 medidas. Os resultados obtidos no tratamento dos lineamentos Tipo 2 encontram-se resumidos na Tabela 1.

Figura 4 Histograma circular de frequência dos lineamentos tipo 2 – Bacia do Rio Três Forquilhas (862 medidas).

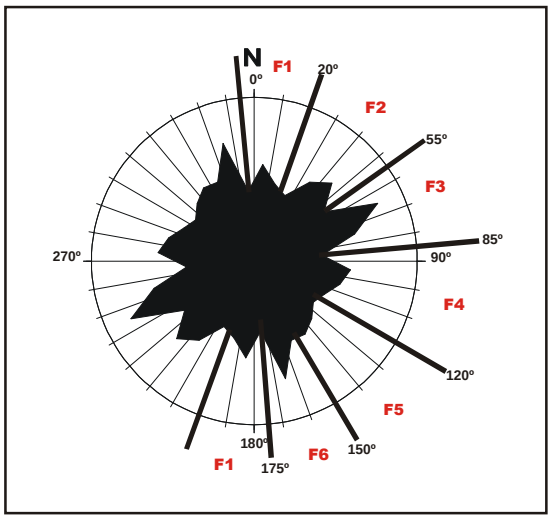


Tabela 1 Distribuição azimutal dos lineamentos Tipo 2 e resultados do tratamento estatístico vetorial.

Família	Faixa Azimutal	*Range*	Nº Medidas	$\theta_m$diag rosetas	*R (m)*	*Cm* (m)	θ_m geográfico	*Sv*
Sem classificação.	_	360	862	_	2.574.900,0	59,74	58º 02’	0,0820
1	355º-20º	25	112	5º	71.991,0	642,77	7º 15’	0,9689
2	21º-55º	34	162	45º	115.395,0	712,32	39º 20’	0,9568
3	56º-85º	29	170	65º	135.247,0	795,57	68º 30’	0,9585
4	86º-120º	34	168	95º	105.148,5	625,88	101º 22’	0,9372
5	121º-150º	29	135	140º	92.285,5	683,60	138º 30’	0,9563
6	151º-155º	24	115	165º	81.762,5	710,98	164º	0,9821

Sendo: *R* o tamanho do vetor resultante; *Cm* o comprimento médio dos lineamentos; θ_m geográfico o valor do azimute geográfico médio obtido; *Sv* a significância do vetor.

Observa-se, na Tabela 1, que o valor obtido para significância do vetor resultante (Sv) é muito baixo (0,0820) quando os dados de lineamentos são tratados como um conjunto único sem classificação. Neste caso, o valor do azimute médio (θ_m geográfico = 58º 02’) é pouco significativo. Da mesma forma, o comprimento do vetor médio resultante ($R\ (m)$ = 2.574.900,0m) também pode ser considerado um valor incoerente. Com base nestes resultados, destaca-se a importância da separação e do tratamento dos dados em populações diferentes.

De forma geral, observa-se que os seis grupos apresentam valores de comprimento médio (Cm) próximos, variando de 625m a 795m, destacando-se a Família 3 (azimute 56º- 85º) com comprimento médio de 795,57m mais expressivo que as demais. Os valores das magnitudes dos vetores resultantes (*R*) estão relacionados proporcionalmente ao número de medidas de cada família.

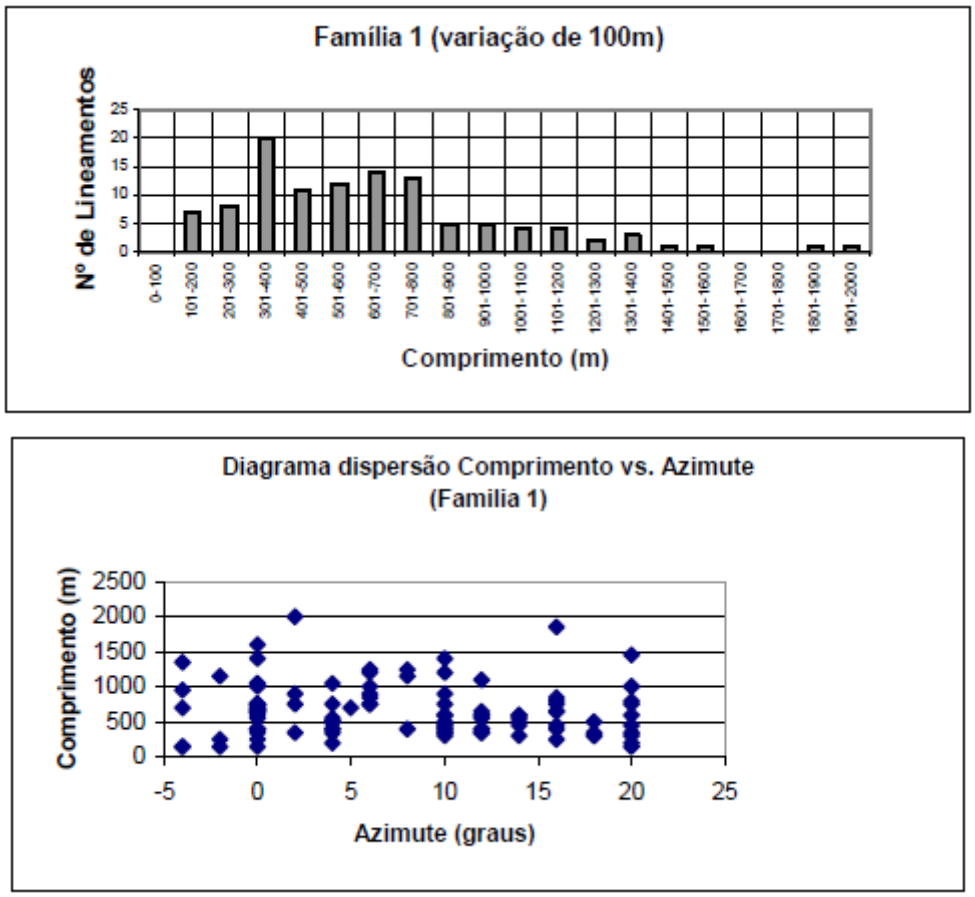
Comparando-se o vetor médio obtido no diagrama de rosetas com aqueles obtidos através de estatística vetorial, nota-se uma pequena diferença devido ao fato da estatística vetorial levar em conta os comprimentos dos lineamentos, o que nos permite obter um valor mais preciso para a direção do vetor médio.

Para cada família, foram construídos gráficos de barras em intervalos de 50, 100, 150 e 200m para ilustrar a relação entre o número de lineamentos e seus comprimentos. Este procedimento serviu para identificar as modas dos comprimentos de cada família. Os melhores resultados foram obtidos para o intervalo de comprimentos de

100m. Além destes, foram construídos diagramas de dispersão comprimento versus azimute para cada família, com o propósito de se identificar quais os azimutes dos lineamentos que apresentam os maiores comprimentos.

Analisando os gráficos da Figura 5, pode-se perceber que, na Família 1 (um), há um grande número de lineamentos de comprimentos menores, num intervalo de 301-400m, bem como nos intervalos de 601-700m e 701-800m. Os lineamentos maiores são em menor número e têm comprimentos variados. No gráfico de dispersão, tem-se que os lineamentos acima de 1001m encontram-se entre os azimutes de 0º e 10º preferivelmente.

Figura 5 Histograma e diagrama de dispersão Família 1

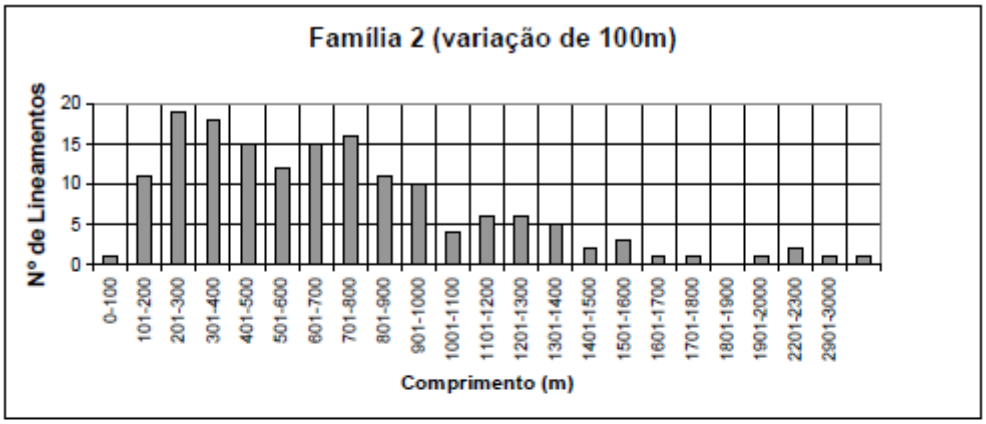


Na Família 2 (Figura 6), os lineamentos de menor comprimento representam o maior número de medidas (intervalos médios 301-400m, 701-800m e 201-300m). Nesta Família, há uma representatividade importante dos lineamentos compreendidos entre os intervalos de 1001 e 1500m. No diagrama de dispersão, os lineamentos de maior comprimento, acima de 1500m, estão entre os azimutes de 35º e 55º.

Na Família 3 (Figura 7), há uma grande população de lineamentos de tamanhos menores nos intervalos médios de 201-300m, 301-400m e 401-500m e uma população também importante no intervalo de 701-800m. Um grande número de lineamentos de grande comprimento está concentrado no intervalo entre 1101 e 1600m, cujos azimutes variam entre o intervalo 55º-65º e em torno de 80º.

Analisando os gráficos da Figura 8, para a Família 4 (quatro), encontra-se um grande número de lineamentos de tamanhos menores nos intervalos 301-400m e no intervalo médio de 601-700m. Os lineamentos de tamanho maior são menos representativos e têm comprimentos variados. No gráfico de dispersão, os maiores lineamentos (acima de 1000m) encontram-se entre os azimutes 85º-95º e em torno de 100º.

Figura 6 Histograma e diagrama de dispersão Família 2



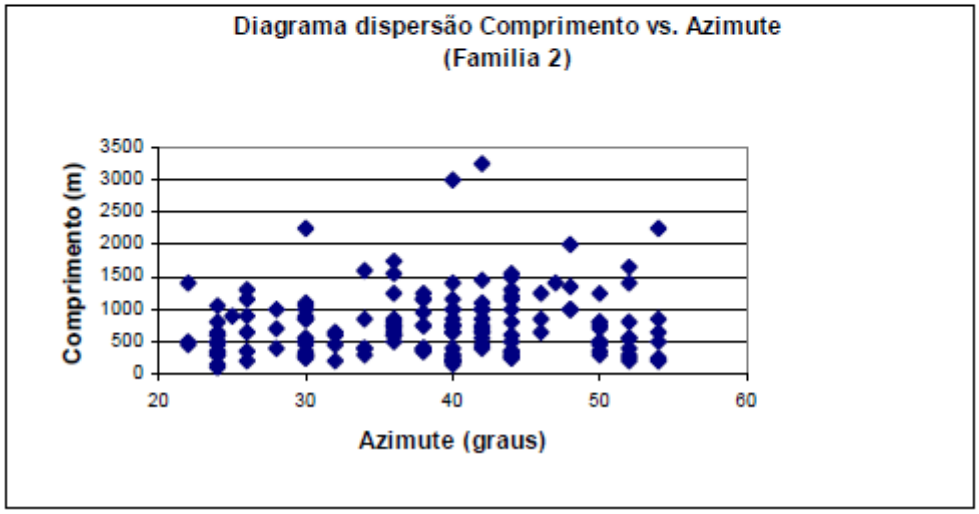
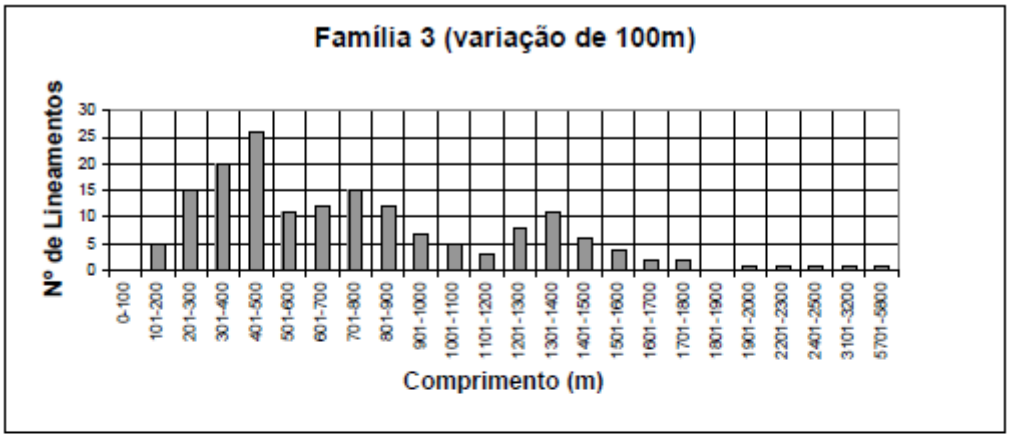


Figura 7 Histograma e diagrama de dispersão Família 3.



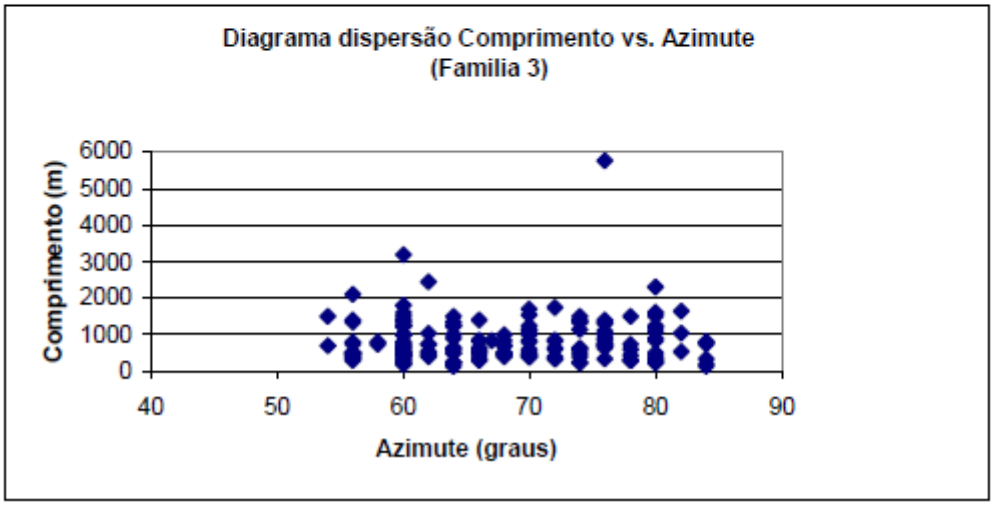
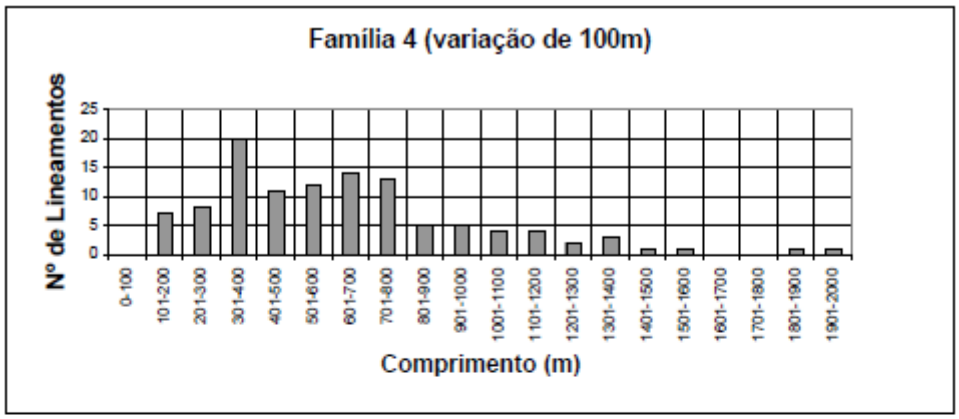
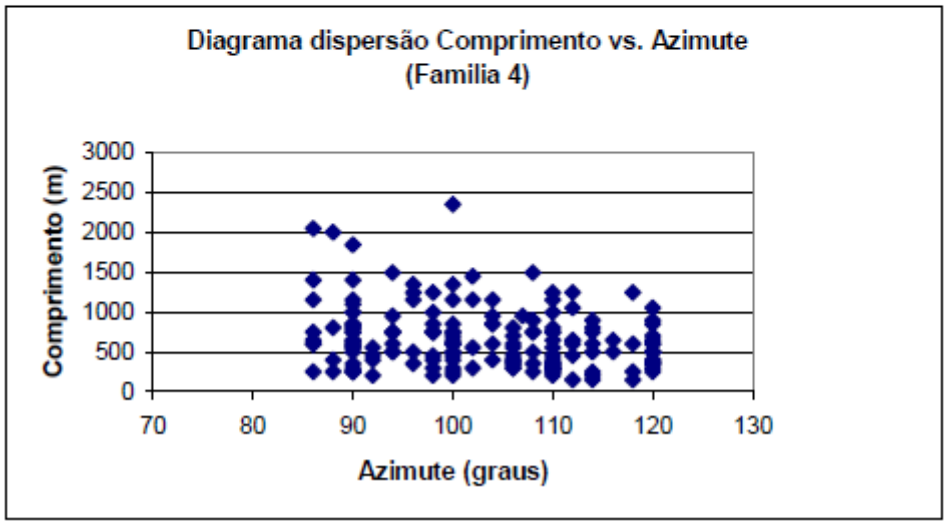


Figura 8 Histograma e diagrama de dispersão Família 4.

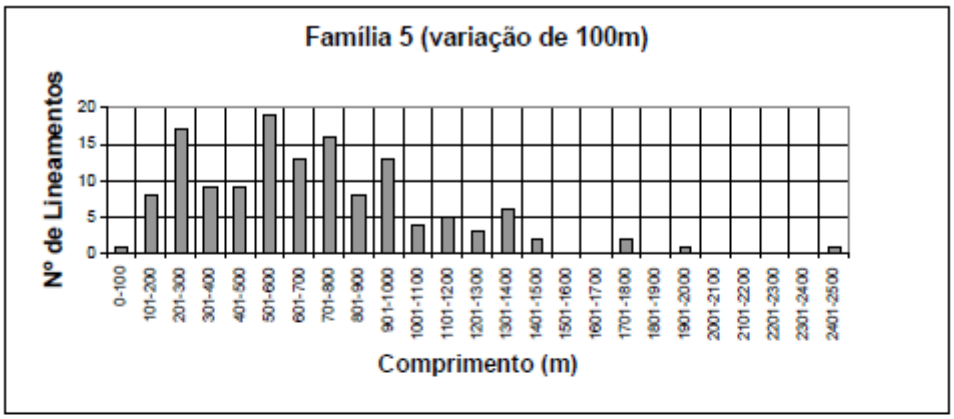




Na Família 5 (Figura 9), observa-se um grande número de lineamentos menores nos intervalos de 201-300m e 501-800m, sendo que estes últimos estão concentrados em torno dos azimutes 130º, 140º e 150º. Os lineamentos de maior comprimento (acima de 1000m) estão dispersos nos intervalos 120º-140º e em torno de 150º. Os lineamentos com comprimentos acima de 1500m encontram-se no intervalo 140º-150º.

Na Família 6 (Figura10), tem-se duas concentrações de lineamentos de tamanhos menores nos intervalos de 301-400m, 401-500m e 801-900m, dispersos entre os azimutes de 155º-170º com maiores concentrações em torno de 160º e 170º. Os lineamentos maiores, entre 1000 e 1500m, encontram-se entre os azimutes 155º-165º. Lineamentos acima de 1500m são observados em torno do azimute 160º.

Figura 9 Histograma e diagrama de dispersão Família 5



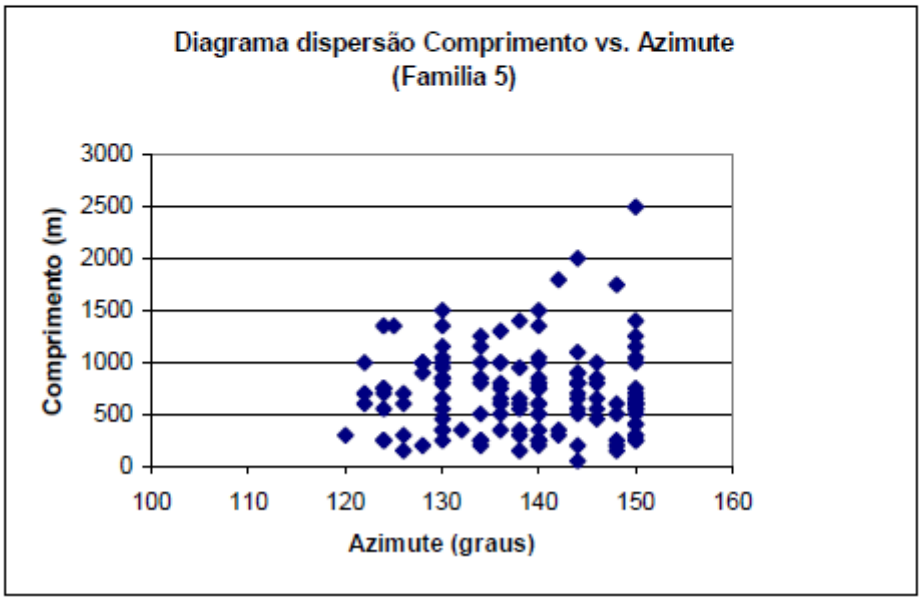
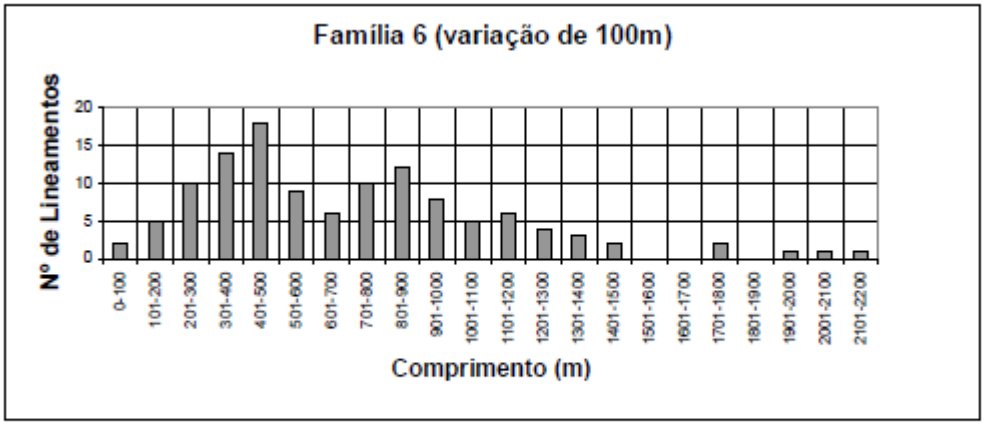
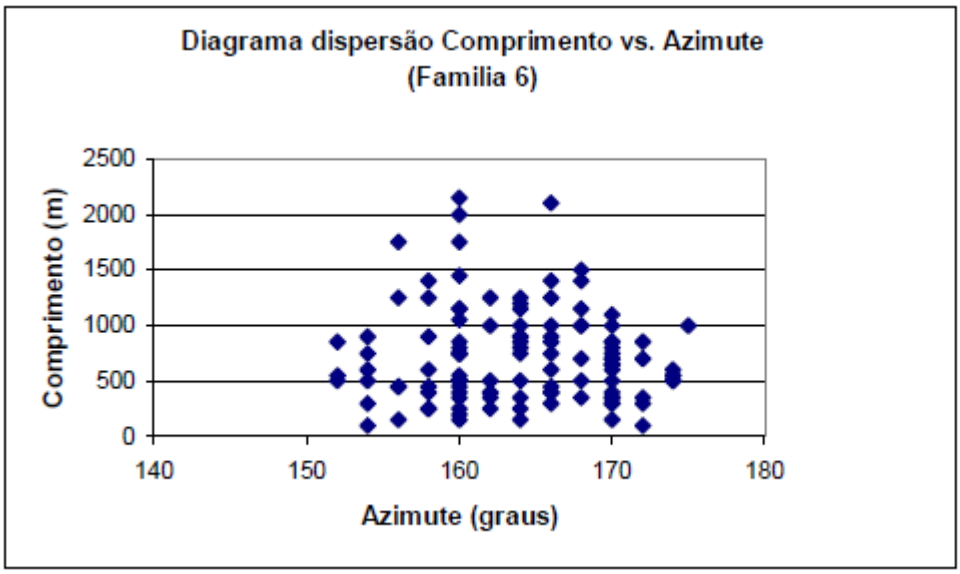


Figura 10 Histograma e diagrama de dispersão Família 6.

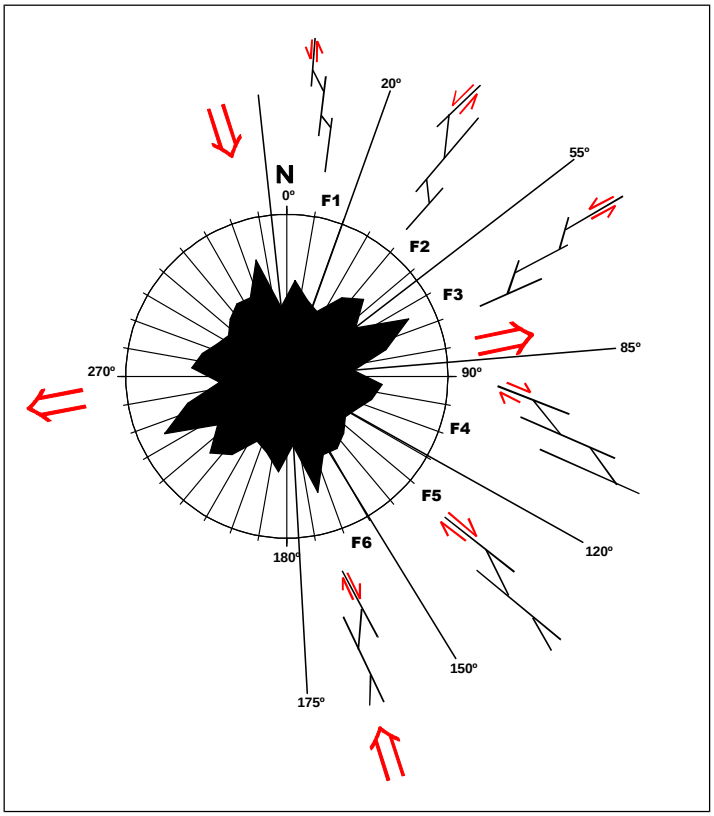




Geometricamente, os lineamentos apresentam-se como ramificações sintéticas (mesmo sentido de deslocamento dos blocos de falha). Acrescente-se que os padrões de ramificação dos lineamentos são constituídos de fraturas principais (mais longas) conectadas a fraturas consideradas como subsidiárias por apresentarem comprimentos menores. Estes padrões podem ser compostos por fraturas de mesma família ou de famílias diferentes, porém a fratura principal (maior tamanho) é que dá o nome à família. As fraturas principais são de natureza cisalhante e as fraturas subsidiárias são também cisalhantes, entretanto apresentam uma componente de tração.

Os histogramas e os diagramas de dispersão apresentados para cada família demonstram a existência destas falhas principais e subsidiárias. Os padrões de ramificações desenvolvidos em cada família analisada podem ser observados na Figura 3 com localização na área de estudo e no Diagrama de Rosetas da Figura 11.

Figura 11 Padrões de ramificações desenvolvidos em cada família. As setas grandes indicam o campo tensional, segundo Reginato (2003), para áreas próximas.



As ramificações da Família 1 (um) apresentam as fraturas principais com direção entre 0º e 10º e as subsidiárias tanto podem pertencer a mesma família como à Família 6 (seis). Na Família 2 (dois), as fraturas principais têm direção entre os azimutes 35º-45º e as subsidiárias podem pertencer à Família 1 (um). Na Família 3 (três), as fraturas principais estão entre os azimutes 60º-70º e as subsidiárias podem pertencer à Família 2 (dois).

No caso da Família 4 (quatro), as ramificações são compostas por fraturas principais com direção entre os azimutes 85º-100º e as fraturas subsidiárias pertencem à mesma Família ou à Família 5 (cinco). Na Família 5 (cinco), as fraturas principais localizam-se entre os azimutes 135º-150º e as subsidiárias podem também pertencer à Família 6 (seis). Nesta Família, 6 (seis), as fraturas principais têm direção entre 155º-165º e as subsidiárias podem pertencer à Família 1 (um).

Nos trabalhos de campo, executados para a consecução deste trabalho, foram mapeadas fraturas de cisalhamento de direção aproximadamente E-W(80º-90º;80º-90º), cujo movimento entre os blocos foi transcorrente destral. Além destas, foram encontrados diques de diabásio alojados em fraturas de atitude130º; 90º. Em todas as Famílias mapeadas, as aberturas das fraturas variaram de fechadas a abertas.

Segundo Reginato (2003), o campo tensional encontrado para a região, entre os municípios de Veranópolis e Caxias do Sul, indica uma direção de compressão NNW (aproximadamente 348º) e uma direção de tração ENE (próximo de 74º), conforme se observa na Figura 11.

Com base nas avaliações realizadas e considerando o campo tensional definido por Reginato (2003) para rochas nas proximidades, é possível estabelecer que as fraturas subsidiárias encontradas na Bacia do Rio Três Forquilhas são transtrativas (têm componente de tração) e que, se forem abertas e não estiverem preenchidas por veios e diques, têm maior capacidade de percolação de água que as demais.

3. Conclusão

A avaliação de lineamentos por meio de estatística vetorial mostra-se eficaz e contribui nos estudos relacionados à geometria das fraturas ou padrões de ramificações. Os dados de lineamento devem ser trabalhados em populações diferentes (Famílias), pois, quando tratados em um único conjunto, os valores obtidos para significância do vetor resultante (*Sv*,), azimute médio (θm geográfico) e comprimento do vetor médio resultante (R (*m*)) são muito baixos ou pouco significativos.

Para a área de estudo, foram extraídos 862 lineamentos individualizados em seis Famílias com direções e comprimentos variados formando ramificações que podem ser confirmadas pelos histogramas e diagramas de dispersão que mostram a existência de falhas principais e subsidiárias.

Com base na análise estrutural realizada na área e considerando o campo tensional definido por Reginato (2003) para rochas próximas desta região, é possível estabelecer que as fraturas subsidiárias, que ocorrem na Bacia do Rio Três Forquilhas, são transtrativas e se forem abertas, sem preenchimento, têm maior capacidade de percolação de água que outras descontinuidades que ocorrem na área.

Referências bibliográficas

CUNHA F.S.S. Análise Estrutural e Estatística de Lineamentos Aplicada à Pesquisa Mineral: o Caso da Região de Porto Nacional (TO). Dissertação. (Mestrado em Engenharia), Escola de Engenharia, UFRGS, Porto Alegre. 1996 109p.

HOBBS, B.E. MEANS, W.D., WILLIAMS, P.F. An outline of Structural Geology. New York: Wiley,1976. 287p.

NUMMER, A.V. 2003 Parâmetros Geológicos-Geotécnicos controladores dos movimentos de massa na RS486/Rota do Sol-Itati. Tese de doutorado apresentada ao programa de Pós-Graduação em Engenharia Civil. Universidade Federal do Rio Grande do Sul, UFRGS, Brasil., 245p.

MOUSTAFA, A.R. 1992. Structural setting of the Sidri-Feiran area eastern side of the Suez Rift. Cairo, Ain Shams University, Middle Eastern Research Center, Earth Science Series, v. 6, p. 44-54.

RAMSAY, J.G.; HUBER, M.I. 1987 The techniques of modern structural geology. Oxford: Academic Press,. v.2, 700p.

REGINATO P.A.R. 2003 Integração de dados geológicos para prospecção de aquíferos fraturados em trecho da Bacia Hidrográfica Taquari-Antas (RS). Tese (Doutorado em Engenharia de Minas, Metalúrgica e de Materiais) - Universidade Federal do Rio Grande do Sul,. 286 f.

STRIEDER, A.J.; AMARO, V.E. 1997 Estruturas de Lineamentos Extraídos de Imagens de Sensores Remotos. EGATEA. Porto Alegre. v.25, n.4,. p. 109-117.

WOODCOCK N.H.; FISCHER M. 1986 Strike-slip duplexes, Journal of Structural Geology v.8 , pp. 725–735

Singer, J.M., Lima, A.C.P., Tanaka, N.I. and González-López, V.A. (2007). To triplicate or not to triplicate? Chemometrics and Intelligent Laboratory Systems, 86, 82-85.

Abstract

The Basin of Três Forquilhas River is located in the northeast of the State of Rio Grande do Sul and embraces the municípalities of Terra de Areia, Itati and part of São Francisco de Paula. Along the stretch of Três Forquilhas River lies part of the State Road RS 486, known as Rota do Sol (Route of the Sun). The geology of the region is composed of volcanic rocks from the Formation Serra Geral and sandstone from the Formation Botucatu. This work presents an analysis of the orientation and pattern of the lineaments that happen in the Basin of Três Forquilhas River, accomplished by the means of vectorial and geometric statistics that, together with structural mapping, geomorphological and lithological analysis, may serve as the basis for predicting the behaviour of the underground water in that region. In this analysis the concept of lineament by Strieder & Amaro (1997) was taken into account, and the lineaments of Type 2 corresponding to the fracture zones have been identified. The length and attitude of the lineaments were stored in an Exel Software database. Out of the data related to the direction of of the leneaments a rose diagram was constructed to separate the main families of structures. Out of the length related data a vectorial statictic evaluation of the lineaments was worked out based on Curray,1956 and Pincus, 1956, quoted by Cunha, 1996, whose objective was to characterize and delimitate each of the azimuthal sets which occur in the area. For the basin area 862 lineaments were extracted and divided into six different familie: F1:355º-20º; F2:21º-55º; F3:56º-85º; F4:86º-120º; F5: 121º-150º; F6: 151º-155º, so that the major number of discontinuities ocurred in the families 2, 3 and 4. The dispersion histograms and diagrams made out for each family demonstrated the existence of main rifts (of major length) and subsidiary rifts (of minor length). The geometric analysis confirms a pattern of ramification formed out of main fractures (longer rifts) connected with subsidiary fractures (shorter length rifts). Based on the evaluation worked out and considering the tensional field defined by Reginato (2003) for the vicinity rocks it is possible to conclude that the subsidiary fractures found in the Basin of Três Forquilhas River are transtractive (characterized by components of traction) and that, in case of being open and not filled by vein and dikes, they are endowed with greater capacity of water percolation than the others.

Um método baseado em combinação de soluções com coassociação para o problema de agrupamento automático

Gustavo Silva Semaan [1]
Cláudio de Carvalho Torres [2]
José André de Moura Brito [3]
Luiz Satoru Ochi [4]

Resumo

Métodos da área de Cluster Analysis podem ser aplicados com duas finalidades: identificar grupos dentro de um conjunto de dados supondo fixado o número de grupos e uma função objetivo, ou identificar o número ideal de grupos mediante avaliação de algum índice de validação. Neste sentido, o presente trabalho traz a proposta de um método de combinação de soluções baseado na Técnica Matriz de Coassociação. Mais especificamente, é utilizado um Método Hierárquico Aglomerativo para a obtenção de padrões em soluções para o Problema de Agrupamento Automático.

A qualidade das soluções obtidas é avaliada mediante a aplicação do Índice Silhueta, que combina coesão e separação. Em uma primeira fase, foram realizados experimentos preliminares com o objetivo de selecionar instâncias que possuem tendência à formação de agrupamentos considerando a utilização da Estatística de Hopkins.

Os resultados apresentados neste trabalho indicam que o método proposto foi capaz de identificar padrões nas soluções do conjunto base, obtendo soluções equivalentes ou de melhor qualidade. Nos comparativos com trabalhos da literatura, os melhores resultados foram obtidos em 14 das 17 instâncias utilizadas e, nas instâncias em que as silhuetas obtidas foram inferiores, a diferença no número de grupos foi de apenas uma unidade.

Palavras Chave: Agrupamento Automático, Comitê de Agrupamentos, Estatística de Hopkins, Índice Silhueta.

[1] Instituto do Noroeste Fluminense de Educação Superior – INFES/UFF, E-mail: gustavosemaan@id.uff.br
[2] Departamento de Ciência da Computação - DCC/UNIPLI - E-mail:claudioct@gmail.com
[3] Escola Nacional de Ciências Estatísticas – ENCE/IBGE - E-mail: jose.m.brito@ibge.gov.br
[4] Instituto de Computação – IC/UFF, E-mail: satoru@ic.uff.br

1. Introdução

A análise de agrupamentos agrega um conjunto de métodos que são aplicados à determinação de grupos a partir de um conjunto de objetos definidos por certas características (atributos). O objetivo é obter grupos que apresentem padrões (características) semelhantes e que possam refletir a forma como os dados são estruturados. Para isso, deve-se maximizar a similaridade (homogeneidade) entre os objetos de um mesmo grupo e minimizar a similaridade entre objetos de grupos distintos [Han and Kamber, 2012] [Larose, 2005] [Goldschimidt e Passos, 2005] [Jain and Dubes, 1988].

Formalmente, o problema clássico de agrupamento pode ser definido da seguinte maneira: dado um conjunto formado por n objetos $\mathrm{X}=\{\mathrm{x}_1,\mathrm{x}_2,...,\mathrm{x}_\mathrm{i},...,\mathrm{x}_\mathrm{n}\}$, com cada objeto $x_i \in X$ possuindo p atributos (dimensões ou características), ou seja, $x_i=\{x_i^1,x_i^2,...,x_i^p\}$, deve-se construir k grupos $C_j(j=1,...,k)$ a partir de X, de forma a garantir que os objetos de cada grupo sejam homogêneos segundo alguma medida de similaridade [Rosseeuw, 1987]. Uma solução (ou partição) pode ser representada como $\pi=\{C_1,C_2,...,C_k\}$. Além disso, devem ser respeitadas as restrições concernentes a cada problema particular abordado [Han and Kamber, 2012] [Ester et al., 1995] [Baum, 1986] [Hruschka and Ebecken, 2001] [Dias and Ochi, 2003]. Apresenta-se, a seguir, o conjunto de restrições que definem o problema clássico de agrupamento. Estas restrições determinam, respectivamente, que: o conjunto X corresponde à união dos objetos dos grupos, cada objeto pertence a exatamente um grupo e todos os grupos possuem pelo menos um objeto.

$$\bigcup_{j=1}^{k} C_j = X \tag{1}$$

$$C_i \bigcap C_l = \varnothing \text{ i,l=1, .., k e i } \neq \text{l} \tag{2}$$

$$C_j \neq \varnothing \text{ j=1, .., k} \tag{3}$$

Para este problema, o número de soluções possíveis, ou seja, o total de maneiras em que os n objetos podem ser agrupados, considerando um número fixo de k grupos, é dado pelo número de Stirling (NS) de segundo tipo [Johnson and Wichern, 2012], e podem ser obtidas pela Equação 4. Para problemas de agrupamento em que o valor de k é desconhecido (agrupamento automático), o número de soluções possíveis aumenta ainda mais. E, neste caso, o número é dado pela Equação 5, que corresponde ao somatório da Equação 4 para o número de grupos variando no intervalo [1,kmax], sendo kmax o número máximo de grupos.

Para ilustrar a ordem de grandeza desse número, no caso de $n=10$ objetos a serem alocados em $k=3$ grupos, o número de soluções a serem consideradas é de $NS(10,3)=9.330$ (Equação 4). Mas considerando apenas o dobro de objetos, ou seja, $n=20$ e $k=3$, o número de soluções possíveis (Equação 4) sobe para $NS(20,3)=580.606.446$. No problema de agrupamento automático estes valores crescem exponencialmente com o aumento da quantidade de objetos (n). Por exemplo, para $n=10$ e a quantidade de grupos no intervalo $k=1,...,n$ o número de soluções possíveis é $NS(10)=115.975$ (Equação 5). Esta característica torna proibitiva a obtenção da solução ótima mediante a aplicação de um procedimento de enumeração exaustiva [Naldi, 2011].

$$NS(n,k)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{n} \tag{4}$$

$$NS(n)=\sum_{i=1}^{k_{\max}}NS(n,i) \tag{5}$$

Conforme [Kumar et al., 2009], as últimas décadas, e em particular os últimos anos, têm sido marcadas pelo desenvolvimento de diversos algoritmos de agrupamento. Por sua vez, estes algoritmos encontram aplicação em diversos domínios, quais sejam: inteligência artificial, reconhecimento de padrões, marketing, economia, ecologia, estatística, pesquisas médicas, ciências políticas, etc.

Esses algoritmos possuem diferentes características, o que pode implicar, por sua vez, em uma grande variedade de soluções possíveis para um mesmo conjunto de dados de entrada. Além disso, alguns deles possuem parâmetros livres que influenciam na solução obtida, o que gera uma diversidade ainda maior de soluções. Porém, nenhum algoritmo é apropriado para todos os tipos de dados, formatos de grupos e aplicações [Naldi, 2011] [Kumar et al., 2009] [Han and Kamber, 2012].

A combinação de agrupamentos (ou comitê de agrupamento, do inglês "*cluster ensemble*") consiste em combinar um conjunto de soluções (denominado conjunto base) em comitês, com o objetivo de obter uma solução que possa aproveitar características das soluções do conjunto base.

Com base nas definições do problema clássico de agrupamentos, define-se a combinação de agrupamentos como: dado um conjunto formado por q soluções do problema de agrupamento, $\Pi = \{\pi_1, \pi_2, ..., \pi_q\}$, deve-se encontrar uma única Solução Consenso [Naldi 2011] [Naldi and Carvalho 2007]. Nesse sentido, a combinação de agrupamentos pode ter diferentes objetivos, como:

Robustez: obter uma *solução consenso* de melhor qualidade que a maioria das soluções do conjunto base ou mesmo uma solução com menor sensibilidade a ruídos e outliers.

Novidade: obter uma solução consenso inédita, que não poderia ser formada com a utilização dos algoritmos de agrupamentos utilizados no processo individualmente.

Reaproveitamento de conhecimento: utilizar o conhecimento obtido para a formação das soluções base para construir a solução consenso.

Consistência: obter uma partição consenso tal que, de alguma forma, esteja em concordância com as partições base.

Desempenho e custo computacional: para reduzir a complexidade (custo) computacional podem ser utilizados algoritmos que utilizem diferentes técnicas e diferentes objetivos para que seus resultados sejam combinados, de forma a produzir uma solução consenso mais robusta que as soluções do conjunto base.

O objetivo relacionado ao Desempenho e Custo Computacional sugere, também, que há espaço para o estudo e o desenvolvimento de novos algoritmos de agrupamento que sejam mais eficientes ou mais apropriados, levando em conta as características específicas de conjuntos de dados. Em muitos casos, inclusive, a análise de "o que é uma boa solução" é subjetiva, tendo em vista as especificidades do problema estudado.

Não obstante, para aumentar a chance de obter sucesso com a utilização das técnicas de comitês de agrupamento, é necessário considerar dois aspectos importantes, sejam eles: a Diversidade, relacionada às soluções que compõem o conjunto base e a Função Consenso, que realiza efetivamente a combinação das soluções. Em relação ao primeiro aspecto apresentado, é necessário que as soluções do conjunto base possuam um grau de diversidade mínimo, de forma a justificar tanto o custo computacional do algoritmo de combinação, quanto formar soluções que atendam a um dos objetivos almejados na utilização da combinação de agrupamentos. Já sobre a função consenso, destacam-se as técnicas: baseadas em Coassociação, em votação e em particionamento de grafos/hipergrafos [Naldi et al., 2009].

Após a apresentação das definições do problema de agrupamento, de agrupamento automático e da técnica de combinação de agrupamentos, devem ser apresentadas as especificidades existentes no método proposto. O presente trabalho está dividido em cinco seções, incluindo a introdução. A seção 2 apresenta uma revisão da literatura concernente aos algoritmos que tratam o problema de agrupamento automático e aos comitês de agrupamentos. Ainda nessa seção é apresentado o índice Silhueta para avaliação das soluções. Já a seção 3 apresenta o algoritmo DBSCAN *(Density-Based Spatial Clustering of Application with Noise)*, utilizado para a obtenção das soluções que compõem o conjunto base, que será submetido ao algoritmo de combinação de agrupamentos. Ainda na seção 3 é apresentado o método proposto neste trabalho. A seção 4 apresenta a *Estatística de Hopkins* (EH), utilizada para identificar se existe tendência à formação de agrupamentos. Nessa mesma seção são apresentados experimentos relacionados à EH. A seção 5 traz os resultados computacionais obtidos considerando, inclusive, os comparativos realizados com algoritmos mais sofisticados da literatura, propostos por [Cruz, 2010]. Por fim, a seção 6 apresenta as conclusões do trabalho e sugere trabalhos futuros.

2. Revisão da Literatura

Conforme [Kumar et al., 2009], talvez um dos problemas de seleção de parâmetros mais conhecido seja o de determinar o número ideal de grupos em um problema de agrupamento automático. Neste sentido, uma revisão da literatura é apresentada e algoritmos heurísticos baseados em metaheurísticas e métodos sistemáticos são relatados.

Considerando o algoritmo clássico de agrupamento baseado em protótipos, o *k-Means*, algumas abordagens são apresentadas com o objetivo de resolver o problema de agrupamento automático. Nesse sentido é possível destacar os algoritmos K´-Means e o *X-Means* [Pelleg and Moore, 2000] [Zalik, 2008].

Entre os trabalhos na literatura que propõem algoritmos baseados em metaheurísticas estão: [Pan and Cheng, 2007] [Ma et al., 2006] [Alves et al. 2006] [Soares and Ochi, 2004] e [Soares, 2004]. Existem, também, as heurísticas baseadas em metaheurísticas que utilizam alguns procedimentos para o refinamento de soluções (buscas locais) baseados no algoritmo *k-Means*. Em um primeiro momento essas heurísticas utilizam algoritmos para construção de grupos, denominados *Grupos Parciais* (temporários, componentes conexos), com o objetivo de unir os objetos mais homogêneos (conforme uma medida de distância). Em seguida são aplicados procedimentos de *Busca Local* e de *Perturbação* nos *Grupos Parciais* produzindo soluções de boa qualidade, conforme a função de avaliação considerada. Dessa forma, os grupos parciais são unidos e formam os Grupos Finais das soluções obtidas [Cruz, 2010] [Tseng and Yang, 2001] [Naldi et al., 2009] [Naldi and Carvalho, 2007] [Hruschka and Ebecken, 2003] [Hruschka et al., 2004a] [Hruschka et al., 2004b] [Alves et al. 2006] [Hruschka et al., 2006].

Segundo [Naldi 2011], de uma forma prática, pode-se definir o procedimento de determinar o número ideal *k* de grupos em um problema de agrupamento automático em dois passos. O primeiro passo consiste em executar diversas vezes algoritmos de agrupamento, considerando que o número *k* de grupos (parâmetro de entrada) irá variar em um intervalo pré-determinado. O segundo passo consiste na utilização de índices de validação para verificar a qualidade das soluções obtidas. Nesse sentido, são apresentadas duas abordagens sistemáticas que atuam na maximização do Índice

Silhueta e que consistem em múltiplas execuções do algoritmo *k-Means*: o *MRk-means* (do inglês *Multiple Runs of k-means*) e o *OMRk-means* (do inglês *Ordered Multiple Runs of k-means*) [Naldi 2011].

Em [Semaan et al., 2013] foi proposto um *Método Sistemático baseado em Densidade* que utiliza um conhecido algoritmo da literatura, o DBSCAN [Ester et al., 1996], para obter soluções para o problema de agrupamento automático. Nesse método, o algoritmo DBSCAN foi adaptado para que os objetos identificados como *outliers* não fossem ignorados, ou seja, eles deveriam ser considerados nas soluções, não violando assim a restrição apresentada na Equação 1 (todos os objetos precisam estar associados a algum grupo). Além disso, foi utilizada uma técnica de calibração automática dos parâmetros de entrada para o algoritmo DBSCAN, denominada *Dist-k* [Kumar et al., 2009].

O trabalho [*Frossyniotis*, *Likas and Stafylopatis*, 2004] traz a proposta de um algoritmo de agrupamento que combina um método de agrupamento não hierárquico com a técnica de *boosting* [James et al., 2013]. Mais especificamente, em cada iteração do *boosting*, um novo conjunto de treinamento é criado considerando a seleção de uma amostra aleatória ponderada do conjunto original de dados. Em seguida, de forma a produzir uma partição, ou seja, um possível agrupamento, aplica-se sobre esta amostra um algoritmo como o *k-means*. Após todas as iterações, a partição final é produzida combinando as múltiplas partições. Os trabalhos de [Vu et at., 2010], [Frossyniots, Pertselakis and Stafylopatis, 2002] e [*Freund and Schapire*, 1996] também constituem-se como boas referência sobre o tema.

Em relação à função de avaliação, foi utilizado o *Índice Relativo Silhueta* (seção 2.1). Os índices relativos, como o próprio nome sugere, têm como finalidade avaliar a qualidade relativa das soluções produzidas por diferentes métodos de agrupamento. Estes índices não têm a propriedade de *monotonicidade*, ou seja, não são afetados pelo aumento ou pela redução do número de grupos da solução [Naldi 2011] [Cruz 2010]. Dessa forma, podem ser utilizados na avaliação de diversas soluções, provenientes de diversos algoritmos.

Os algoritmos de agrupamento baseados em densidade têm como objetivo a determinação de grupos (regiões) de alta densidade de objetos separados por regiões de baixa densidade. Nesse contexto, as soluções do conjunto base, que serão submetidas ao método proposto no presente trabalho, foram obtidas com a aplicação do DBSCAN.

2.1 A Silhueta

O Índice Silhueta foi proposto por [Rousseeuw, 1987] e tem por finalidade determinar a qualidade das soluções com base na proximidade entre os objetos de determinado grupo e na distância desses objetos ao seu grupo mais próximo. Dessa forma ele combina as ideias de coesão e de separação. O índice deve ser calculado para cada objeto e possibilita identificar se o objeto está alocado ao grupo mais adequado. A silhueta da solução é a média das silhuetas dos objetos. A seguir, os quatro passos explicam como calcular a silhueta de uma solução:

1. Neste trabalho d_{ij} (Equação 6) corresponde à Distância Euclidiana entre dois objetos x_i e x_j, e p é a quantidade de atributos dos objetos. Para cada objeto x_i calcula-se a sua distância média $a(x_i)$ (Equação 7) em relação a todos os demais objetos do mesmo grupo. Na Equação 7, $|C_w|$ representa a quantidade de objetos do grupo C_w, ao qual o objeto x_i pertence.

$$\mathrm{d_{ij}} = \sqrt{\sum_{\mathrm{q=1}}^{\mathrm{p}} \left(\mathrm{x_i^q} - \mathrm{x_j^q}\right) 2} \qquad (6)$$

$$a(x_i) = \frac{1}{|C_w| - 1} \sum d_{ij} \quad \forall x_j \neq x_i \quad x_j \in C_w \qquad (7)$$

2. A Equação 8 apresenta a distância entre o objeto x_i e os objetos do grupo C_t, em que $|C_t|$ é a quantidade de objetos do grupo C_t. Para cada objeto x_i calcula-se a sua distância média em relação a todos os objetos dos demais grupos ($b(x_i)$) (Equação 9) e busca-se a menor distância (grupo externo mais próximo).

$$d(x_i, C_t) = \frac{1}{|C_t|} \sum d_{ij} \qquad \forall x_j \in C_t \tag{8}$$

$$b(x_i) = \min d(x_i, C_t) \qquad C_t \neq C_w \qquad C_t \in C \tag{9}$$

3. O índice silhueta do objeto x_i ($s(x_i)$) pode ser obtido pela Equação 10 e seu valor está no intervalo [-1,1].

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}} \tag{10}$$

4. O cálculo da silhueta de uma solução S é a média das silhuetas de cada objeto, conforme apresenta a Equação 11, em que n é a quantidade de objetos da solução. Essa função deve ser maximizada. Valores positivos de silhueta indicam que o objeto está bem localizado em seu grupo, enquanto valores negativos indicam que o objeto deveria ser alocado a outro grupo.

$$\max Silhueta(S) = \frac{1}{n} \sum_{i=1}^{n} s(x_i) \tag{11}$$

3. Método baseado em combinação de soluções com coassociação

Na introdução deste trabalho foram apresentadas as definições dos problemas de agrupamento, agrupamento automático e da técnica de combinação de agrupamentos, bem como alguns dos objetivos almejados em sua utilização.

Foram apresentados, também, os aspectos necessários para o sucesso da utilização da combinação, que são: função consenso e diversidade.

O método proposto no presente trabalho está associado à aplicação de uma função consenso baseado em coassociação para obtenção de soluções para o problema de agrupamento automático. A presente seção apresenta o método proposto bem como o algoritmo DBSCAN, utilizado na formação das soluções do conjunto base.

3.1 DBSCAN

Os algoritmos de agrupamento baseados em densidade têm como objetivo a determinação de grupos (regiões) de alta densidade de objetos separados por regiões de baixa densidade. Nesse contexto, o algoritmo DBSCAN [Ester et al., 1996] é um dos mais conhecidos da literatura e possui uma complexidade computacional $O(n^2)$. Trata-se de um algoritmo simples, eficiente, e que contempla conceitos importantes, que servem de base para qualquer abordagem baseada em densidade.

O DBSCAN utiliza-se de um conceito de densidade tradicional baseada em centro. Mais especificamente, a densidade de um objeto x_i é a quantidade de objetos em um determinado raio de alcance (distância) de x_i, incluindo o próprio objeto. Este algoritmo possui como parâmetros de entrada o raio (raioDBSCAN) e a quantidade mínima de objetos em um determinado raio (qtdeObjetos). A densidade de um objeto, portanto, depende do raio especificado.

Deve-se, então, calibrar o parâmetro raioDBSCAN para que o seu valor não seja tão alto de forma que todos os objetos tenham densidade n (solução com apenas um grupo), e nem tão baixo em que todos os objetos terão densidade 1 (solução com n grupos denominados singletons). A abordagem da densidade baseada em centro realiza a classificação dos objetos em:

Interiores ou Centrais: objetos que pertencem ao interior de um grupo baseado em densidade. Deve possuir uma quantidade de objetos em seu raio raioDBSCAN igual ou superior ao parâmetro qtdeObjetos - 1.

Limítrofes: não é um objeto central, mas é alcançável por ao menos um objeto central, ou seja, está dentro do raio de vizinhança de algum objeto central.

Ruídos: demais objetos que não são Centrais e nem estão na vizinhança de um objeto central.

O algoritmo DBSCAN consiste nos seguintes passos:

1. Classificar os objetos como Objetos Centrais, Limítrofes ou Ruídos.

2. Eliminar os objetos que sejam classificados como Ruídos.

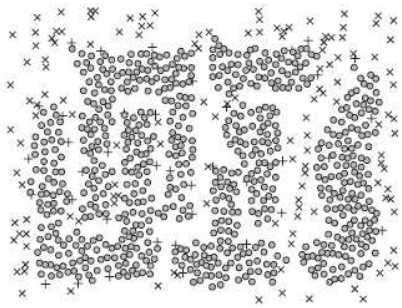
3. Adicionar arestas entre Objetos Centrais x_i e x_j que estejam dentro do raioDBSCAN, ou seja, quando a distância entre esses objetos é menor ou igual ao raioDBSCAN.

4. Tornar cada grupo de objetos de centro um grupo separado.

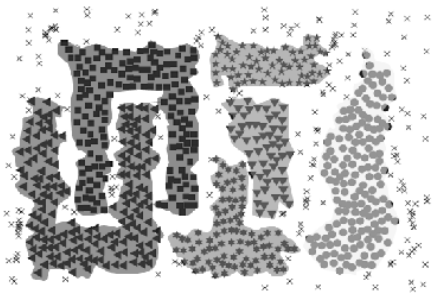
5. Atribuir cada Objeto limítrofe a um dos grupos dos seus objetos centrais associados.

Como base nestas informações, a Figura 1 ilustra a classificação de cada um dos objetos em *Ruído, Limítrofe ou Interior*. Essa mesma figura apresenta também uma solução obtida com a execução do DBSCAN, em que é possível observar que objetos identificados como dos tipos Interior ou Limítrofes formam grupos, enquanto objetos do tipo Ruído permanecem isolados e não fazem parte de nenhum grupo.

Figura 1: Classificação de 3000 objetos de duas dimensões pelo DBSCAN [Kumar et al., 2009]



(x)Ruído (+)Limítrofe (o)interior



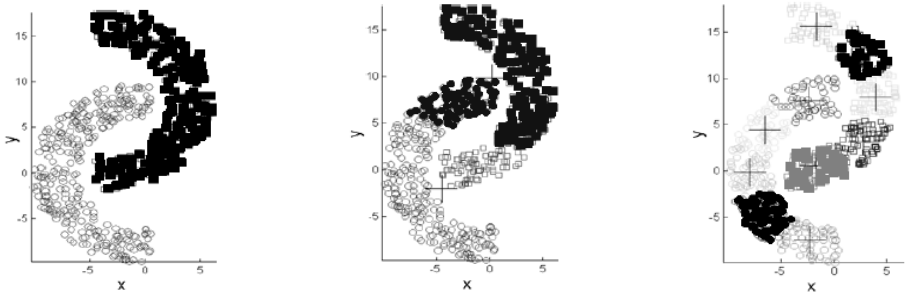
Solução obtida

Tendo em vista que o DBSCAN é um algoritmo baseado em densidade, o mesmo é imune a ruídos, uma vez que esses objetos são identificados e ignorados (não pertencem a grupos). Além disso, o algoritmo pode trabalhar com grupos de tamanhos diferentes (número de objetos) e formas arbitrárias. Dessa forma, ele é capaz de identificar grupos que não poderiam ser encontrados mediante a aplicação de outros algoritmos, como por exemplo, o *k-means*.

A Figura 2 (a) apresenta os objetos iniciais de uma instância em que é possível observar dois grupos, que poderiam ser obtidos com o algoritmo DBSCAN.

Já as Figuras 2 (b) e (c) apresentam resultados obtidos com a utilização do k-means. Esse algoritmo tem como característica a obtenção de grupos esféricos e de tamanhos semelhantes em relação aos raios dos centróides. É possível observar que mesmo para k = 2 os grupos da Figura 2 (a) não foram obtidos.

Figura 2: (a) Objetos Originais; (b) e (c) Soluções do *k*-Means [Kumar et al., 2009].



(a) Objetos Originais **(b)** Solução do *k-Means* para k=2 **(c)** Solução do *k-Means* para k=10

3.2 Método Proposto

O método proposto neste trabalho consiste na utilização da técnica para comitê de agrupamentos baseada na matriz de coassociação [Naldi, 2011]. Essa técnica calcula a similaridade s_{ij} entre dois objetos x_i e x_j por meio do número de grupos compartilhados entre eles nas soluções do *Conjunto Base*. Para isso, com base em uma matriz de coassociação, um algoritmo para agrupamento hierárquico realiza cortes, formando soluções, que são avaliadas por meio da utilização do Índice Silhueta. A quantidade de grupos da solução com o maior valor para esse índice é considerada a

ideal. Com o objetivo de demonstrar o método proposto, um estudo de caso é apresentado.

A Tabela 1 apresenta soluções para o problema de agrupamento automático, em que as soluções π_1, π_2 e π_3 possuem três grupos e a solução π_4 possui quatro grupos (e nessa solução os grupos C_3 e C_4 são singletons). Para cada solução do conjunto base é gerada uma matriz solução. Em seguida, uma matriz consenso é obtida por meio da soma de todas as matrizes solução.

A Tabela 2 apresenta uma matriz que representa a solução π_1. Nessa matriz, o valor 1 indica que os objetos x_i e x_j estão em um mesmo grupo (para i $\neq$ j). Por exemplo, os objetos x_1 e x_2 estão no mesmo grupo, enquanto os objetos x_1 e x_3 estão em grupos distintos (em destaque). Já a Tabela 3 apresenta uma matriz que representa a solução π_4. Nessa tabela, em destaque, estão as células (x_2, x_2) e (x_4, x_4). Para estas matrizes, temos que o valor 1 associado a uma entrada da diagonal principal indica que objeto xi está alocado em um grupo com outro(s) objeto(s).

Caso contrário, esse objeto está em um grupo singleton (um grupo com apenas um objeto). É importante ressaltar que esta matriz é simétrica.

Tabela 1: Exemplo de um conjunto base (soluções).

Soluções	Grupos
π_1	C_1={x_1,x_2}, C_2={x_3, x_4, x_5}, C_3={x_6, x_7}
π_2	C_1={x_1,x_2, x_3}, C_2={x_4, x_5}, C_3={x_6, x_7}
π_3	C_1={x_1,x_2}, C_2={x_3, x_4}, C_3={x_5, x_6, x_7}
π_4	C_1={x_1,x_6}, C_2={x_3, x_5, x_7}, C_3={x_2}, C_4={x_4}

Tabela 2: Representação da solução π_1 em uma matriz.

π_1	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1	**1**	**0**	0	0	0	0
x_2		1	0	0	0	0	0
x_3			1	1	1	0	0
x_4				1	1	0	0
x_5					1	0	0
x_6						1	1
x_7							1

Tabela 3: Representação da solução π_4 em uma matriz.

π_4	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1	0	0	0	0	1	0
x_2		**0**	0	0	0	0	0
x_3			1	0	1	0	1
x_4				**0**	0	0	0
x_5					1	0	1
x_6						1	0
x_7							1

Tabela 4: Matriz consenso.

M	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	4	**3**	1	0	0	1	0
x_2		3	1	0	0	0	0
x_3			4	2	2	0	1
x_4				3	2	0	0
x_5					4	1	2
x_6						4	**3**
x_7							4

Tabela 5: Matriz consenso.

M	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	4	**3**	1	0	0	1	0
x_2		3	1	0	0	0	0
x_3			4	**2**	**2**	0	1
x_4				3	**2**	0	0
x_5					4	1	**2**
x_6						4	**3**
x_7							4

A partir da soma dos valores das matrizes solução que representam as soluções do conjunto base (Tabela 1), uma matriz consenso é formada. Com base na matriz consenso apresentada na Tabela 4, aplica-se o algoritmo hierárquico para a obtenção de soluções consenso. Os valores de corte submetidos ao algoritmo hierárquico são valores inteiros no intervalo $\left[2\,|\Pi|\right]$, em que $|\Pi|$ é a quantidade de soluções do conjunto base. Para cada valor de corte, a solução formada é avaliada por meio da utilização do índice silhueta. Destaca-se, que mesmo com valores de cortes diferentes, pode-se obter a mesma solução, ou seja, diferentes valores de corte submetidos ao algoritmo hierárquico podem resultar em uma mesma solução.

Após a aplicação do algoritmo hierárquico para todos os valores de corte, a solução final será escolhida com base nos valores das silhuetas das soluções obtidas. Esse algoritmo hierárquico utiliza uma Estratégia Aglomerativa, ou seja, inicialmente todos os objetos formam grupos singletons ($C_1 = \{x_1\}$, $C_2 = \{x_2\}$, $C_3 = \{x_3\}$, $C_4 = \{x_4\}$, $C_5 = \{x_5\}$, $C_6 = \{x_6\}$ e $C_7 = \{x_7\}$), sendo estes grupos unidos com base na matriz consenso.

A Tabela 4, por exemplo, apresenta a matriz consenso em que as células em destaque indicam valores equivalentes ou superiores ao valor de corte submetido como parâmetro (corte = 3) para células m_{ij}, $i < j$, (linha i e coluna j). Os objetos x_1 e x_2 pertencem a um mesmo grupo em três das quatro soluções do conjunto base. O mesmo ocorre com os objetos x_6 e x_7. Dessa forma, os grupos $C_1 = \{x_1\}$ e $C_2 = \{x_2\}$ são unidos em um novo grupo C_1, e os grupos $C_6 = \{x_6\}$ e $C_7 = \{x_7\}$ são unidos em um novo grupo C_2. Assim, a solução obtida com a aplicação do algoritmo hierárquico considerando o valor de corte 3 possui 5 grupos ($k = 5$): $C_1 = \{x_1, x_2\}$, $C_2 = \{x_6, x_7\}$, $C_3 = \{x_3\}$, $C_4 = \{x_4\}$ e $C_5 = \{x_5\}$.

Já com base na Tabela 5, considerando o valor de corte 2, a solução obtida foi $C_1 = \{x_1, x_2\}$, $C_2 = \{x_3, x_4, x_5, x_6, x_7\}$ ($k = 2$). A seguir, com base no valor de corte 2, são relatados os passos executados pelo algoritmo hierárquico:

- Inicialmente a solução é: C_1={ x_1}, C_2={ x_2 }, C_3={ x_3}, C_4={ x_4}, C_5={ x_5}, C_6={ x_6} e C_7= {x_7 }. (k = 7)
- Objetos x_1 e x_2 devem pertencer ao mesmo grupo ($m_{12} \geq 2$). Então deve ocorrer a união dos grupos C_1 e C_2 .
    - **Solução parcial:** C_1={ x_1, x_2 }, C_3={ x_3}, C_4={ x_4}, C_5={ x_5}, C_6={ x_6}, C_7={ x_7 }. (k = 6)
- Objetos x_3 e x_4 devem pertencer ao mesmo grupo ($m_{34} \geq 2$). Então deve ocorrer a união dos grupos C_3 e C_4 .
    - **Solução parcial**: C_1={ x_1, x_2 }, C_3={ x_3, x_4}, C_5={ x_5}, C_6={ x_6}, C_7={ x_7 }. (k = 5)
- Objetos x_3 e x_5 devem pertencer ao mesmo grupo ($m_{35} \geq 2$). Então deve ocorrer a união dos grupos C_3 e C_5 .
    - **Solução parcial:** C_1={ x_1, x_2 }, C_3={ x_3, x_4, x_5}, C_6={ x_6}, C_7={ x_7 }. (k = 4)
- Objetos x_4 e x_5 já pertencem a um mesmo grupo.
- Objetos x_5 e x_7 devem pertencer ao mesmo grupo ($m_{57} \geq 2$). Então deve ocorrer a união dos grupos C_3 e C_7 .
    - **Solução parcial:** C_1={ x_1, x_2 }, C_3={ x_3, x_4, x_5, x_7 }, C_6={ x_6 }. (k = 3)
- Objetos x_6 e x_7 devem pertencer ao mesmo grupo ($m_{67} \geq 2$). Então deve ocorrer a união dos grupos C_3 e C_6.
    - **Solução parcial:** C_1={ x_1, x_2 }, C_3={ x_3, x_4, x_5, x_6, x_7 }. (k = 2)
- Nenhuma união deve ocorrer. Última solução parcial é a solução obtida.
    - **Solução obtida:** C_1={ x_1, x_2 }, C_3={ x_3, x_4, x_5, x_6, x_7 }. (k = 2)

É importante destacar que a ordem das operações (uniões) realizadas não altera o resultado final.

4. Estatística de Hopkins

O Teste de Tendência de Agrupamento, também descrito como um Teste de Aleatoriedade Espacial, como o próprio nome sugere, consiste em verificar se em uma instância existe uma tendência à formação de agrupamentos de um dado conjunto de objetos X com n unidades $X = \{x_1, x_2, ..., x_n\}$, em um espaço p-dimensional, $x_i = \{x_{i1}, x_{i2}, ..., x_p\}$ [Banerjee 2004] [Han and Kamber, 2012]. Nesse sentido, a Estatística de Hopkins (EH) utiliza um critério interno, em que nenhuma informação a priori é necessária para a realização das análises. Além do conjunto X, dois outros conjuntos são considerados nessa abordagem, sejam eles:

X^*: trata-se de uma amostra do conjunto X ($X^* \subset X$) com m objetos que são selecionados de maneira aleatória.

A: possui m objetos construídos artificialmente com valores aleatórios no espaço de cada uma das p-dimensões.

Após a definição dos conjuntos de objetos utilizados, devem ser apresentadas as distâncias utilizadas:

w_j : distância entre um objeto $x_j^* \in X^*$ até o objeto de $X/\{x_j^*\}$ mais próximo.

u_j : distância entre um objeto $a_j \in A$ até o objeto mais próximo em X .

Na Equação 12, para cada objeto, são consideradas as distâncias w_j e u_j . Busca-se a maximização de H , cujo valor pertence ao intervalo [0,1]. Em uma instância em que objetos estão em grupos bem definidos, coesos e bem separados, a distância média entre os objetos é pequena. Isso implica, por sua vez, que o somatório de w_j tende a ser próximo de 0 e, consequentemente, H é próximo de 1. Já em instâncias em que os objetos estão dispersos no espaço, os somatórios de w_j e u_j são próximos, ou seja, o valor de H é próximo a 0,5.

$$\max H = \frac{\sum_{j=1}^{m} u_j}{\sum_{j=1}^{m} u_j + \sum_{j=1}^{m} w_j} \tag{12}$$

Conforme [Banerjee 2004], existem três classes em que a instância pode ser classificada:

- **Objetos são regularmente espaçados:** instância sem tendência à formação de agrupamentos. Em resultados da literatura, para essa classe, o valor de H variou no intervalo (0;0,3].
- **Objetos distribuídos de maneira aleatória no espaço:** indica que o conjunto de objetos não tem uma estrutura propícia para o agrupamento (H próximo a 0,5).
- **Existe uma tendência à formação de agrupamentos:** existem grupos bem definidos. Em resultados da literatura, para essa classe, o valor de H variou no intervalo [0,7;1).

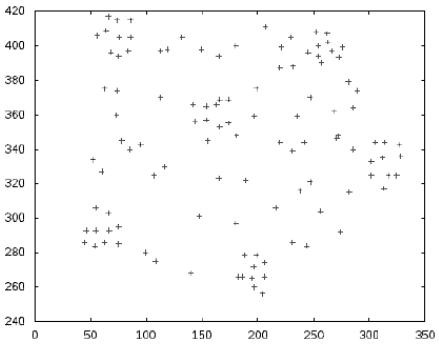
4.1. Experimentos Computacionais: Estatística de Hopkins

Como foi apresentado no início da seção, a EH utiliza uma amostra do conjunto de objetos da instância ($X^* \subset X$) e um conjunto de objetos artificiais A, cujos atributos possuem valores aleatórios no espaço de cada uma das p-dimensões. Uma vez que fatores aleatórios foram considerados (tanto em X^* quanto em A), tornou-se necessário

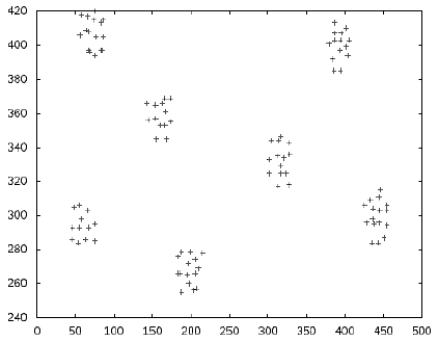
o desenvolvimento de um algoritmo que realizasse diferentes execuções com o objetivo de obter estatísticas para os valores de H. Nesse sentido, a cada iteração o algoritmo seleciona objetos para a formação do conjunto X^* e constrói objetos artificiais para o conjunto A.

As Figuras 3 (a) e (b) apresentam duas instâncias e as medianas dos valores de H para a instância 100p7c1 classificada como "*não comportada*" que possui 112 objetos e para a instância 100p7c, em que é possível identificar 7 grupos bem definidos (coesos e bem separados), com 100 objetos. Nesse exemplo o algoritmo foi executado 1000 vezes.

Figura 3: Instâncias 100p7c1 e 100p7c. H* é a mediana dos valores H em 1000 iterações.



(a) *100p7c1*: *H**=0,516



(b) *100p7c*: *H**=0,831

No experimento da presente subseção foram consideradas 51 instâncias propostas e utilizadas por [Cruz, 2010], que possuem entre 100 e 2000 objetos. Além disso, 63% das instâncias utilizadas possuem grupos bem definidos (denominadas "*comportadas*") e as demais 37% são as instâncias denominadas "*não comportadas*", conforme a classificação indicada por [Cruz, 2010]. É importante ressaltar que a classificação proposta pelo autor foi realizada em uma análise visual, durante a construção das instâncias. Dessa forma, a hipótese é que uma instância classificada como *comportada* também seja classificada como instância *com tendência à formação de agrupamentos*.

Embora possam surgir outras hipóteses relacionando instâncias *não comportadas* ou mesmo as instâncias sem tendência *à formação de agrupamentos*, o foco do presente trabalho está em utilizar o método baseado em coassociação apenas em instâncias consideradas *comportadas* e que possuam *tendência à formação de agrupamentos.*

Em experimentos preliminares foram utilizados conjuntos de amostras (X^*) com tamanhos 1%, 3%, 5%, 10% e 15% em relação à quantidade de objetos da instância. Além disso, foram consideradas as seguintes quantidades de execuções: 10, 100, 500 e 1000. Após a análise dos resultados obtidos nos experimentos preliminares, foi selecionada uma configuração em que o tamanho da amostra é de 1% e o algoritmo foi executado 100 vezes. Essa escolha deve-se ao fato da compatibilidade entre os resultados concernentes ao valor de H e, além disso, do reduzido custo computacional necessário. A Tabela 6 apresenta as estatísticas dos resultados obtidos com a configuração selecionada considerando, separadamente, apenas as instâncias "*comportadas*" e as instâncias "*não comportadas*".

Tabela 6: Resultados da Estatística de Hopkins com 1% de amostra e 100 execuções.

	Estatística Hopkins (H)		
Instância	**Não Comportadas**	**Comportadas**	**Tempo (s)**
Maior	0,95	***0,96***	0,03
Menor	0,45	***0,77***	0,00
Média	0,66	***0,90***	0,01
Mediana	0,64	***0,90***	0,00

Ainda com base na Tabela 6, a EH identificou a Tendência de Agrupamentos em todas as instâncias consideradas comportadas, em que a média e a mediana foram 0,9.

Em relação ao conjunto de instâncias não comportadas a média e a mediana foram inferiores a 0,7. É importante ressaltar que os valores extremos (menor e maior), embora sejam apresentados na tabela, não são considerados na análise.

A explicação para isto é que, eventualmente, configurações dos objetos do conjunto de amostra e dos objetos artificiais podem resultar em falsos positivos, ou seja, indicar tendência à formação onde não existe (H =0,95 para uma das instâncias não comportadas). É importante ressaltar que uma instância classificada como não comportada pelo autor (em [Cruz, 2010]) não necessariamente corresponde a uma instância sem tendência à formação de agrupamentos.

A partir dos resultados e análises realizadas na presente subseção foram selecionadas 19 instâncias para a realização dos experimentos com o método proposto no presente trabalho, em que todas são classificadas como comportadas e possuem tendência à formação de agrupamentos. Essas instâncias possuem entre 100 e 2000 objetos e entre 3 e 26 grupos.

5. Experimentos Computacionais

É uma prática comum em abordagens sistemáticas, para a identificação da quantidade ideal de grupos em problemas de agrupamento automático, utilizar $k = 2,...,k_{\max}$, sendo $k_{\max} = \sqrt{n}$ (vide [Pal and Bezdek, 1995][Pakhira et al., 2005][Campello et al., 2009]. Em [Han and Kamber, 2012], entretanto, um método simples para a estimativa do número ideal de grupos consiste em utilizar valores inteiros de k próximos a $\sqrt{n/2}$, na expectativa de que cada grupo possua cerca de $\sqrt{2n}$ objetos. Com o objetivo de cobrir os intervalos apresentados na literatura, no presente trabalho foi considerado $k = 2,...,\sqrt{n}$.

O método proposto considera um conjunto S, composto por 28 soluções obtidas por meio da utilização do *Método de Classificação Baseado em Densidade* proposto por [Semaan, et al., 2013]. No método do presente trabalho, a cada iteração, um subconjunto de soluções $S^* \subset S$ deve ser selecionado. Para isso, foi utilizado o método de *seleção por torneio*, muito utilizado em trabalhos relacionados à metaheurística *Algoritmo Genético* [Reeves, 2010] [Linden, 2012]. Este subconjunto corresponde a um conjunto base constituído por cinco soluções.

Nos experimentos realizados, o torneio selecionou três soluções, e a melhor entre elas é adicionada em S^*. Uma vez que esse subconjunto possui tamanho cinco, a matriz consenso é a soma das matrizes das cinco soluções de S^*. Nesse sentido, o algoritmo hierárquico será executado com cinco valores de corte (corte = 1,…,5) e, para cada valor de corte, a solução obtida deve ser avaliada por meio da utilização do índice silhueta. A melhor solução obtida entre os diferentes valores de corte é armazenada. É importante destacar que são consideradas apenas soluções válidas (viáveis), em que a quantidade de grupos está no intervalo pré-determinado ($k = 2,...,\sqrt{n}$).

A Tabela 7 apresenta comparativos entre as soluções selecionadas (conjunto base S^*) e os resultados obtidos com o método proposto. A coluna *Conjunto Base* possui duas colunas, com a média e a maior silhueta do conjunto S^* para cada instância. Já a coluna *Método Proposto* apresenta o gap (Equação 13) entre a melhor solução obtida no método proposto e os resultados da coluna *Conjunto Base*. Conforme os resultados

apresentados nesta tabela, em 15 das 19 instâncias os resultados obtidos foram melhores que a média de S^* ($gap > 0$). Além disso, para as demais quatro instâncias, os resultados foram equivalentes à média do conjunto base ($gap = 0$).

Ainda conforme a Tabela 7, em relação aos melhores resultados, o método proposto empatou com o melhor resultado do conjunto S^* em 16 instâncias e alcançou resultados superiores para 3 instâncias. Essa melhoria indica que o método, mesmo sem possuir buscas locais para o refinamento de soluções, com base nos padrões obtidos nas soluções de S^*, foi capaz de formar soluções novas. Além disso, destaca-se que a utilização do método resultou em soluções de qualidade equivalente ou superior ao conjunto de soluções submetido como entrada.

$$gap = s_{obtido} - s_{referencia} \quad (13)$$

A Tabela 8 apresenta comparativos entre os resultados do método proposto e de alguns métodos da literatura. A coluna *Algoritmos da Literatura* possui os números de grupos (k) e *Índices Silhueta* das melhores soluções obtidas por [Cruz, 2010] e [Semaan, et al., 2013]. Destaca-se que embora em [Cruz, 2010] sejam propostos vários algoritmos heurísticos, os *Algoritmos Evolutivos* destacaram-se. Já em [Semaan et al., 2012] foi proposto um *Método de Classificação Baseado em Densidade*. A coluna *Método Proposto* apresenta o número de grupos e os *gaps* do melhor resultado obtido com o método proposto em relação aos métodos da literatura que foram considerados.

Com base nos comparativos com o método sistemático proposto em [Semaan et al., 2012], os resultados obtidos pelo método proposto foram equivalentes ou superiores em 16 das 17 instâncias ($gap_2 \geq 0$). Em relação à única instância em que a quantidade de grupos foi diferente (100p10c), a solução obtida pelo método proposto foi superior ($gap_2 > 0$), e está em conformidade com o número apresentado pelo trabalho da literatura com o melhor resultado [Cruz, 2010].

Tabela 7: Comparativos entre as soluções selecionadas (conjunto base) e os resultados obtidos com o método proposto.

	Conjunto Base (Silhueta)		Método Proposto (*gap*)	
Instância	**Média**	**Maior**	**Média**	**Maior**
100p3c	0,768	0,786	**0,017**	**0**
100p7c	0,828	0,834	**0,006**	**0**
100p10c	0,544	0,692	**0,289**	**0,142**
200p4c	0,772	0,772	0	**0**
300p3c	0,751	0,766	**0,016**	**0**
400p3c	0,775	0,799	**0,023**	**0**
500p3c	0,812	0,825	**0,013**	**0**
600p15c	0,762	0,781	**0,019**	**0**
700p4c	0,693	**0,786**	**0,104**	**0,010**
800p23c	0,696	0,787	**0,084**	**0**
900p5c	0,714	0,716	**0,002**	**0**
900p12c	0,791	0,841	**0,034**	**0**
1000p6c	0,736	0,736	0	**0**
1000p14c	0,808	**0,808**	0	**0**
1300p17c	0,778	0,806	**0,027**	**0**
1800p22c	0,698	0,791	**0,093**	**0**
1900p24c	0,713	0,788	**0,081**	**0,006**
2000p11c	0,713	0,713	0	**0**
2000p26c	0,735	0,789	**0,054**	**0**

Tabela 8: Comparativos com resultados da literatura.

	Algoritmos da Literatura						
	[Cruz, 2010][1]		**[Semaan et al., 2013][2]**		**Método Proposto**		
Instância	***k***	**Silhueta**	***k***	**Silhueta**	***k***	***gap*[1]**	***gap*[2]**
100p3c	3	0,786	3	0,786	**3**	**0**	**0**
100p7c	7	0,834	7	0,834	**7**	**0**	**0**
100p10c	10	0,834	*8*	*0,691*	**10**	**0**	**0,143**
200p4c	4	0,773	4	0,773	**4**	**0**	**0**
300p3c	3	0,766	3	0,766	**3**	**0**	**0**
400p3c	3	0,799	3	0,799	**3**	**0**	**0**
500p3c	3	0,825	3	0,825	**3**	**0**	**0**
600p15c	15	0,781	15	0,781	**15**	**0**	**0**
700p4c	4	0,797	4	**0,797**	**4**	**0**	**0**
800p23c	23	0,787	23	0,787	**23**	**0**	**0**
900p5c	5	0,716	5	0,716	**5**	**0**	**0**
900p12c	12	0,841	12	0,841	**12**	**0**	**0**
1000p6c	6	0,736	6	0,736	**6**	**0**	**0**
1000p14c	14	0,831	15	**0,831**	15	-0,023	-0,023
1300p17c	17	0,830	18	0,806	18	-0,024	**0**
1800p22c	22	0,804	23	0,791	23	-0,013	**0**
2000p11c	11	0,713	11	0,713	**11**	**0**	**0**

* **As instâncias *1900p24c* e *2000p26c* foram disponibilizadas pelo autor de [Cruz, 2010] mas não têm resultados relatados na literatura.**

Segundo os resultados relatados para as instâncias 700p4c e 1000p14c nas Tabelas 7 e 8, é possível observar que as melhores soluções apresentadas por [Semaan et al., 2012] para essas instâncias não foram selecionadas para os respectivos conjuntos S^*.

Esta situação pode ter ocorrido pelo fato de a seleção de soluções realizada pelo torneio ser aleatória e, além disso, não ter sido utilizado nenhum procedimento de elitismo [Linden, 2012] (não é garantido que a melhor solução S será adicionada em S^*). Entretanto, mesmo sem a seleção da melhor solução e também sem a utilização de procedimentos de refinamento (buscas locais), o método proposto foi capaz de obter o melhor resultado da literatura para a instância 700p4c apenas utilizando os padrões identificados.

Conforme os resultados de [Cruz, 2010], também apresentados na Tabela 8, é possível observar que o número de grupos identificado como *ideal* no método proposto é equivalente aos relatados em 14 das 17 instâncias. Além disso, para as 3 instâncias em que os números não foram equivalentes, a diferença foi de apenas uma unidade. Essa pequena diferença deve ser destacada, pois, ainda em [Cruz, 2010], são apresentados métodos exatos para o Problema de Agrupamento em que o número de grupos submetido ao método está no intervalo [k-3, k+3] (k obtido por meio da utilização de uma heurística baseada em Algoritmo Genético).

Dessa forma, todos os números de grupos identificados pelo método proposto estão no intervalo utilizado por uma abordagem exata utilizada em [Cruz, 2010].

6. Conclusões e trabalhos futuros

O presente trabalho trouxe a proposta de um método baseado em combinação de soluções que considera a técnica de matriz de coassociação para a identificação do número ideal de grupos em problemas de agrupamento automático. Para isso, foi utilizado o índice silhueta, que combina as características como coesão e separação.

A técnica utilizada calcula a similaridade entre dois objetos por meio do número de grupos compartilhados entre eles nas soluções do conjunto base. Em seguida, com base em uma matriz de coassociação, um algoritmo para agrupamento hierárquico realiza cortes, formando soluções que foram avaliadas por meio do índice silhueta.

Os resultados apresentados neste estudo indicam que o método proposto foi capaz de identificar padrões nas soluções do conjunto base, obtidas com a utilização do método de classificação baseado em Densidade proposto em [Semaan et al., 2012]. Em relação aos índices de silhueta do conjunto base, as soluções obtidas com a implementação do método proposto foram equivalentes ou superiores.

Além disso, nos comparativos com trabalhos da literatura em relação ao número de grupos, os melhores resultados foram obtidos em 14 das 17 instâncias utilizadas e, nas 3 instâncias em que as silhuetas obtidas foram inferiores, a diferença no número de grupos foi de apenas uma unidade.

Não obstante, de forma a reforçar ainda mais esta análise, em trabalhos futuros serão efetuados novos experimentos com mais instâncias da literatura. Uma proposta para trabalhos futuros é a utilização de buscas locais nos grupos formados em cada solução obtida pelo algoritmo hierárquico aglomerativo. Assim, para cada valor de corte, os grupos da solução podem ser unidos com o objetivo de maximizar o índice silhueta e obter novas soluções, percorrendo um novo espaço de busca.

No que concerne ao método proposto na seção 3.2, será desenvolvido um novo trabalho em que as soluções produzidas pelo algoritmo de agrupamento hierárquico serão avaliadas mediante a utilização do coeficiente de correlação cofenética [Bussab, Miazaki e Andrade, 1990]. Além disso, adotar-se-á nesse novo trabalho a distância de Mahalanobis para avaliar a dissimilaridade entre os objetos.

Referências bibliográficas

Alves et al. 2006] Alves, V., R. Campello, & E. Hruschka (2006). Towards a fast evolutionary algorithm for clustering. In IEEE Congress on Evolutionary Computation, 2006, Vancouver, Canada, pp. 1776–1783.

Banerjee 2004] Banerjee, A. Validating clusters using the hopkins statistic. IEEE International Conference on Date of Conference, 2004.

Baum, 1986] Baum, E.B. Iterated descent: A better algorithm for local search in combinatorial optimization problems. Technical report Caltech, Pasadena, CA. Manuscript, 1986.

Bussab, Miazaki e Andrade, 1990] Bussab, W.O., Miazaki, E.S., Andrade D.F. (1990). Introdução à Análise de Agrupamentos, IME- USP, 9º Simpósio Brasileiro de Probabilidade e Estatística.

Campello et. al., 2009b] Campello, R.J.G.B., Hruschka, E.R., Alves, V.S. (2009) On comparing two sequences of numbers and its applications to clustering analysis. Information Sciences 129(8).

Campello et al., 2009] Campello, R. J. G. B., E. R. Hruschka, & V. S. Alves (2009). On the efficiency of evolutionary fuzzy clustering. Journal of Heuristics 15 (1), 43–75.

Campello et al., 2009b] Campello, R.J.G.B., Hruschka, E.R., Alves, V.S. On comparing two sequences of numbers and its applications to clustering analysis. Informatin Sciences 129(8), 2009.

Cruz, 2010] Cruz, M. D. O Problema de Clusterização Automática. Tese de Doutorado, COPPE\UFRJ, Rio de Janeiro, 2010.

Dias and Ochi, 2003] Dias, C.R.; & Ochi, L. S. Efficient Evolutionary Algorithms for the Clustering Problems in Directed Graphs. Proc. of the IEEE Congress on Evolutionary Computation (IEEE-CEC), 983-988. Canberra, Austrália, 2003.

Ester et al., 1995] Ester, M., Kriegel, H.-P., and Xu, X., A Database Interface for Clustering in Large Spatial Databases, In: Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining (KDD-95), pp. 94- 99, Montreal, Canada, August, 1995.

Ester et al., 1996] Ester, M., H.-P. Kriegel, J. Sander, & X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp. 226–231.

Freund and Schapire, 1996] Freund, Y., Schapire, R. (1996). Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning. Bari, Italy, pp. 148-156.

Frossyniotis, Likas and Stafylopatis, 2004] Frossyniotis, D., Likas, A., Stafylopatis, A. (2004). A clustering method based on boosting. Pattern Recognition Letters, 25, pp. 641-654.

Frossyniots, Pertselakis and Stafylopatis, 2002] Frossyniots, D., Pertselakis, M., Stafylopatis, A. (2002). A multiclustering fusion algorithm. In: Proceedings of the Second Hellenic Conference on Artificial Intelligence, April 11-12 LNAI 2308. Springer-Verlag, Thessaloniki, Greece, pp. 225-236.

Goldschimidt e Passos, 2005] Goldschimidt R.; Passos, E. Data Mining: um guia prático. Editora Campus, Rio de Janeiro: Elsevier, 2005.

Han and Kamber, 2012] Han, J., e Kamber, M., Cluster Analysis. In: Morgan Kaufmann. Publishers (eds.), Data Mining: Concepts and Techniques, 3 ed., chapter 8, New York, USA, Academic Press, 2012.

Hruschka and Ebecken, 2001] Hruschka, E. R., Ebecken, N. F. F. A Genetic algorithm for cluster analysis. IEEE Transactions on Evolutionary Computation, 2001.

Hruschka and Ebecken, 2003] Hruschka, E. R. & Ebecken, N. F. F. (2003). A genetic algorithm for cluster analysis. Intelligent Data Analysis 7 (1), 15–25.

Hruschka et al., 2004a] Hruschka, E. R., R. J. G. B. Campello, & L. N. de Castro (2004a). Evolutionary algorithms for clustering gene-expression data. In Proc. IEEE Int. Conf. on Data Mining, Brighton/England, pp. 403–406.

Hruschka et al., 2004b] Hruschka, E. R., R. J. G. B. Campello, & L. N. de Castro (2004b). Improving the efficiency of a clustering genetic algorithm. In Advances in Artificial Intelligence - IBERAMIA 2004: 9th Ibero-American Conference on AI, Puebla, Mexico, November 22-25. Proceedings, Volume 3315, pp. 861–870. Springer-Verlag GmbH, Lecture Notes in Computer Science.

Hruschka et al., 2006] Hruschka, E. R., R. J. G. B. Campello, & L. N. de Castro (2006). Evolving clusters in gene-expression data. Information Sciences 176 (13), 1898–1927.

Jain and Dubes, 1988] Jain, A. & R. Dubes (1988). Algorithms for Clustering Data. Prentice Hall.

James et al., 2013] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning – with Applications in R. Springer.

Johnson A.R. and Wichern D.W., 2012]. Applied Multivariate Statistical Analysis. Prentice Hall. Sixth Edition.

Kumar et al., 2009] Kumar, V.; Steinbach, M.; Tan, P. N. Introdução ao Data Mining - Mineração De Dados. Ciência Moderna.

Larose, 2005] Larose, D. T. Discovering Knowledge in Data, An Introduction to Data Mining. John Wiley & Sons, 2005.

Linden, R., 2012] Algoritmos Genéticos, Editora Ciência Moderna.

Ma et al., 2006] Ma, P. C. H., K. C. C. Chan, X. Yao, & D. K. Y. Chiu (2006). An evolutionary clustering algorithm for gene expression microarray data analysis. IEEE Trans. Evolutionary Computations 10 (3), 296–314.

Naldi et al., 2009] Naldi, M. C.; Faceli, K.; Carvalho, A. C. P. L. F.. Uma Revisão Sobre Combinação de Agrupamentos. Revista de Informática Teórica e Aplicada, v. 16, p. 25-51, 2009.

Naldi and Carvalho, 2007] Naldi, M. C. & A. C. P. L. F. Carvalho (2007). Clustering using genetic algorithm combining validation criteria. In Proceedings of the 15th European Symposium on Artificial Neural Networks, ESANN 2007, Volume 1, pp. 139–144. Evere.

Naldi, 2011] Naldi, C. N. Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados. Tese de Doutorado, USP - São Carlos, 2011.

Pakhira et al., 2005] Pakhira, M., S. Bandyopadhyay, & U. Maulik (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. Fuzzy Sets Systems 155 (2), 191–214.

Pal and Bezdek, 1995] Pal, N. & J. Bezdek (1995). On cluster validity for the fuzzy c-means model. IEEE Transactions of Fuzzy Systems 3 (3), 370–379.

Pan and Cheng, 2007] Pan, S. & K. Cheng (2007). Evolution-based tabu search approach to automatic clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C - Applications and Reviews 37 (5), 827–838.

Pelleg and Moore, 2000] Pelleg, D. & A. Moore (2000). X-means: extending k-means with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727–734.

Reeves 2010] Reeves, C. R. Genetic algorithms. In Glover, F. and Kochenberger, G., editors, Handbook of Metaheuristics, pages 109–139. Kluwer Academic Publishers, 2010.

Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

Semaan et al., 2012] Semaan, G. S., Cruz, M.D., Brito, J. A. M., and Ochi, L. S. (2012) "Proposta de um método de classificação baseado em densidade para a determinação do número ideal de grupos em problemas de clusterização", vol. 10 número 4.

Semaan, 2013] Semaan, G. S. (2013) Algoritmos para o Problema de Agrupamento Automático, Tese de Doutorado, UFF - Niterói.

Soares and Ochi, 2004] Soares, S. S. R. F., Ochi, L. S. Um Algoritmo Evolutivo com Reconexão de Caminhos para o Problema de Clusterização Automática. in XII Latin Ibero American Congress on Operations Research, 2004, Havana. Proc. of the XII CLAIO (em CD-ROM). ALIO, 2004. v.1, p. 7 -13.

Soares, 2004] Soares, A. S. R. F. Metaheurísticas para o Problema de Clusterização Automática, Dissertação de Mestrado, UFF - Niterói, 2004.

Tseng and Yang, 2001] Tseng, L. & . Yang, S.B. (2001). A genetic approach to the automatic clustering problem. Pattern Recognition 34, 415–424.

Vu et al., 2010] Vu, V., Labroche, N., Bouchon-Bernier, B. (2010). Boosting Clustering by Active Constraint Selection. In Proceeding of: ECAI 2010- 19th European Conference on Artificial Intelligence, Lisbon, Portugal.

Zalik, 2008] An Efficient K'-Means Clustering Algorithm, Pattern Recognition Letters 29, 2008.

Agradecimentos

Os autores agradecem ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Combinação em série e em paralelo de modelos de Redes Neurais e Regressão Logística – Um estudo de caso em Cross-Selling

Sabrina Zanatta Grebin [1]
Lisiane Priscila Roldão Selau [2]

Resumo

Como resultado ao crescente desenvolvimento tecnológico, computadores mais potentes tornam possível o armazenamento diário de grande quantidade de dados. As técnicas de mineração de dados surgem como uma alternativa inteligente e eficaz para transformá-los em conhecimento. Este trabalho se propôs a resolver um problema real de venda-cruzada (*cross-selling*) de uma instituição financeira e, com o objetivo de contribuir para o desenvolvimento das técnicas de *data mining*, realizou-se uma comparação entre duas técnicas consagradas, regressão logística e redes neurais, e entre duas formas de combinação das mesmas, em série (*hybrid*), na qual a regressão logística é utilizada para selecionar as variáveis que irão entrar na rede neural, e em paralelo (*ensemble*), na qual os resultados das técnicas individuais são combinados com base em suas decisões. As comparações entre o desempenho das técnicas individuais e dos métodos de combinação indicam que os métodos de combinação são ligeiramente superiores às técnicas de modelagem individuais.

Palavras-chave: Cross-Selling, Regressão Logística, Redes Neurais, Hybrid, Ensemble.

[1] *Departamento de Estatística – Instituto de Matemática – UFRGS.*
[2] *Departamento de Estatística – Instituto de Matemática – UFRGS.*

1. Introdução

Como resultado ao crescente desenvolvimento tecnológico nas últimas décadas, computadores cada vez mais potentes tornam possível o armazenamento diário de grande quantidade de dados. Empresas dos mais diversos setores e de todos os tamanhos buscam transformar essa grande massa de dados em informação útil para a tomada de decisão. Para tanto, são necessárias ferramentas e técnicas capazes de extrair esse conhecimento. Devido à crescente concorrência do mercado, empresas buscam antecipar as necessidades e preferências dos seus clientes. Segundo Barry e Linoff (2004), para melhorar o relacionamento com os clientes é preciso aprender o seu comportamento, para que com base nesse conhecimento seja possível aumentar a rentabilidade do negócio, de modo a se obter clientes mais satisfeitos e fiéis.

Desde 1960, o processamento de dados vem sendo migrado para sistemas cada vez mais sofisticados e poderosos, evoluindo, na década de 1970, para sistemas relacionais e seguindo em direção ao desenvolvimento de sistemas de banco de dados avançados (*advanced database systems*), armazenamento de dados (*data warehousing*) e mineração de dados (*data mining*), surgindo, na década de 1980, com análises avançadas de dados (HAN et al., 2012).

Simultaneamente ao avanço tecnológico, o constante aumento dos conjuntos de dados e, consequentemente, das variáveis que estão sujeitas a se relacionarem, vêm tornando os métodos tradicionais de análise (como correlações bivariadas) menos eficientes e, por vezes, inadequados (SILVA, 2009). As técnicas de *data mining* surgem então como uma alternativa inteligente e eficaz de identificar padrões de comportamento dos dados, transformando-os em conhecimento. Segundo Berry e Linoff (2004), elas tem por finalidade a classificação, estimação, previsão e agrupamento.

Segundo Fayyad *et al*.(1996), dentre as finalidades da mineração de dados, as duas mais utilizadas na prática são: a previsão, em que o valor previsto para determinada variável de interesse é explicado pelas demais variáveis presentes no banco de dados; e a descrição, em que são descritos os padrões existentes nas relações entre elas. Não raro, o objetivo está tanto em encontrar o padrão de relacionamento entre as variáveis quanto em prever um valor futuro para as mesmas. Alguns exemplos são: modelos de venda para estratégias de marketing, modelos de risco de crédito, segmentação de clientes, previsão de inadimplência e controle de fraudes.

Neste estudo, o foco da estratégia de *data mining* é a venda-cruzada (*cross-selling*) que, segundo Dyche (2001), se caracteriza pela venda de um produto ou serviço adicional como resultado de uma compra anterior e, quando corretamente realizada, significa vender o produto certo ao cliente certo. É consenso no marketing de relacionamento que é menos dispendioso e mais rentável investir mais nos clientes atuais do que adquirir um novo cliente.

Entre as técnicas mais utilizadas para a previsão e descrição estão a Regressão Logística, Análise Discriminante, Árvores de Decisão e Redes Neurais. Sendo que Regressão Logística e Redes Neurais têm sido as mais utilizadas (SELAU, 2012). Com o intuito de se obter ganhos no desempenho e na explicação do modelo obtido por meio dessas técnicas, autores vêm estudando formas de combiná-las. Dentre elas, a combinação em série, também conhecida como modelo híbrido (*hybrid*), e a combinação em paralelo, ou simplesmente combinação de previsões (*ensemble*). Segundo Selau (2012), a combinação em série consiste em utilizar em sequência duas técnicas distintas, com o intuito de melhorar o poder preditivo. A combinação em paralelo, que, segundo Werner (2005) vem sendo estudada desde Bates e Granger (1969), é um método bastante utilizado para a previsão em séries temporais com o intuito de minimizar os erros de previsão do modelo estimado e consiste basicamente em combinar os resultados obtidos por meio de duas ou mais técnicas individuais.

Nesse contexto, o objetivo deste trabalho é comparar o desempenho individual das técnicas de Regressão Logística e Redes Neurais e de duas formas de combinação: em série, em que a Regressão Logística é utilizada para selecionar as variáveis que irão entrar na Rede Neural, e em paralelo, em que os resultados das técnicas individuais são combinados com base em suas decisões. Para tanto, este trabalho se propõe a resolver um problema real de *cross-selling* de uma instituição financeira brasileira.

Este trabalho está estruturado em cinco seções, sendo a primeira a introdução já exposta, onde foram apresentadas as considerações iniciais acerca do problema de pesquisa. Na segunda seção são apresentados os conceitos e a fundamentação teórica dos modelos de *cross-selling*, das técnicas de modelagem e dos métodos de combinação utilizados no desenvolvimento deste trabalho. Na terceira seção é descrita a metodologia de pesquisa, bem como as etapas propostas para a construção dos modelos. A seção quatro ilustra os resultados e comparações obtidas por meio do método proposto. Por

fim, as principais conclusões e considerações são apresentadas na seção cinco, além de sugestões para trabalhos futuros.

2. Fundamentação Teórica

Nesta seção é apresentada uma revisão, com base na literatura, sobre as técnicas e métodos a serem utilizados no desenvolvimento deste trabalho.

2.1 Modelos de *cross-selling*

Para Berry e Linoff (2004), é preciso que as grandes empresas imitem as pequenas, criando um relacionamento com seus clientes, baseado no conhecimento sobre eles, para melhor atendê-los. Segundo HAN *et al*. (2012), no processo de descoberta de conhecimento em base de dados, conhecido por KDD (*Knowledge Discovery in Databases*), primeiramente os dados são preparados para mineração (que inclui limpeza dos dados, combinação com outras bases de dados, seleção e transformação de variáveis), então é realizada a mineração de dados e os padrões descobertos são transformados em conhecimento.

Para se estabelecer um relacionamento de longo prazo com o mercado, duas técnicas são utilizadas no Marketing de Relacionamento: *up-selling* e *cross-selling*. O conceito de *up-selling* é aumentar o valor da venda de um mesmo produto, já os modelos de *cross-selling*, traduzidos na literatura como "venda casada de produtos" ou simplesmente "venda-cruzada", buscam nos dados o conhecimento necessário para identificar o perfil de clientes mais propensos a adquirir um ou mais produtos adicionais (BERRY e LINOFF, 2004).

O objetivo ao se realizar uma modelagem de *cross-selling* é concentrar mais esforços nos clientes atuais, cujo custo de aquisição é menor em relação a clientes novos, porque os clientes atuais já possuem um relacionamento com a empresa. O uso adequado da técnica consiste em identificar que produto ou serviço oferecer a qual cliente e em que momento, com o intuito de que estes venham a adquirir mais produtos ou serviços com a empresa, se tornando clientes fiéis. Consequentemente, o bom emprego de *cross-selling* implica em um aumento da satisfação dos clientes, da sua fidelidade e lucratividade (KAMAKURA *et al*., 1991).

Assim como nas outras aplicações de *data mining*, os trabalhos em *cross-selling* também utilizam com bastante frequência as técnicas de Regressão Logística e de Redes Neurais. Como, por exemplo, Kisahleitner (2008) que compara o desempenho das técnicas de Regressão Logística, Árvores de Decisão e Redes Neurais na aquisição de clientes não correntistas para cartão de crédito de uma empresa do setor financeiro. Os resultados mostram que, para este estudo, as três técnicas utilizadas obtiveram desempenho similar. Silva (2009) e Adorno e Bueno (2011) fazem uso das técnicas de Redes Neurais e Regressão Logística que são comparadas para desenvolver um modelo de propensão ao consumo de um produto de crédito pessoal. Ambos destacam que as diferenças entre os melhores modelos originados de cada técnica não são muito significativas, porém o modelo de Redes Neurais obteve desempenho melhor. Knott *et al*. (2002) utilizaram-se de dados de um banco de varejo com o objetivo de aumentar a quantidade de produtos adquiridos por clientes, prevendo qual é o próximo produto que cada cliente possui maior probabilidade de comprar. Os autores utilizaram Regressão Logística, Logit Multinomial, Análise Discriminante e Redes Neurais com o intuito de desenvolver um modelo de *cross-selling* com a abordagem de *next-product-to-buy* (NPTB).

2.2 Técnicas de Modelagem Individuais

2.2.1 Regressão Logística

O modelo de Regressão Logística se caracteriza por ter como resposta uma variável dicotômica, sendo essa uma das diferenças entre este e o modelo de Regressão Linear (HOSMER e LEMESHOW, 1989). O objetivo da técnica é modelar a probabilidade de ocorrência de um evento, em que a ausência e a presença são usualmente denotadas por 0 e 1, respectivamente (DINIZ e LOUZADA, 2012). A Regressão Logística possui muitas vantagens frente à análise discriminante quando se tem um modelo com resposta dicotômica, pois não depende de suposições tão rígidas, tais como a normalidade das variáveis explicativas e a igualdade de matrizes de variâncias e covariâncias dos grupos (HAIR *et al*., 2005).

De acordo com Hair *et al*. (2005), ela possui muitas similaridades com a Regressão linear, como, por exemplo, os testes estatísticos e a capacidade de inserir efeitos não lineares no modelo (com a criação de variáveis *dummies*), mas difere no sentido de ter como objetivo prever a probabilidade de um evento ocorrer. No caso específico de *cross-selling*, o modelo define a relação entre a probabilidade de um cliente adquirir mais de um produto ou serviço e um conjunto de fatores ou atributos que o caracterizam. Esta relação é definida pela transformação *logit.*

A transformação *logit* consiste em aplicar a função logaritmo no *odds ratio*, ou razão de chances, que é a probabilidade do indivíduo assumir o evento de interesse quando a característica está presente (1), comparado com a ausência da mesma (0) (DINIZ e LOUZADA, 2012). A transformação *logit* é definida pela expressão apresentada na Equação 1.

$$\text{logit}\left(p_i\right) = \ln\left\{\frac{p_i}{1-p_i}\right\} = \beta_o + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}, \tag{1}$$

Em que p_i é a probabilidade de ocorrência do evento de interesse (variável resposta e é definida como

$$P\left(Y=1\right) = p_i = \frac{\exp\left(\beta_o + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}\right)}{1+\exp\left(\beta_o + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}\right)}, \tag{2}$$

onde $x_{1,i},\cdots,x_{k,i}$ são as covariáveis, ou variáveis explicativas, e $\beta_o,\beta_1,\cdots,\beta_k$ são os coeficientes do modelo. O efeito das covariáveis sobre a variável resposta é medido por meio dos seus respectivos coeficientes. Quando positivo indica um aumento na probabilidade prevista de ocorrer o evento, e negativo o efeito é contrário. No estudo de *cross-selling*, isso significa dizer que a variável que possui coeficiente positivo aumenta a probabilidade prevista de o cliente adquirir mais de um produto ou serviço, enquanto a variável que possui coeficiente negativo diminui a probabilidade prevista de que a compra aconteça.

A transformação *logit* possui natureza não linear e, então, os parâmetros do modelo são estimados por meio do método da máxima verossimilhança. De acordo com Hosmer e Lemeshow (1989), para testar a significância dos coeficientes, uma alternativa é a razão de verossimilhança, e Hair *et al.* (2005) sugerem a estatística de Wald.

Apesar da facilidade na explicação dos resultados e flexibilidade de utilização da técnica de Regressão Logística, ela está sujeita, assim como os outros modelos de Regressão, à multicolinearidade. De acordo com Hair *et al.* (2005), o ideal é que se tenha alta correlação das variáveis explicativas com a variável resposta, mas com baixa correlação entre elas próprias. Utilizar variáveis explicativas altamente correlacionadas no modelo pode resultar em estimativas errôneas dos coeficientes (HOSMER e LEMESHOW, 1989). Para avaliar a colinearidade, duas medidas são frequentemente utilizadas: o valor de tolerância e o seu inverso, o VIF – fator de inflação de variância, (HAIR *et al.*, 2005). Uma alternativa é o uso do método *stepwise* para seleção das variáveis, muito utilizado em Regressão linear e encontrado na maioria dos pacotes estatísticos, que consiste basicamente em incluir ou excluir variáveis do modelo de acordo com a sua importância, evitando a inclusão de variáveis altamente correlacionadas.

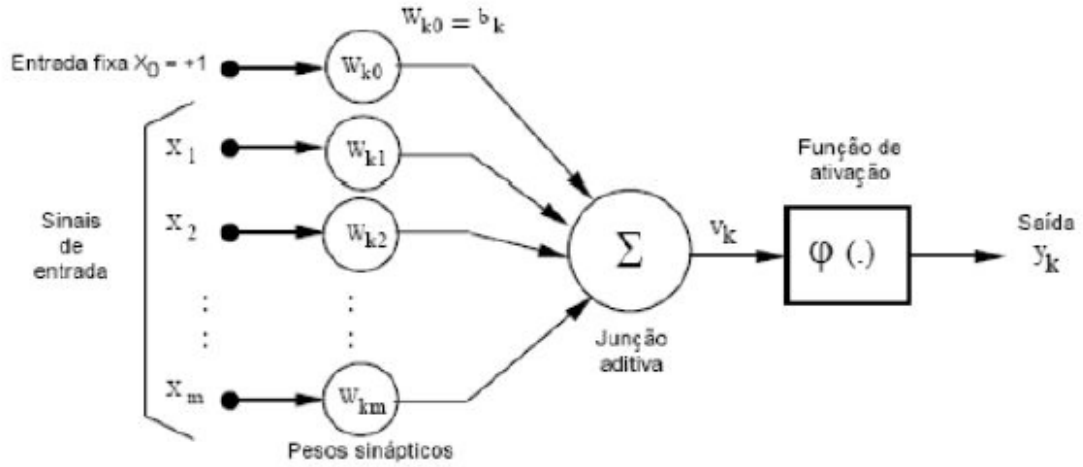
2.2.2 Redes Neurais

O cérebro humano possui muitas habilidades, dentre elas, a capacidade de relacionar situações novas com experiências passadas e de identificar, entender e interpretar características com grande eficiência. De acordo com Haykin (2001), desejando entender e reproduzir uma máquina que se aproximasse do cérebro humano, Mc Culloch e Pitts, Hebb e Rosenblat iniciaram entre os anos de 1940 e 1950, os primeiros trabalhos sobre Redes Neurais Artificiais (RN) que foram originados de estudos de Inteligência Artificial (IA). Uma RN busca explicar, com base na forma funcional das observações, a relação entre a variável resposta e as variáveis explicativas, possuindo o processo de aprendizagem como um grande diferencial em relação às demais técnicas de *data mining*.

A Rede Neural se assemelha ao cérebro humano no sentido de que o conhecimento é adquirido por meio de um processo de aprendizagem e armazenado por pesos sinápticos, nos quais o elemento fundamental é o neurônio, também chamado de nó. O uso de Redes Neurais possui benefícios como, por exemplo, a capacidade de lidar com a não linearidade das relações, realizar o processo de treinamento até que não haja mais mudanças significativas nos pesos sinápticos e a habilidade de se adaptar a modificações do ambiente na qual foi treinada, podendo alterar seus pesos sinápticos em tempo real. Os neurônios de uma Rede Neural são uniformes, o que permite que as mesmas teorias e algoritmos de aprendizagem sejam utilizados em diferentes aplicações (HAYKIN, 2001).

A Figura 1 apresenta o modelo não linear de um neurônio, comumente chamado de perceptron. Ele possui três elementos que processam a informação de entrada gerando as informações de saída: (*i*) um conjunto de sinapses; (*ii*) um somador; e (*iii*) uma função de ativação. Esse processamento é feito multiplicando cada variável de entrada por seu respectivo peso sináptico e então o valor resultante é processado por uma função de ativação, que restringe a amplitude de saída em um intervalo finito, resultando na informação que será a entrada para o nó seguinte. Este modelo também inclui um viés, que tem o efeito de aumentar (viés positivo) ou diminuir (viés negativo) os valores de entrada da função de ativação (HAYKIN, 2001).

FIGURA 1 - Modelo não linear de um neurônio [Haykin, 2001]



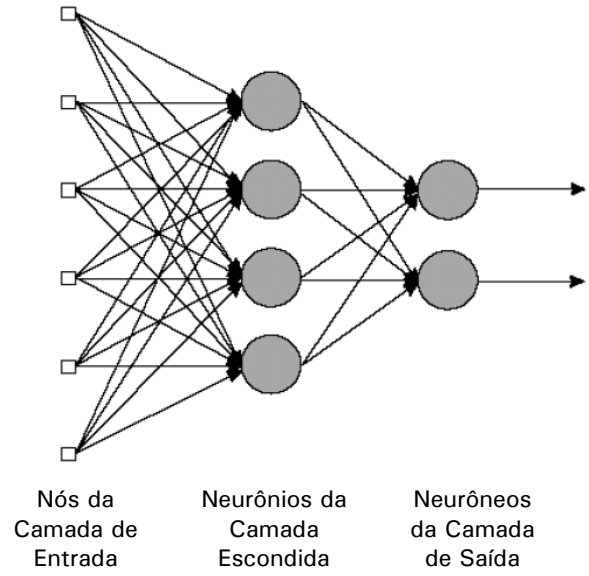
O modelo da Figura 1 pode ser descrito matematicamente com a Equação 3, que possui semelhança com o modelo de Regressão não linear múltipla, em que as entradas do neurônio $x_1, \cdots, x_m$ são as variáveis explicativas, a saída y_k é a variável resposta, os pesos sinápticos $w_{k1}, \cdots, w_{km}$ são os coeficientes de Regressão, $w_{k0} = b_k$, também chamado de viés, é o intercepto e funciona como uma constante da função aditiva, e $\varphi(.)$ é a função de ativação não linear.

$$y_k = \varphi(u_k + b_k) \tag{3}$$

$$\text{Onde: } u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{4}$$

A primeira camada de uma Rede Neural é a camada de entrada, as intermediárias são as camadas escondidas e a última é a camada de saída. O número de camadas e a quantidade de neurônios na rede devem ser determinados conforme a natureza do problema. Geralmente, o aumento do número de neurônios na rede é utilizado quando uma característica específica é importante, para assegurar um grau de precisão maior na tomada de decisão (HAYKIN, 2001). De acordo com Hair *et al*.(2005), as variáveis métricas necessitam de um neurônio para cada variável, enquanto que as não métricas precisam ser codificadas, criando categorias representadas por uma variável binária, que serão representadas, cada uma, por um neurônio. A Figura 2 apresenta uma Rede Neural na sua forma mais simples, uma rede de camada intermediária única.

Figura 2 - Modelo estrutural de camada única [Haykin, 2001]



Os fatores principais que caracterizam uma RN são: (*i*) arquitetura da rede: divididas em três classes, em que a primeira delas, redes alimentadas adiante com camada única, é a mais simples e consiste na projeção da camada de entrada sobre a de saída, mas não o inverso; a segunda, redes alimentadas diretamente com múltiplas camadas, se diferencia da primeira pela presença de neurônios na camada escondida e; a terceira, redes recorrentes, se diferencia por permitir a realimentação; (*ii*) tipo de treinamento: os parâmetros do modelo podem ser estimados de duas formas; pelo treinamento supervisionado, que se caracteriza por ter uma saída para cada vetor de entrada; ou por meio do não supervisionado, que não possui uma saída para cada entrada. O algoritmo mais utilizado no treinamento supervisionado é o Backpropagation, que busca encontrar os pesos sinápticos que minimizam o erro e (*iii*) função de ativação: as fundamentais são a função de limiar, função linear por partes e função sigmoide. Dado que a entrada de um neurônio na camada escondida é a combinação linear das saídas anteriores, as funções lineares não seriam adequadas. Portanto, no caso de modelos com neurônios na camada escondida, normalmente se usa a função Logística. Já na camada de saída a função varia de acordo com a natureza do problema. (HAYKIN, 2001).

De acordo com Selau (2012), como resultado da combinação não linear dos pesos sinápticos que ocorre na camada escondida, o modelo de Rede Neural não informa a influência direta das variáveis explicativas na variável resposta, dificultando na interpretação dessas variáveis. Nesse sentido, Hair et al. (2005) sugerem que se utilize a técnica de Redes Neurais para previsão e classificação quando o interesse está somente na precisão do modelo.

2.3 Métodos de Combinação

2.3.1 Combinação em série (*Hybrid*)

Conhecida na literatura como modelo híbrido, a combinação em série é um método que vem sendo foco de estudos recentes e que tem mostrado resultados promissores, sendo utilizada principalmente com o intuito de minimizar os inconvenientes das técnicas de inteligência artificial e melhorar a classificação, previsão e desempenho dos modelos (LEE *et al.*, 2002; HSIEH, 2005; LEE e CHEN, 2005; CHEN *et al.*, 2009; SELAU, 2012; TSAI e CHEN, 2010). Segundo Tsai e Chen (2010), o modelo híbrido em dois estágios consiste basicamente na combinação em sequência de duas técnicas diferentes de agrupamento ou de classificação. De acordo com Ghodselahi (2011), para desenvolver um modelo híbrido, a primeira técnica é utilizada para orientar o processamento da segunda.

Existem diversas maneiras de se combinar técnicas de modelagem individuais para o desenvolvimento de modelos híbridos. Ao se utilizar a abordagem híbrida em dois estágios, os tipos de métodos de combinação possíveis são: combinação de duas técnicas de classificação; combinação de duas técnicas de agrupamento; uma técnica de agrupamento combinada com uma de classificação; e uma técnica de classificação combinada com uma de agrupamento (TSAI e CHEN, 2010).

Lee *et al*. (2002) combinaram duas técnicas de classificação com o objetivo de melhorar a solução inicial e aumentar a precisão de classificação de um modelo de pontuação de crédito, fazendo uso da análise discriminante para selecionar as variáveis preditoras significativas que são então utilizadas como as variáveis de entrada do modelo de Redes Neurais, utilizando ainda, o resultado da previsão obtida na análise discriminante como informação extra na camada de entrada da Rede Neural. Comparando o desempenho dos modelos desenvolvidos, o modelo híbrido convergiu mais rápido e obteve uma maior acurácia que os modelos obtidos com uso das duas técnicas de classificação.

Com vista à necessidade de se comparar a eficiência entre as diferentes formas de combinações de modelos híbridos, no estudo pioneiro de Tsai e Chen (2010) foram desenvolvidos os quatro tipos de combinação de modelos híbridos em dois estágios. Como resultado, o melhor método de combinação indicado é entre duas técnicas de classificação, para o qual foram utilizadas a Regressão Logística e Redes Neurais, sendo que a Regressão Logística seleciona as variáveis significativas que serão utilizadas como nós de entrada na Rede Neural.

Não existe na literatura um método para selecionar as variáveis de entrada da Rede Neural (LEE *et al*., 2002). Desta forma, a modelagem híbrida torna-se uma alternativa eficiente, na qual essas variáveis são selecionadas por meio de outra técnica que é utilizada na modelagem em um estágio anterior a Rede Neural.

2.3.2 Combinação em paralelo (Ensemble)

O primeiro estudo acerca da combinação em paralelo, conhecida como combinação de previsões ou simplesmente *ensemble*, surgiu com Bates e Granger (1969), sendo consagrado em diversos estudos posteriores (CLEMEN, 1989; MAKRIDAKIS e HIBON, 2000; HIBON e EVGENIOU, 2005; WERNER, 2005; CONSTANTINE e PAPALARDO, 2010; MARTINS, 2011) como um método eficaz para se reduzir os erros gerados com a previsão. Segundo Clemen (1989), ao invés de escolher a melhor técnica de previsão a ser utilizada, são definidas, de acordo com o objetivo do estudo, quais técnicas poderiam aumentar a acurácia da previsão, de modo que cada técnica pode contribuir capturando algum tipo de informação intrínseca aos dados.

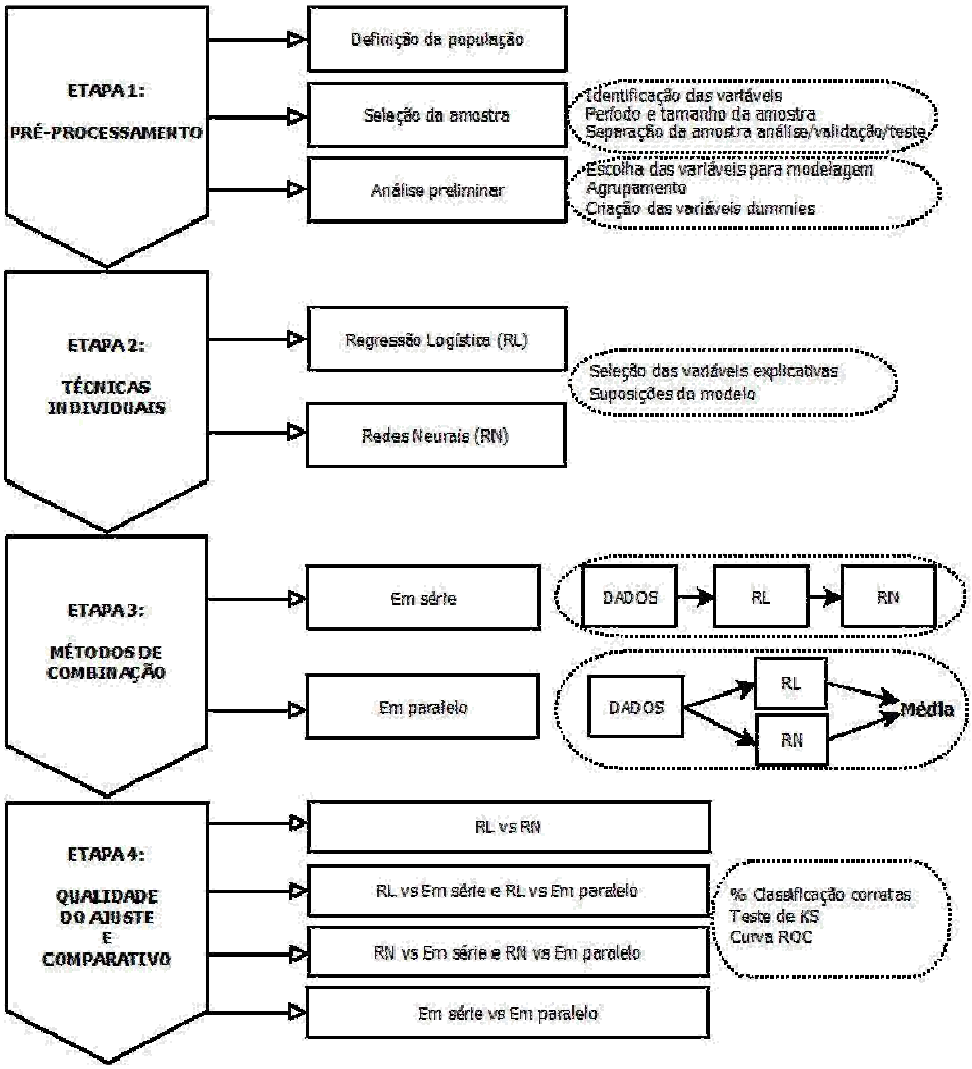
O método de combinação em paralelo consiste em combinar as decisões de diferentes técnicas de previsão individuais aplicadas a um mesmo conjunto de dados (KITTLER *et al*., 1998). Em *data mining*, esse método é bastante utilizado também na combinação de diferentes algoritmos de aprendizagem, como por exemplo, em Redes Neurais. De acordo com Haykin (2001), estudos apontam que modelos obtidos a partir da combinação de diferentes redes resultam em melhor desempenho quando comparados aos modelos obtidos com a rede que apresenta melhor desempenho individual.

São elencadas três etapas para o desenvolvimento de classificadores baseados na combinação em paralelo, sendo que, primeiramente são escolhidas as técnicas de classificação que geram os modelos individuais, posteriormente os modelos que obtiveram melhor desempenho são então combinados gerando por fim uma única previsão. Dado que o principal objetivo ao se combinar técnicas individuais é aumentar a acurácia da previsão, não há vantagem em combinar técnicas com características de previsão similares. Segundo Polikar (2006), existem diversas maneiras para realizar a combinação, entre elas, os métodos algébricos, como média aritmética ou ponderada, mediana, soma, produto, mínimo e máximo, e também os métodos baseados em votação, como por exemplo, a votação majoritária simples, na qual mais de duas técnicas são comparadas quanto a sua decisão. De acordo com Clemen (1989), métodos de combinação simples funcionam razoavelmente bem quando comparados a métodos mais complexos.

3. Método

A metodologia proposta para este estudo é uma adaptação dos métodos propostos por Berry e Linoff (2004) e por Selau (2012). É composta por quatro etapas, em que a primeira, chamada pré-processamento, consiste na definição da população alvo, seleção da amostra e na escolha e categorização das candidatas a variáveis explicativas. Na segunda etapa são desenvolvidos os modelos de *cross-selling* por meio das técnicas de classificação individuais: Regressão Logística e Redes Neurais. Na terceira etapa são aplicados os métodos de combinação em série e em paralelo. A quarta e última etapa tem por objetivo realizar um comparativo entre os modelos desenvolvidos nas duas etapas anteriores. Estas etapas são ilustradas na Figura 3 e são descritas na sequência.

Figura 3 - Método Proposto



3.1 Pré-processamento

A etapa de pré-processamento engloba desde a definição da população alvo e seleção dos dados, período de tempo e tamanho da amostra até a validação dos dados, agrupamento e seleção das variáveis que serão utilizadas no desenvolvimento dos modelos. Esta etapa é dividida em três passos: (*i*) Definição da população; (*ii*) Seleção da amostra; e (*iii*) Análise preliminar.

3.1.1 Definição da população

A escolha da população alvo e dos dados a serem utilizados devem contemplar os objetivos do estudo. Muitas vezes a seleção das variáveis relevantes ao problema de pesquisa e o volume e período de tempo a serem analisados é realizada de acordo com a disponibilidade dos dados. Primeiramente torna-se necessário avaliar o volume e a qualidade dos dados e definir quais variáveis podem descrever o cliente, como por exemplo, variáveis sócio-demográficas: idade, sexo, estado civil, renda, profissão, região de residência, entre outras. A escolha dessas variáveis pode ser baseada na literatura, no conhecimento de especialistas ou de acordo com as variáveis disponíveis no banco de dados da empresa. Considerando que o modelo de *cross-selling* proposto é baseado no comportamento de compra do cliente após adquirir o primeiro produto, ou seja, o produto pelo qual o cliente se direciona até o estabelecimento para efetuar a compra, todos os dados utilizados são oriundos das operações deste produto.

3.1.2 Seleção da amostra

A seleção da amostra tem impacto direto na qualidade do modelo final, tanto em relação a sua representatividade perante a população quanto à sua dimensão, visto que amostras maiores tendem a obter resultados melhores. Um ponto crucial a ser avaliado é a consistência e preenchimento dos dados, de modo que dados incorretos devem ser eliminados para que não sejam realizadas análises errôneas e inadequadas. É fundamental então que seja feita uma exploração prévia dos dados, avaliando todos os campos quanto ao seu conteúdo, qualidade de preenchimento, consistência e presença de observações faltantes (*missing*).

Com o intuito de se obter uma amostra mais representativa, o conjunto de dados é dividido de acordo com a natureza da variável de interesse, formando-se então dois grupos distintos: o grupo de indivíduos com resposta positiva (evento de interesse): clientes que adquiriram o segundo produto; e o grupo com resposta negativa: clientes que não adquiriram o segundo produto. Segundo Berry e Linoff (2004), em grande parte das aplicações de *data mining* é observado um desbalanceamento entre o primeiro e o segundo grupo, dado que o número de ocorrências do evento de interesse é consideravelmente menor. Uma alternativa à amostragem aleatória simples, que pode não ser eficiente quando se observa eventos raros (ou até mesmo nem tão raros), é

aumentar a proporção do evento menos frequentes de modo a se obter o equilíbrio na amostra e minimizar os problemas de discriminação. Para Berry e Linoff (2004), quando se tem uma variável resposta binária, valores entre 20% e 30% para a categoria mais rara na amostra produzem bons resultados, já Thomas (2002) sugere uma proporção 1:1.

É necessário ainda, que mais de uma amostra seja retirada da mesma população, pois ao se utilizar a mesma amostra para desenvolver e testar o modelo pode-se obter informações equivocadas quanto ao seu desempenho, como por exemplo, concluir que o modelo está bom quando, na verdade, ele se ajusta bem apenas para aquelas observações. O conjunto de dados é usualmente dividido em três amostras independentes: *(i*) análise (ou treinamento): é a amostra sobre a qual o modelo será construído; (*ii*) validação: é a amostra utilizada para medir a capacidade de generalização do modelo; e (*iii*) teste: é a amostra que servirá para avaliar o desempenho do modelo perante novos dados, considerando o desbalanceamento inicial entre os grupos. De acordo com Berry e Linoff (2004) a partição 60%-30%-10% (análise-validação-teste) tem bom desempenho em casos práticos.

3.1.3 Análise preliminar

A análise preliminar consiste em explorar os dados com o objetivo de adquirir algum conhecimento sobre o comportamento das variáveis e qualidade dos dados. Esta etapa contempla tarefas importantes como a escolha das variáveis para entrar na modelagem, o agrupamento dos atributos de variáveis e a criação de variáveis *dummies*. Para escolher quais variáveis irão seguir para o processo de modelagem, é analisada a associação entre cada variável explicativa e a variável resposta utilizando o risco relativo (RR), calculado a partir de tabelas de contingência. O cálculo do RR consiste em dividir o percentual de clientes do grupo 1 (adquiriram o segundo produto) pelo percentual de clientes do grupo 2 (não adquiriram o segundo produto) para cada atributo de cada variável explicativa. Quanto maior a diferença entre os percentuais dos que adquiriram o segundo produto e dos que não adquiriram, maior será a utilidade dessa variável no modelo.

Selau (2012) sugere a escala proposta por Lewis (1992) e Hand e Henley (1997) como método para agrupar os atributos de acordo com o risco que o cliente possui de adquirir o segundo produto após ter adquirido o primeiro, definida como: péssimo – RR < 0,50; muito mal – RR entre 0,50 e 0,67; mau – RR entre 0,67 e 0,90; neutro – RR entre 0,90 e 1,10; bom – RR entre 1,10 e 1,50; muito bom – RR entre 1,50 e 2,00; e excelente – RR maior que 2,00, de modo que esse risco seja homogêneo dentro de cada categoria da variável e heterogêneo entre elas. Após esse agrupamento são então criadas as variáveis *dummies*, que assumem os valores 1 (o cliente está neste grupo de risco) ou 0 (o cliente não está nesse grupo de risco).

Segundo (SICSÚ, 2010), além do risco relativo, há outros métodos que podem ser utilizados para categorização de variáveis, como, por exemplo, o valor da informação e o peso da evidência. O valor da informação (IV) utiliza-se da soma dos pesos das evidências (WOE), que é obtido pelo logaritmo natural da probabilidade do evento de interesse em uma dada categoria, pela probabilidade de não ocorrência do evento na categoria (Equação 5), ponderados após o agrupamento das categorias, conforme Equação 6.

$$WOE = \ln\left[\frac{P(c \mid Bom)}{P(c \mid Mau)}\right] \tag{5}$$

$$IV = \sum [P(c \mid Bom) - P(c \mid Mau)]\ x\ WOE \tag{6}$$

3.2 Técnicas de Modelagem Individuais

Para construção dos modelos individuais, é necessário primeiramente se avaliar as suposições das técnicas de Regressão Logística e Redes Neurais, escolher as variáveis explicativas e, por fim, avaliar o seu ajuste aos dados. Como resultado, é obtido um score para cada cliente que classifica a sua propensão a adquirir o segundo produto após adquirir o primeiro.

3.2.1 Regressão Logística

A técnica de Regressão Logística possui o pressuposto de ausência de multicolinearidade. Uma forma de contornar este pressuposto é utilizar o método *stepwise*, que seleciona automaticamente a combinação de variáveis que melhor explicam o modelo, de forma que variáveis altamente correlacionadas tendem a não ser selecionadas conjuntamente. Esse método é geralmente utilizado porque, na presença de duas ou mais variáveis explicativas altamente correlacionadas, o método seleciona dentre elas a que possui maior influência na variável resposta e as demais não entram no modelo.

3.2.2 Redes Neurais

A técnica de Redes Neurais não possui pressupostos a serem analisados, pois o modelo é gerado por meio do processo de aprendizagem. Pelo mesmo motivo, não existe um número de nós, de camadas escondidas ou um algoritmo fixo para que sejam aplicados a qualquer conjunto de dados. Desta maneira, é necessário realizar diversos arranjos possíveis destas características e das variáveis em estudo para avaliar qual o modelo que obtém melhor desempenho.

3.3 Métodos de Combinação

Para aplicar os métodos de combinação em série e em paralelo serão utilizados os modelos a serem construídos conforme descrito na Etapa 2. Ao final, são avaliados os ajustes destes métodos ao conjunto de dados.

3.3.1 Combinação em série (*Hybrid*)

No método de combinação em série serão informados ao nó de entrada da Rede Neural apenas as variáveis explicativas que foram significativas no modelo obtido com a técnica de Regressão Logística. Esta é a única diferença no processo em relação ao modelo de Rede Neural tradicional, uma vez que, após a entrada dos dados a modelagem segue da mesma forma.

3.3.2 Combinação em paralelo (*Ensemble*)

No método de combinação em paralelo, é combinado, para cada cliente, o score que determina a sua propensão a adquirir o segundo produto, obtido pelo modelo de Regressão Logística e pelo modelo de Rede Neural, originando um novo *score* para o cliente. Essa combinação é realizada por meio da média aritmética. Outras formas de combinação estão disponíveis, mas não serão abordadas neste estudo.

3.4 Qualidade do ajuste e comparativo

Diversas medidas estatísticas são utilizadas para testar o desempenho e medir a qualidade dos modelos com o objetivo de comparar os métodos empregados e escolher o melhor entre eles. As medidas abordadas neste trabalhão são: (*i*) percentual de classificação corretas; (*ii*) índice Kolmogorov-Smirnov (KS); e a (*iii*) curva ROC.

O percentual de acerto do modelo nas classificações dos clientes propensos ou não a adquirir o segundo produto é avaliado pela divisão da quantidade de clientes corretamente classificados pelo total de clientes da amostra.

O teste não paramétrico de Kolmogorov-Smirnov (KS) para 2 amostras independentes tem por objetivo determinar se as duas amostras provém da mesma população. Consiste basicamente em obter a máxima diferença entre as distribuições acumuladas das duas amostras. De acordo com Picinini *et. al*. (2003) uma diferença entre as distribuições acumuladas maiores que 30% e taxas de acerto superior a 65% são consideradas satisfatórias na classificação dos dois grupos.

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica da sensibilidade versus especificidade e tem por objetivo primário mensurar a capacidade do modelo em reduzir os erros tipo I e tipo II. Para este estudo, entende-se por sensibilidade a probabilidade de identificar os clientes propensos a adquirir o segundo produto, já especificidade é a probabilidade de identificar os clientes não propensos, sendo que o valor de corte para estes valores é a probabilidade a partir da qual o especialista considera o cliente propenso ou não. No entanto, a medida que é analisada de fato é a área sob a Curva ROC (AUROC – Area Under ROC), também conhecida por discriminação. Uma discriminação perfeita vale 1, e uma discriminação de 0,5 significa que o modelo não agrega valor, pois estimular os clientes aleatoriamente teria o mesmo resultado.

4. Resultados

Os resultados da aplicação do método proposto em uma base de dados real são expostos nesta seção e seguem a mesma ordenação da metodologia descrita na seção 3. O *software* utilizado para as análises é o *SAS Enterprise Miner* versão 4.3.

4.1 Pré-processamento

4.1.1 Definição da população

A população de interesse são os clientes de uma instituição financeira brasileira que, após contratarem o produto crédito pessoal (CP), podem contratar também o produto seguro de vida. Com o intuito de classificar os clientes da instituição quanto à sua propensão a adquirir o segundo produto, a base de dados contempla todos os clientes que adquiriram o crédito pessoal entre 01 de julho de 2011 e 31 de julho de 2012, totalizando 61.365 clientes. A base é composta por todas as variáveis dos clientes disponíveis na base de dados da empresa (preenchidas na proposta em que é feita a solicitação de crédito pessoal) e também informações referentes à venda do primeiro produto; estas variáveis estão descritas na Tabela 1. A variável resposta é dicotômica e define se o cliente adquiriu ou não o seguro de vida depois de ter contratado o crédito pessoal.

Tabela 1: Variáveis disponíveis para o modelo

Variável	Descrição
Sexo	Feminino ou masculino
Idade	Idade, em anos, no dia da contratação do CP
Estado Civil	Casado, solteiro, divorciado, viúvo e outros
Renda	Valor da renda (R$)
Órgão de trabalho	Órgão em que o cliente trabalha
Grupo do órgão	Critérios estabelecidos pela instituição
CEP Residencial	CEP do local de residência do cliente
Banco	Banco que o cliente é correntista
Origem da venda	Venda interna ou externa
Pessoa captação	Vendedor pessoa física ou jurídica
Valor Liberado CP	Valor liberado no CP (R$)

4.1.2 Seleção da amostra

Dados inconsistentes, incorretos ou faltantes representaram menos de 0,01% da base total e, portanto, foram excluídos da análise. Os clientes foram divididos aleatoriamente em amostras de análise, validação e teste, na proporção de 60%, 30% e 10%, respectivamente. Considerando o desbalanceamento inicial entre os grupos, em que 30% dos clientes adquiriram o seguro de vida (evento de interesse) e 70% não adquiriram, optou-se então, com o intuito de se obter o equilíbrio na amostra e minimizar os problemas de discriminação, dividir os clientes aleatoriamente na proporção 1:1, tanto na amostra de análise como na amostra de validação. Para a amostra de teste, a verdadeira proporção foi mantida. A amostra de análise é então constituída por 27.352 clientes, a amostra de validação por 6.838 e a amostra de teste por 6.136 clientes, resultado em uma proporção de 68%, 17% e 15%, respectivamente.

4.1.3 Análise Preliminar

As 11 variáveis selecionadas na seção 4.1.1 como candidatas a compor o modelo foram analisadas individualmente e a variável "estado civil" foi excluída da análise. Isso porque, conclui-se que esta informação não é preenchida corretamente no momento da proposta. No caso da Regressão Logística em particular, faz-se necessário a criação de variáveis *dummies* para as variáveis qualitativas, para que estas sejam melhor interpretáveis na equação resultante. As variáveis *dummies* foram criadas para as 10 variáveis candidatas restantes, sendo que, para aquelas variáveis que apresentam um grande número de atributos, as *dummies* foram criadas com base nas categorias de risco relativo apresentadas na seção 3.1.3, de modo que, para cada variável nominal (órgão de trabalho, grupo de órgão, CEP e banco) se obteve sete agrupamentos de atributos de acordo com o risco de adquirir o segundo produto. Também com base no risco relativo, foram criadas classes para as variáveis de natureza numérica (idade, renda e valor liberado) de modo a se evitar problemas decorrentes da não linearidade. No total, foram geradas 58 variáveis *dummies* que serão as variáveis explicativas para a construção do modelo desenvolvido com a Regressão Logística.

4.2 Técnicas de Modelagem Individuais

4.2.1 Regressão Logística

Para a construção do modelo logístico utilizou-se o método stepwise, com níveis de significância para a entrada e saída de variáveis de 5%. Com o uso deste método garantiu-se o atendimento da suposição de ausência de multicolinearidade entre as variáveis explicativas. Para que algumas variáveis tivessem significância foi necessário agrupar *dummies* com risco próximo, como por exemplo, agrupar uma com risco péssimo e outra com risco muito mau. Ao final, 19 variáveis *dummies* foram significativas para compor o modelo final. A variável resposta Y define se o cliente possui propensão a adquirir um seguro de vida após a aquisição do crédito pessoal (Y = 1) ou não (Y = 0). A equação que retorna a probabilidade de um cliente vir a comprar o seguro de vida (segundo produto) após a compra do produto crédito pessoal (primeiro produto) é apresentada na Equação 7. A especificação das variáveis é apresentada no Quadro 1.

$$P(Y=1) = \frac{1}{1 + \exp(-1{,}1586 - 0{,}5632\,DGCEPR12 + 0{,}2610\,DGCEPR56 + 0{,}4516\,DGCEPRE2 + 0{,}1055\,DGCEPRE4 + 0{,}6773\,DGCEPRE7 + 0{,}1590\,DGGORGAO6 - 0{,}0531\,DGORGAO2 + 0{,}1610\,DGORGAO7 - 0{,}2996\,DIDAD1 + 0{,}0893\,DIDAD10 - 0{,}1178\,DIDAD3 + 0{,}0442\,DIDAD7 + 0{,}0538\,DIDAD8 + 0{,}1136\,DPSCAPTF + 0{,}0322\,DSEXOF - 0{,}7072\,DVENDE + 0{,}0964\,DVLIB4 + 0{,}0725\,DVLIB7 + 0{,}0932\,DVLIB8)} \quad (7)$$

Para a interpretação do modelo, é necessário se observar o sinal dos coeficientes de cada uma das variáveis; coeficientes positivos indicam que clientes com aquela característica têm maior probabilidade de adquirir o segundo produto, enquanto coeficientes negativos representam diminuição na probabilidade de adquirir o segundo produto. Observando a Equação 7, pode-se concluir, por exemplo, que o cliente que possui residência no grupo de CEP de risco excelente (DGCEPRE7) ou que trabalhe em um dos órgãos do grupo de risco excelente (DGORGAO7) tem um aumento na probabilidade de adquirir o segundo produto. Já um cliente que contratou o primeiro produto com a venda externa (DVENDE) ou que trabalha em um dos órgãos do grupo de risco muito mau (DGORGAO2) tem uma diminuição na probabilidade de adquirir o segundo produto.

Quadro 1: Especificação das variáveis utilizadas no modelo de Regressão Logística.

Y = propensão de o cliente vir a adquirir um seguro de vida após a aquisição do crédito pessoal	
DGCEPRE2	CEP residencial com desempenho muito mau
DGCEPRE4	CEP residencial com desempenho bom
DGCEPRE7	CEP residencial com desempenho excelente
DGCEPR12	CEP residencial com desempenho péssimo ou muito mau
DGCEPR56	CEP residencial com desempenho bom ou muito bom
DGORGAO2	Órgão com desempenho muito mau
DGORGAO7	Órgão com desempenho excelente
DGGORGAO6	Grupo de órgão com desempenho muito bom
DIDAD1	Idade até 21 anos
DIDAD3	Idade entre 27 e 30 anos
DIDAD7	Idade entre 44 e 50 anos
DIDAD8	Idade entre 51 e 60 anos
DIDAD10	Idade entre 66 e 70 anos
DPSCAPTF	Pessoa captação é do tipo física
DSEXOF	É do sexo feminino
DVENDE	Origem de venda é externa
DVLIB4	Valor liberado no crédito pessoal entre R$ 1.100 e R$ 1.400
DVLIB7	Valor liberado no crédito pessoal entre R$ 2.000 e R$ 3.000
DVLIB8	Valor liberado no crédito pessoal superior a R$ 3.000.

4.2.2 Redes Neurais

O modelo neural foi construído por meio do treinamento supervisionado utilizando o algoritmo *backpropagation* e a função de ativação Logística. Foram desenvolvidos diversos modelos variando a quantidade de neurônios (de 1 a 58 neurônios) na camada escondida. A partir dos modelos neurais desenvolvidos, se optou pelo modelo com 31 neurônios na camada escondida, pois este possui o maior percentual de classificações corretas, maior valor de KS a de área abaixo da curva ROC. A Tabela 2 mostra os resultados das três melhores redes construídas.

Tabela 2: Melhores redes construídas para o modelo neural – amostra de teste

N de neurônios	% Classificação	KS	ROC
18	72,39%	0,4850	0,8000
31	72,80%	0,4860	0,8000
32	71,91%	0,4840	0,7960

4.3 Métodos de Combinação

4.3.1 Combinação em série (*Hybrid*)

O modelo de combinação em série foi construído utilizando o resultado da Regressão Logística como entrada na Rede Neural. As 19 variáveis *dummies* que foram significativas para compor o modelo logístico descrito na seção 4.2.1 foram informadas ao modelo neural como variáveis explicativas. As demais variáveis não são informadas com o intuito de minimizar o tempo de aprendizado da rede e facilitar a interpretação dos resultados. Analogamente ao processo realizado para a construção do modelo neural na seção 4.2.2, o modelo de combinação em série foi construído por meio do treinamento supervisionado utilizando o algoritmo *backpropagation* e a função de ativação Logística. A partir dos modelos neurais desenvolvidos (de 1 a 58 neurônios na camada escondida), se optou pelo modelo com 30 neurônios na camada escondida, por possuir o maior percentual de classificações corretas, maior valor de KS e de área abaixo da curva ROC, dentre as redes testadas. A Tabela 3 mostra os resultados dos três melhores modelos construídos.

Tabela 3. Melhores redes construídas para a combinação em série – amostra de teste

N de neurônios	% Classificação	KS	ROC
29	72,83%	0,4870	0,8030
30	72,85%	0,4870	0,8030
33	72,67%	0,4850	0,8000

4.3.2 Combinação em paralelo (*Ensemble*)

A combinação em paralelo foi construída utilizando o *score* obtido com o modelo logístico construído em 4.2.1 e o *score* obtido com o modelo neural construído em 4.2.2. Primeiramente são obtidos os *scores* por meio das técnicas de modelagem individuais para cada cliente e então estes valores são combinados por meio de média aritmética. Assim, um cliente que possui, por exemplo, um *score* de 0,80 via Regressão Logística e 0,86 via Redes Neurais, utilizando a combinação em paralelo esse cliente passa a ter um *score* de 0,83.

4.4 Qualidade do ajuste e comparativo

A avaliação de cada um dos modelos construídos se dá mediante o emprego das medidas já expostas na seção 3.4; são elas: percentual de classificação corretas, teste de Kolmogorov-Smirnov (KS) e curva ROC. Essas medidas são avaliadas nas três amostras (análise, validação e teste) e são expostas na Tabela 4, com o intuito de realizar um comparativo entre as técnicas e métodos utilizados no desenvolvimento do modelo de *cross-selling*.

Tabela 4: Medidas de desempenho para as três amostras

Amostra	Medida de Desempenho	RL	RN	Em série	Em paralelo
Análise	% Classificações corretas	75,61%	75,08%	75,07%	75,03%
	KS	0,4860	0,5033	0,5029	0,5023
	ROC	0,8030	0,8091	0,8116	0,8078
Validação	% Classificações corretas	74,14%	74,82%	74,92%	74,80%
	KS	0,4829	0,4975	0,4993	0,4970
	ROC	0,8029	0,8069	0,8086	0,8090
Teste	% Classificações corretas	73,26%	72,80%	72,85%	72,93%
	KS	0,4790	0,4861	0,4873	0,4882
	ROC	0,7965	0,7999	0,8033	0,8054

A primeira das medidas apresentada na Tabela 4 mostra o percentual de classificações corretas para cada uma das amostras, sendo que a técnica ou método que melhor classifica os clientes na amostra de teste é o modelo logístico, com 73,26% de classificações corretas. Tanto na amostra de análise quanto na amostra de validação e de teste, os percentuais de acerto total encontrados para todas as técnicas e métodos são superiores a 65%, indicando que todos possuem boa capacidade de classificação.

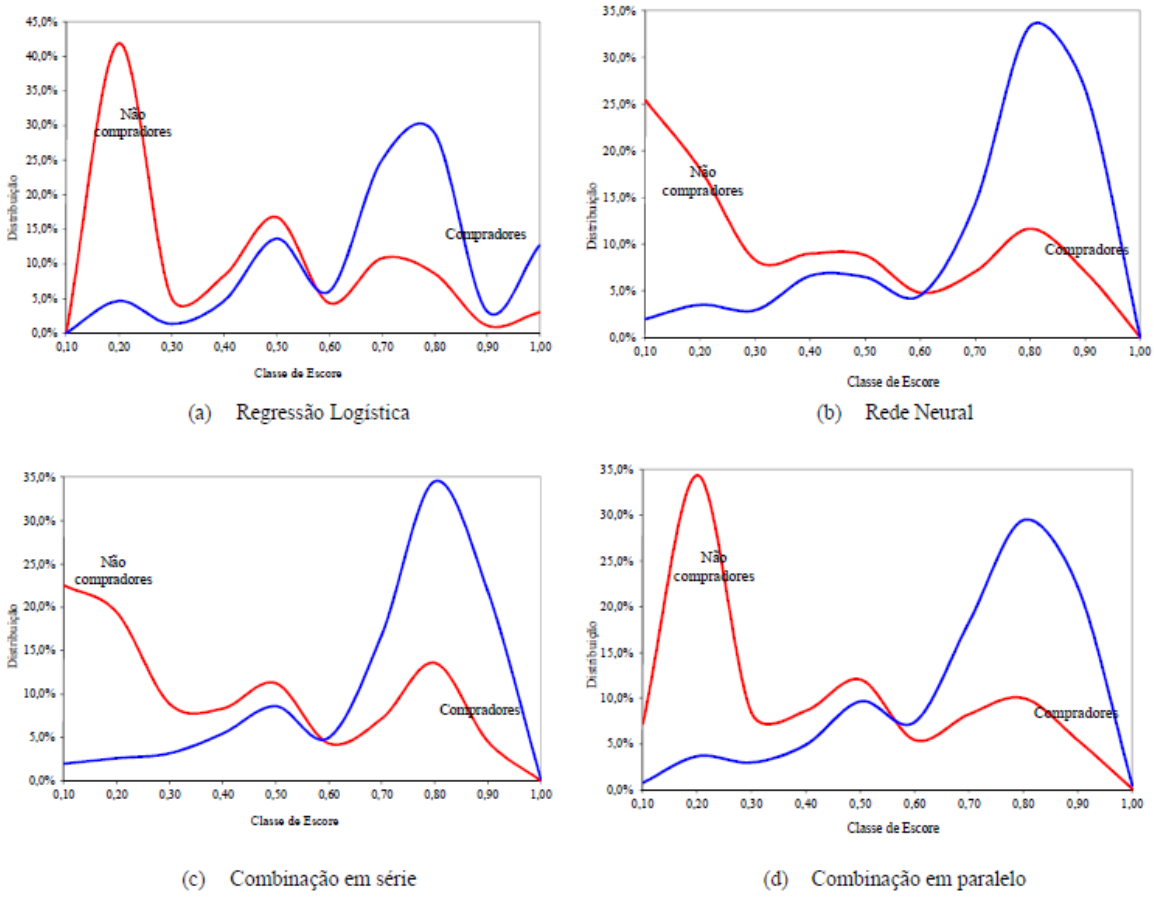
A segunda medida observada é o valor do teste de KS, que tem por objetivo determinar se duas amostras provêm de uma mesma população. Espera-se que as duas amostras (clientes propensos e clientes não propensos a adquirir o segundo produto) sejam oriundas de populações distintas, pois assim o modelo é eficiente em separar os dois grupos de clientes. Os quatro modelos construídos apresentaram diferença entre as distribuições acumuladas dos clientes propensos e dos não propensos maior de 0,30 de modo que estes podem ser considerados eficientes. O maior valor de KS para a amostra de teste é o obtido com o método de combinação em paralelo, o que significa dizer que

este método alcançou os maiores níveis de diferença entre as distribuições acumuladas dos clientes propensos e dos não propensos.

A terceira e última medida a ser analisada é a área abaixo da curva ROC, que mostra uma boa capacidade de discriminação para todos os modelos, indicando que a capacidade de identificar corretamente os propensos (sensibilidade), assim como a capacidade de identificar os não propensos (especificidade) está bem ajustada, sendo que o método de combinação em paralelo apresentou um valor ligeiramente maior.

Uma avaliação dos modelos construídos, na amostra de teste, é apresentada na Figura 4 com a distribuição dos dois grupos de clientes: compradores e não compradores do segundo produto. Analisando o comportamento das curvas de distribuição dos clientes compradores e não compradores, verifica-se que os quatro modelos conseguem separar os dois grupos de clientes, já que é possível observar a tendência de que os clientes não compradores se concentram à esquerda da escala e os compradores à direita.

Figura 4 - Distribuição de compradores e não compradores por classe de score



A Tabela 5 apresenta uma comparação entre os modelos construídos e a escolha, dentre estes, daquele que melhor se ajustou aos dados, segundo os resultados das medidas apresentadas para verificar a qualidade de ajuste. Foi escolhido como melhor modelo aquele que obteve desempenho superior na amostra de teste em pelo menos duas das medidas de ajuste apresentadas na tabela 4.

Tabela 5: Comparativo dos modelos construídos

Técnica/Método	Melhor ajuste
Regressão Logística *vs* Rede Neural	Rede Neural
Regressão Logística *vs* Em série	Em série
Rede Neural *vs* Em série	Em série
Regressão Logística *vs* Em paralelo	Em paralelo
Rede Neural *vs* Em paralelo	Em paralelo
Em série *vs* Em paralelo	Em paralelo

Ao analisar as três medidas empregadas para comparação das técnicas e métodos abordados, observa-se, na comparação entre as técnicas de modelagens individuais, que os resultados obtidos com o modelo neural foram ligeiramente superiores aos obtidos com a Regressão Logística. Ao analisar os dois métodos de combinação percebe-se que o método de combinação em paralelo obteve resultados ligeiramente superiores em relação ao método de combinação em série. As comparações entre as técnicas individuais e os métodos de combinação indicam que os dois métodos de combinação são ligeiramente superiores às duas técnicas de modelagem individuais. Cabe ressaltar ainda, que os métodos de combinação possuem um objetivo em comum, que é o de melhorar a acuracidade da previsão, porém, o método de combinação em série possui a vantagem da facilidade na interpretação das variáveis obtidas através da Regressão Logística, o que não acontece no método de combinação em paralelo.

5. Considerações Finais

Este artigo desenvolveu um método para resolução de um problema real de *cross-selling,* buscando identificar que clientes, após terem adquirido um primeiro produto (crédito pessoal), têm maior probabilidade de adquirir também um segundo produto (seguro de vida). Para tanto, foram apresentadas e comparadas duas das técnicas mais utilizadas em *data mining* para previsão e classificação de clientes, Regressão Logística e Redes Neurais, e dois métodos de combinação dessas técnicas, combinação em série e em paralelo, com o intuito de se chegar ao modelo que melhor classifica os clientes quanto à sua propensão a adquirir um segundo produto. Foram apresentados todos os passos necessários para a obtenção dos modelos utilizando as técnicas e métodos abordados, detalhando desde a obtenção das variáveis e das amostras, a modelagem e aplicação das técnicas e também dos métodos de combinação para se chegar ao melhor modelo. Esse trabalho traz uma abordagem bastante inovadora principalmente ao que se refere aos modelos de *cross-selling* e às comparações das técnicas individuais de modelagem com os dois métodos de combinação; técnicas e métodos estes que podem ser utilizados nas mais diversas aplicações de *data mining*.

Os dois objetivos principais foram alcançados: resolver um problema real de *cross-selling* de uma instituição financeira e contribuir na pesquisa de técnicas de *data mining*. A efetividade desta proposta para a empresa se dá pelo fato de que a utilização do modelo de *cross-selling* elimina a subjetividade da análise tradicional, aproveitando as informações expostas para direcionar as campanhas de marketing. Uma vez que foi possível identificar o perfil dos clientes mais propensos a adquirir o produto, atribuindo a estes clientes um *score* que determina a sua propensão a adquiri-lo, esta análise se faz de grande utilidade também para possíveis aprimoramentos do produto. Além disso, a padronização do procedimento de decisão e o direcionamento correto das campanhas diminuem os gastos e aumentam o retorno das mesmas, aumentando a rentabilidade da empresa e possibilitando uma maior eficiência no atendimento aos clientes.

Os métodos de combinação apresentaram desempenho ligeiramente superior às técnicas individuais, sendo que o método de combinação em paralelo obteve maior ajuste aos dados em relação ao método de combinação em série, e, consequentemente, aos modelos individuais. Diferenças entre os modelos logístico e neural e os métodos de combinação em série e em paralelo empregados parecem pouco relevantes, mas servem de indícios para comprovar a melhora dos modelos combinados sobre os modelos individuais. Nesse sentido, uma possibilidade de trabalho futuro que se coloca é a comparação entre tais modelos em outras bases de dados, com o intuito de reforçar os resultados apresentados.

Além disso, cabe ressaltar ainda que, algumas operações possuem um nível de criticidade alto para a empresa e, mesmo diferenças singelas de acuracidade das previsões podem resultar em ganhos ou perdas significantes para a empresa. A escolha do melhor modelo a ser utilizado deve ir ao encontro das necessidades da empresa, dado que o melhor resultado encontrado em termos de ajuste se deu com a aplicação do método de combinação em paralelo, porém, este não possui grandes diferenças quando comparado ao método de combinação em série, método este que possui a vantagem da facilidade na interpretação das variáveis obtidas através da Regressão Logística, o que não acontece no método de combinação em paralelo.

No decorrer do desenvolvimento deste artigo surgiram algumas questões que não foram abordadas, mas que foram consideradas importantes como sugestões para trabalhos futuros: (*i)* fazer uso de algoritmos genéticos para encontrar os parâmetros ótimos das Redes Neurais (quantidade ótima de neurônios da camada oculta e quantidade de camadas ocultas), pois o método utilizado de tentativa e erro é limitado, no sentido que os parâmetros encontrados podem não ser os melhores, resultando em um poder de classificação menor do que poderia se obter ao otimizar a rede; (*ii*) utilizar outras formas de combinar os *scores* no método de combinação em paralelo, como por exemplo, a média ponderada, na qual se atribui um peso maior para aquele modelo que obteve um melhor desempenho individual; (*iii*) combinar redes com parâmetros diferentes, como por exemplo, diferentes algoritmos de aprendizagem, número de neurônios e quantidade de camadas ocultas; e (*iv*) investigar o quanto representa de ganho efetivo para a empresa usar os modelos criados pelos métodos de combinação comparados com os modelos criados utilizando as técnicas individuais.

Referências bibliográficas

ADORNO, C. F.; BUENO, J. F. Modelos de Propensão: Oferta de Crédito Pessoal, 2011.

BATES, J. M.; GRANGER, C. W. J. The combination of forecasts. Operational Research Quarterly. v. 20, n. 4. p. 451-468, 1969.

BERRY, M. J. A; LINOFF, G. S. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. 2.ed. New York: John Wiley & Sons, 2004.

CHEN, W.; MA, C.; MA, L. Mining the customer credit using hybrid support vector machine technique. Expert Systems with Applications, v.36, n.4, p.7611-7616, 2009.

CLEMEN, R. T.; Combining forecasts: A review and annotated bibliography. International Journal of Forecasting. v. 5, p. 559-583, 1989.

COSTANTINE, C.; PAPPALARDO, C. A Hierarchical procedure for combination of forecasts. International Journal of Forecasting. v. 26, p. 725-743, 2010.

DINIZ, C.; LOUZADA, F. Modelagem Estatística para Risco de Crédito. In: Minicurso no XX SINAPE – Simpósio Nacional de Probabilidade e Estatística, João Pessoa-PB, 2012.

DYCHE, J. The CRM handbook: a business guide to customer relationship management. Reading, MA: Addison-Wesley, 2001.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine. v. 17 n. 3, 1996.

GHODSELAHI, A. A Hybrid support vector machine ensemble model for credit scoring. International Journal of Computer Applications, v.17, n.5, 2011.

MORAES, L. G. Uma abordagem alternativa de behavioral scoring usando modelagem híbrida de dois estágios com Regressão Logística e Redes Neurais. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Análise multivariada de dados. 5.ed. Porto Alegre: Bookman, 2005.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: concepts and techniques. 3.ed. San Francisco: Morgan Kaufmann, 2012.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. Journal of Royal Statistical Society, Series A, v.160, n.3, p.523-541, 1997.

HAYKIN, S. Redes Neurais: princípios e prática. Trad. Paulo Martins Engel. 2.ed. Porto Alegre: Bookman, 2001.

HIBON, M.; EVGENIOU, T. To combine or not to combine: selecting among forecasts and their combinations. International Journal of Forecasting. v. 21, p. 15-24, 2005.

HOSMER, D. W.; LEMESHOW, S. Applied logistic regression. New York: John Wiley & Sons, 1989.

HSIEH, N. C. Hybrid mining approach in the design of credit scoring models. Expert Systems with Applications, v.28, p. 655-665, 2005.

HSIEH, N. C.; HUNG, L. P. A data driven ensemble classifier for credit scoring analysis. Expert Systems with Applications, v.37, n.1, p. 534-545, 2010.

KAMAKURA, W. A.; RAMASWAMI, S.; SRIVASTAVA, R. Applying Latent Trait Analysis in the Evaluation of Prospects for Cross-Selling of Financial Services. International Journal of Research in Marketing, v. 8, p. 329-349, 1991.

KISAHLEITNER, M. Análise de Técnicas de Data Mining na aquisição de clientes de cartão de crédito não correntistas. 93f. Dissertação (Mestrado em Administração) – Fundação Getúlio Vargas, São Paulo, 2008.

KITTLER, J.; HATEF, M.; DUIN, R. P. W.; MATAS, J. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, v.20, n.3, p.226-239, 1998.

KNOTT, A.; HAYES, A.; NESLIN, S. A. Next-product-to-buy models for cross-selling applications. Journal of Interactive Marketing, v. 16, 2002.

LEE, T.; CHIU, C.; LU, C.; CHEN, I. Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications, v.23, n.3, p.245-254, 2002.

LEE, T.S.; CHEN, I. F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, v.28, p.743-752, 2005.

LEWIS, E. M. An introduction to credit scoring. San Rafael: Fair, Isaac and Co., Inc. 1992.

MAKRIDAKIS, S. G.; HIBON, M. The M3-Competition: results, conclusions and implications. International Journal of Forecasting, v. 16, p. 451-476, 2000.

MARTINS, V. L.M Comparação de Combinação de Previsões correlacionadas e não correlacionadas com as suas previsões individuais: um estudo com séries industriais. 100f. Dissertação (Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011.

NISBET, R.; ELDER, J.; MINER, G. Handbook of Statistical Analysis and Data Mining Applications.Amsterdam, NL: Academic Press, 2009.

OLSON, D. L.; DELEN, D. Advanced Data Mining Technique. New York: Springer, 2008.

PICININI, R.; OLIVEIRA, G. M. B.; MONTEIRO, L. H. A. Mineração de critério de credit scoring utilizando algoritmos genéticos. In: VI Simpósio Brasileiro de Automação Inteligente, Bauru, SP, 2003.

POLIKAR, R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, v.6, n.3, p.21-45, 2006.

SELAU, L. P. R. Modelagem para Concessão de Crédito a pessoas físicas em empresas comerciais: da decisão binária para a decisão monetária. 111f. Tese (Doutorado em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

SICSÚ, A. L. Desenvolvimento, implantação, acompanhamento. São Paulo: Blucher, 2010.

SILVA, V. F. Modelos de Propensão ao Consumo baseados em Redes Neuronais Artificiais, o caso particular do Crédito Pessoal. 105f. Dissertação de Mestrado (Estatística e Gestão da Informação) – Universidade Nova de Lisboa, Lisboa, 2009.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. Credit Scoring and Its Applications, Philadelphia: SIAM.

TSAI, C. F.; CHEN, M. L. Credit rating by hybrid machine learning techniques. Applied Soft Computing, v.10, p.374-380, 2010.

WERNER, L. Um Modelo Composto para Realizar Previsão de Demanda Através da Integração da Combinação e de Previsões e Ajuste Baseado na Opinião. 166f. Tese de Doutorado (Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

Agradecimento:

Ao SAS Institute, por disponibilizar a licença do software para que este trabalho pudesse ser realizado.

* Este artigo faz parte do trabalho de conclusão do curso de Bacharelado em Estatística pela Universidade Federal do Rio Grande do Sul.

Abstract

As a result of the increasing technological development, more powerful computers make it possible to store large amount of data everyday. The data mining techniques emerge as a smart alternative and effective way to turn them into knowledge. This work proposes to solve a real cross-selling problem of a financial institution, and aiming to contribute to the development of data mining techniques, we carried out a comparison between two consecrated techniques, logistic regression and neural networks, and between two ways of combining them, in series (hybrid), where logistic regression is used to select the variables that will enter the neural network, and in parallel (ensemble), where the results of the individual techniques are combined based on their decisions. The comparisons between the performance of individual techniques and combination methods show that the combination methods are slightly better than individual modeling techniques.

Keywords: Cross-Selling, Logistic Regression, Neural Networks, Hybrid, Ensemble.

Planejamento de amostras domiciliares no Brasil explorando a malha setorial do Censo Demográfico 2010

Sâmela Batista Arantes [1]
Pedro Luis do Nascimento Silva [2]

Resumo

Os desenhos amostrais complexos estão presentes nas pesquisas domiciliares amostrais realizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Uma das fases do planejamento amostral dessas pesquisas é o cálculo do tamanho de amostra. Este trabalho apresenta uma proposta de cálculo de tamanhos amostrais utilizando o Efeito do Plano Amostral (EPA) para quatro estratégias amostrais. Todas as quatro estratégias amostrais são conglomeradas em dois estágios, onde em um primeiro estágio são selecionados os setores censitários e no segundo os domicílios.Os resultados mostram que os planos amostrais PPT são, em geral, mais eficientes, pois tiram proveito da correlação entre os totais das variáveis e o tamanho do setor. Além disso, planos amostrais sem reposição apresentam uma certa vantagem em relação ao plano amostral com reposição para sorteio dos setores.

Palavras-chave: Pesquisas amostrais domiciliares, plano amostral complexo, efeito do plano amostral, coeficiente de correlação intraclasse, tamanhos amostrais.

[1] Escola Nacional de Ciências Estatísticas. E-mail: samela.arantes@ibge.gov.br
[2] Escola Nacional de Ciências Estatísticas. E-mail: pedronsilva@gmail.com

1. Introdução

A principal motivação deste trabalho é orientar o cálculo de tamanhos de amostra para pesquisas amostrais domiciliares como as que são realizadas pelo Instituto Brasileiro de Geografia e Estatística. Para o planejamento destas pesquisas domiciliares o IBGE utiliza o cadastro do Censo Demográfico mais recente, que fornece parâmetros populacionais e a malha setorial atualizada.

Na maior parte das pesquisas domiciliares os planos amostrais são conglomerados em dois estágios, onde em um primeiro estágio são selecionados setores censitários e no segundo estágio são selecionados domicílios. Em consequência, o dimensionamento amostral das pesquisas domiciliares envolve definir tamanhos amostrais de setores e domicílios. Estes tamanhos são obtidos em função de precisão e custos, além de considerar as questões operacionais.

O Efeito do Plano Amostral (EPA) é uma ferramenta útil para auxiliar no cálculo de tamanhos de amostra. EPAs são obtidos para alguns planos amostrais conglomerados em função do coeficiente de correlação intraclasse, que é uma medida da homogeneidade dentro dos conglomerados.

A seção 2 apresenta o referencial teórico de alguns planos amostrais conglomerados em dois estágios, onde são apresentadas fórmulas gerais para calcular variâncias de estimadores de totais sob esses planos. A seção 3 apresenta algumas características das principais pesquisas domiciliares realizadas pelo IBGE descrevendo os planos amostrais de cada uma. A seção 4 apresenta os conceitos de Efeito do Plano Amostral e coeficiente de correlação intraclasse e a relação dessas medidas com o cálculo de tamanho de amostra. A seção 5 descreve o Censo Demográfico realizado pelo IBGE. A seção 6 apresenta estimativas de Efeitos do Plano Amostral obtidas a partir do Censo Demográfico 2010 com o auxílio de alguns coeficientes de correlação intraclasse. A seção 7 apresenta a abordagem proposta para cálculo de tamanhos amostrais em pesquisas domiciliares como as que realiza o IBGE.

A seção 8 apresenta as considerações finais.

2. Referencial teórico dos planos amostrais

2.1 Amostragem Conglomerada em Pesquisas Domiciliares

A amostragem por conglomerados é bastante útil quando o objetivo é otimizar os recursos para se chegar à unidade elementar final em pesquisas amostrais domicilares. Neste tipo de amostragem as unidades elementares estão reunidas em grupos que podem conter outros subgrupos. Na amostragem por conglomerados em 2 estágios é feita a seleção de grupos e dentro dos grupos selecionados é feita a seleção de unidades elementares.

Os planos amostrais conglomerados possuem algumas vantagens:

- Podem ser usados quando não existe cadastro das unidades elementares mas existe cadastro dos conglomerados;
- Como a coleta é concentrada em apenas alguns conglomerados o custo de deslocamento para coleta das unidades elementares é menor;
- A supervisão da coleta das unidades elementares fica facilitada e há um aumento do controle sobre a qualidade dos dados.

A principal desvantagem é o aumento da variabilidade dos estimadores em comparação com planos amostrais que não envolvam conglomeração. O Efeito do Plano Amostral é uma ferramenta útil para comparar a precisão de estimadores com a de planos amostrais por amostragem aleatória simples (AAS).

Quatro estratégias amostrais baseadas em amostragem conglomerada em 2 estágios foram consideradas neste trabalho:

- Estratégia 1 (E1): emprega seleção de setores (1º estágio) e domicílios (2º estágio) por Amostragem Aleatória Simples (AAS) combinada com o estimador de Horvitz-Thompson;
- Estratégia 2 (E2): emprega seleção de setores (1º estágio) e domicílios (2º estágio) por Amostragem Aleatória Simples (AAS) combinada com o estimador de razão;

- Estratégia 3 (E3): faz a seleção dos setores (1º estágio) com probabilidade proporcional ao tamanho (PPT) com reposição e a seleção de domicílios (2º estágio) por AAS, combinada com estimador natural;
- Estratégia 4 (E4): compreende a seleção de setores (1º estágio) com PPT de Pareto e a seleção de domicílios por AAS (2º estágio), combinada com o estimador do tipo Horvitz- Thompson.

As estratégias amostrais E1, E2 e E3 foram consideradas no trabalho de Silva e Moura (1990) onde coeficientes de correlação intraclasse, EPAs e tamanhos amostrais foram calculados para nove regiões metropolitanas considerando dados do Censo Demográfico de 1980. Neste trabalho será considerada também a E4 pois é o método de seleção da Amostra Mestra do Sistema Integrado de Pesquisas Domiciliares implantado pelo IBGE a partir de 2012.

A amostragem de Pareto é um dos métodos de amostragem por ordenação que utiliza probabilidade proporcional ao tamanho (Rosén, 1997). Para o sorteio das unidades é utilizada a distribuição de Pareto para geração das variáveis de ordenação que auxiliam na seleção da amostra.

Amostragem por ordenação consiste na associação de um número aleatório a cada unidade da população, e em sequência na seleção das unidades associadas aos menores números aleatóriosAmostragem PPT de Pareto é um dos métodos de seleção das unidades primárias de amostragem (UPAs) considerado nesta dissertação. Mais detalhes e passos para o sorteio de UPAS através do método de seleção de Pareto estão no apêndice A.1.

2.2 Notações para população e amostra

Nesta seção apresenta-se a notação a ser usada neste trabalho:

AAS	Amostragem Aleatória Simples sem reposição
AC2	Amostragem Conglomerada em 2 estágios;
UPA	Unidade Primária de Amostragem, isto é, unidade a ser selecionada no primeiro estágio de seleção;
USA	Unidade Secundária de Amostragem, isto é, unidade a ser selecionada no segundo estágio de seleção;

$U = \{1,2,...,i,...,M\}$	O conjunto U representa a população de UPAs;
M	Número de (UPAs) na população U;
y	A variável de interesse a ser observada nas unidades da amostra selecionada;
y_{ij}	Valor da variável y para USA j da UPA i;
N_i	Número de USAs na UPA i;
$N = \sum_{i \in U} N_i$	Número de USAs na população;
$U_i = \{1,2,..., j,..., N_i\}$	População de unidades secundárias da UPA i;
$Y_i = \sum_{j \in U_i} Y_{ij}$	O valor total da variável y para a i-ésima UPA;
$Y = \sum_{i=1}^{M} Y_i = \sum_{i \in U} Y_i$	O total populacional de uma variável de interesse y;
$\overline{Y}_i = Y_i / N_i$	A média da variável de interesse y na UPA i;
$\overline{Y}_C = Y / M$	O total médio da variável y por UPA;
$\overline{\overline{Y}} = Y / N$	A média populacional da variável y por USA;
m	Número de UPAs selecionadas no primeiro estágio;
$s = \{i_1, i_2,..., i_m\}, \quad s \subset U$	A amostra de UPAs selecionada da população U;
n_i	Número de USAs selecionadas na UPA i;
$n = \sum_{i \in s} n_i$	Número total de USAs selecionadas no segundo estágio;
$\pi_i = \Pr(i \in s)$	A probabilidade de inclusão da i-ésima UPA na amostra s;
$a_i = \{j_1, j_2,..., j_{n_i}\}, \quad a_i \subset U_i$	Amostra de USAs selecionadas da UPA i;
$A = \bigcup_{i \in s} a_i$	é o conjunto das USAs selecionadas para a amostra

$\pi_{ij} = \Pr[i \in s, j \in a_i] = \Pr[i \in s].\Pr[j \in a_i \mid i \in s] = \pi_i . \pi_{j|i}$ A probabilidade de inclusão da USA j da UPA i na amostra;

$\pi_{j/i} = \Pr(j \in a_i / i \in s)$ A probabilidade de inclusão da USA j na amostra a_i dado que a UPA i foi selecionada para a amostra s.

2.3 Estimadores de totais dos planos conglomerados em dois estágios

O estimador de Horvitz-Thompson para o total populacional Y é dado por (Cochran,1977, eq. 11.7):

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{\hat{Y}_i}{\pi_i} = \sum_{i \in s} \frac{1}{\pi_i} \sum_{j \in a_i} \frac{y_{ij}}{\pi_{j|i}} = \sum_{i \in s} \sum_{j \in a_i} w_{ij}^{HT} y_{ij} \qquad (1)$$

onde

$\hat{Y}_i = \sum_{j \in a_i} \frac{y_{ij}}{\pi_{j|i}}$ é um estimador HT do total Y_i da UPA i;

$w_{ij}^{HT} = \pi_{ij}^{-1} = \pi_i^{-1} \pi_{j|i}^{-1}$ é o peso natural (básico, do desenho) associado à USA j da UPA i.

Este estimador foi utilizado nas estratégias amostrais E1 e E4, e também como base para o estimador de razão considerado na estratégia E2.

Um estimador de razão (Särndal, Swensson & Wretman, 1992, res. 5.6.3) para o total Y baseado no tamanho das UPAs é dado por

$$\hat{Y}_R = N \left(\sum_{i \in s} \hat{Y}_i / \pi_i \right) / \left(\sum_{i \in s} N_i / \pi_i \right) \qquad (2)$$

quando N for conhecido.

Este estimador foi utilizado na estratégia amostral E2.

Um estimador natural para o total Y sob a estratégia E3 segundo Cochran (1977, eq. 11.31) é dado por:

$$\hat{Y} = \frac{1}{m} \sum_{i \in s} \frac{\hat{Y}_i}{p_i} \qquad (3)$$

Onde $p_i = \frac{N_i}{N}$ é a probabilidade de seleção da UPA i em um único sorteio, quando os tamanhos das UPAs N_i são usados para definir as probabilidades de seleção na estratégia E3.

2.4 Variância de estimadores de totais em planos conglomerados em dois estágios

A variância de estimadores de total sob planos conglomerados em dois estágios pode ser decomposta como a soma da componente da variância devida ao sorteio das unidades primárias de amostragem (UPAs) e a componente da variância devida ao sorteio das unidades secundárias de amostragem (USAs). A variância do estimador de Horvitz-Thompson sob estes planos (Cochran, 1977, eq. 10.7) é:

$$V_{AC2}\left(\hat{Y}_{HT}\right) = V_1\left[E_2\left(\sum_{i\in s}\frac{\hat{Y}_i}{\pi_i}\right)\right] + E_1\left[V_2\left(\sum_{i\in s}\frac{\hat{Y}_i}{\pi_i}\right)\right] = V_{UPA} + V_{USA} \qquad (4)$$

Sendo V_{UPA} a componente da variância de $\hat{Y}_{HT}$ originada da amostragem de unidades primárias no estágio 1 e V_{USA} a componente da variância de $\hat{Y}_{HT}$ originada da amostragem de unidades secundárias no estágio 2 .

Quando o sorteio de UPAs é feito com probabilidades proporcionais aos tamanhos N_i, seguem os resultados para as variâncias dos estimadores de total considerados na seção 2.3.

Variância do estimador de total sob a estratégia amostral E1 (Cochran 1977, eq. 11.22):

$$V_{E1}\left(\hat{Y}_{HT}\right) = M^2\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{(M-1)}\sum_{i\in U}\left(Y_i - \overline{Y}_C\right)^2 + \frac{M}{m}\sum_{i\in U}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2 \qquad (5)$$

Onde $S_i^2 = \frac{1}{N_i - 1}\sum_{j\in U_i}\left(y_{ij} - \overline{Y}_i\right)^2$.

Variância do estimador de total sob a estratégia amostral E2 (Cochran 1977, eq. 11.28):

$$V_{E2}\left(\hat{Y}_{HT}\right)= M^2\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{(M-1)}\sum_{i\in U} N_i^2\left(\overline{Y}_i-\overline{\overline{Y}}\right)^2 + N^2\left(\frac{1}{n}-\frac{1}{m\overline{N}}\right)\sum_{i\in U}\frac{N_i}{N}S_i^2 \qquad (6)$$

Onde $\overline{N}=\frac{N}{M}$.

A variância para o estimador de total sob a estratégia amostral E3 (Silva e Moura, 1990 p.30) é dada por:

$$V_{E3}\left(\hat{Y}\right)= \frac{1}{m}\sum_{i\in U}\left(\frac{Y_i}{p_i}-Y\right)^2 p_i + \frac{1}{m}\sum_{i\in U}\frac{N_i^2}{p_i}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2 \qquad (7)$$

A variância do estimador Horvitz-Thompson para o total sob a estratégia amostral E4 (ver apêndice A.2) é dada por:

$$V_{E4}\left(\hat{Y}_{HT}\right)= \frac{M}{M-1}\left\{\sum_{i\in U}\frac{Y_i^2}{\pi_i}(1-\pi_i) - \frac{\left[\sum_{i\in U} Y_i(1-\pi_i)\right]^2}{m-\sum_{i\in U}\pi_i^2}\right\} + \sum_{i\in U}\frac{1}{\pi_i}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2 \qquad (8)$$

Com $\pi_i = m\frac{N_i}{N}$.

3. Principais Pesquisas Amostrais Domiciliares realizadas pelo IBGE

Os principais aspectos comuns do desenho amostral das pesquisas amostrais domiciliares do IBGE envolvem estratificação, conglomeração, probabilidades desiguais de seleção e calibração para ajuste de pesos. A amostragem por conglomerados é utilizada em mais de um estágio de seleção, geralmente com a seleção de setores censitários como UPAs e posteriormente com a seleção de domicílios como USAs.

Pesquisa Nacional por Amostra de Domicílios (PNAD): O desenho amostral da PNAD envolve estratificação e conglomeração em um, dois ou três estágios de seleção, dependendo do estrato de seleção. Veja por exemplo Lila (2004).

Pesquisa Mensal de Emprego (PME): o plano da PME emprega estratificação e conglomeração em dois estágios. A estratégia amostral E3 desta dissertação e utilizada para aproximar variâncias de estimadores de total na PME.

Sistema Integrado de Pesquisas Domiciliares (SIPD) e Amostra Mestra: O SIPD (ver IBGE 2007b) representa a integração de pesquisas domiciliares realizadas pelo IBGE. O plano amostral da Pesquisa Nacional por Amostra de Domicílios Contínua coincide com a estratégia amostral E4, a menos de sua estratificação mais detalhada.

Pesquisa de Orçamento Familiar (POF): Na edição 2008/2009 da POF os setores censitários foram selecionados da Amostra Mestra com seleção por amostragem aleatória simples de setores. Portanto seu plano amostral também é similar ao da estratégia E4.

4. Efeitos do Plano Amostral e coeficientes de correlação intraclasse

O efeito do plano amostral (EPA) é uma medida da eficiência estatística de um desenho amostral complexo comparando com um desenho que utilize amostragem aleatória simples (AAS). É uma razão de variâncias: a variância de um estimador sob um plano amostral complexo e a variância do estimador sob AAS.

Segundo Silva e Moura (1990), o coeficiente de correlação íntraclasse "é uma medida do grau de homogeneidade ou similaridade existente entre os elementos pertencentes a determinados grupos, a união dos quais constitui a população".

Relação entre coeficiente de correlação intraclasse e EPA

Em um plano amostral conglomerado em dois estágios o Efeito do Plano Amostral é definido como a variância do estimador AC2 sobre a variância do estimador sob uma amostra aleatória simples (AAS). Silva e Moura (1990) mostram que esta razão pode ser aproximada para alguns planos AC2 por uma função do coeficiente de correlação intraclasse da seguinte forma:

$$EPA\left(\hat{Y}_{AC2}\right)=\frac{V_{AC2}\left(\hat{Y}\right)}{V_{AAS}\left(\hat{Y}\right)}\approx 1+\left(\bar{n}-1\right)\rho \qquad (9)$$

Onde $\bar{n}=\frac{n}{m}$ **e** ρ é um coeficiente de correlação intraclasse adequado.

Quando (9) é satisfeita, o tamanho da amostra n_{AC2} em um plano conglomerado em dois estágios é aproximado pela seguinte relação:

$$n_{AC2}\approx \quad n_{AAS} \quad . \quad EPA\left(\hat{Y}_{AC2}\right) \qquad (10)$$

onde n_{AAS} é o tamanho amostral sob AAS.

Segundo a expressão (10) o tamanho amostral de USAs em um plano AC2 é aproximadamente o tamanho amostral calculado por AAS multiplicado pelo efeito do plano amostral correspondente. Através dessa relação o cálculo de tamanhos de amostra em pesquisas por AC2 fica facilitado.

4.1 Efeito do Plano Amostral para a estratégia E1

Para a estratégia amostral E1, Silva e Moura (1990) mostraram que o efeito do plano amostral pode ser aproximado por uma função do coeficiente de correlação intraclasse dada por:

$$EPA\left(E1,\hat{Y}_{HT}\right)=\frac{V_{E1}\left(\hat{Y}_{HT}\right)}{V_{AAS}\left(\hat{Y}_{HT}\right)}\approx\frac{\sigma_a^2}{\sigma_b^2}\left[1+\rho_e\left(\bar{n}-1\right)\right] \qquad (11)$$

onde:

$$\sigma_a^2=\frac{\sigma_e^2}{\bar{N}^2}+\frac{\bar{N}-1}{\bar{N}}\frac{M}{N}\sigma_d^2, \quad ,$$

$$\rho_e=\frac{\sigma_e^2-\sigma_d^2}{\sigma_e^2+\left(\bar{N}-1\right)\sigma_d^2}=1-\frac{\bar{N}\sigma_d^2}{\sigma_e^2+\left(\bar{N}-1\right)\sigma_d^2},\ \sigma_i^2=\frac{N_i-1}{N_i}S_i^2,$$

$$\sigma_e^2 = \frac{1}{M}\sum_{i\in U}\left(Y_i - \overline{Y}_C\right)^2 \text{ e } \sigma_d^2 = \frac{1}{M}\sum_{i\in U}\frac{N_i^2}{N_i - 1}\sigma_i^2$$

$$V_{AAS}\left(\hat{Y}_{HT}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 \qquad (12)$$

Onde: $S_y^2 = \frac{1}{N-1}\sum_{i\in U}\sum_{j\in U}\left(y_{ij} - \overline{\overline{Y}}\right)^2$

Podemos observar que ρ_e é uma função de variâncias entre $\left(\sigma_e^2\right)$ e dentro $\left(\sigma_d^2\right)$ dos conglomerados, e que se a variância dentro dos conglomerados for nula, este coeficiente assume o valor 1, indicando a homogeneidade máxima dentro dos conglomerados, onde todos os elementos dentro de um conglomerado têm valores idênticos da variável y.

4.2 Efeito do Plano Amostral para a estratégia E2

Para a estratégia amostral E2, onde é utilizado um estimador do tipo razão, Silva e Moura (1990) mostraram que o EPA é aproximado pela expressão:

$$EPA\left(E2, \hat{Y}_R\right) = \frac{V_{E2}\left(\hat{Y}_R\right)}{V_{AAS}\left(\hat{Y}_{HT}\right)} \approx \frac{\sigma_c^2}{\sigma_b^2}\left[1 + \rho_f\left(\overline{n} - 1\right)\right] \qquad (13)$$

Onde:

$$o_o^2 = \frac{1}{\overline{\overline{N}}^2}\frac{1}{M}\sum_{i\in U}\left(Y_i - N_i\overline{\overline{Y}}\right)^2 + \frac{\overline{N} - 1}{\overline{N}}\frac{1}{N}\sum_{i\in U}\frac{N_i^2}{N_i - 1}\sigma_i^2 \qquad \rho_f = 1 - \frac{\overline{N}\sigma_d^2}{\sigma_{e,2}^2 + \left(\overline{N} - 1\right)\sigma_d^2} \text{ e}$$

$$o_{e,2}^2 = \frac{1}{M}\sum_{i\in U}\left(Y_i - N_i\overline{\overline{Y}}\right)^2$$

4.3 Efeito do Plano Amostral para a estratégia E3

Para a estratégia amostral E3 Silva e Moura (1990) mostraram que:

$$EPA\left(E3,\hat{Y}\right)=\frac{V_{E3}\left(\hat{Y}\right)}{V_{AAS}\left(\hat{Y}_{HT}\right)}\approx 1+\left(\overline{n}-1\right)\rho_C \qquad (14)$$

Onde $\rho_C=1-\dfrac{\frac{1}{N}\sum_{i\in U}N_i\sigma_i^2\left(1+\frac{1}{N_i-1}\right)}{\sigma^2}$ e $o^2=\dfrac{1}{N}-\sum_{i\in U}N_i\sigma_i^2+\dfrac{1}{N}\sum_{i\in U}N_i\left(Y_i-\overline{\overline{Y}}\right)^2$.

4.4 Efeito do Plano Amostral para a estratégia E4

O efeito do plano amostral para a estratégia amostral E4 é obtido através da razão de variâncias:

$$EPA\left(E4,\hat{Y}_{HT}\right)=\frac{V_{E4}\left(\hat{Y}_{HT}\right)}{V_{AAS}\left(\hat{Y}_{HT}\right)} \qquad (15)$$

Onde

$V\left(\hat{Y}_{E4}\right)$ **é dado pela expressão em (8) com** $\pi_i=m\dfrac{N_i}{N}$ e $V_{AAS}\left(\hat{Y}_{HT}\right)$ **é dado em (12).**

5. O Censo Demográfico 2010

5.1 Aspectos Gerais

Os dados utilizados neste trabalho se referem aos microdados do Censo Demográfico 2010 da Unidade de Federação de Minas Gerais. O Censo Demográfico brasileiro é composto por dois questionários: o questionário básico e o questionário de amostra.

O questionário básico é formado por um conjunto pequeno de perguntas fundamentais sobre o domicílio e seus moradores na data de referência e foi aplicado nos domicílios que não foram selecionados para a amostra. Já o questionário da amostra continha toda a investigação presente no questionário básico mais algumas características do domicílio e outras informações sociais, econômicas e demográficas dos moradores (IBGE, 2012b). Foi aplicado em uma parcela previamente definida da população estabelecida pela fração amostral adotada.

Foram utilizadas algumas variáveis do questionário básico e algumas do questionário de amostra para ilustrar os métodos propostos. Para o cálculo de coeficientes de correlação intraclasse e EPAs são necessárias algumas estatísticas resumo por setor, tais como:

- Total da variável de interesse y por UPA $\left(Y_i\right)$;
- Variância da variável por UPA $\left(S_i^2\right)$;
- Tamanho de cada UPA $\left(N_i\right)$.

Para as variáveis do questionário básico, estas quantidades foram calculadas e usadas diretamente nas expressões dos EPAs apresentadas na seção 4. Já para as variáveis investigadas somente no questionário da amostra os totais e variâncias são desconhecidos e tiveram que ser estimados. A estimação dos EPAs é descrita na seção 5.2.

5.2 Estimação dos EPAs a partir da amostra do Censo Demográfico

O questionário da Amostra do Censo Demográfico 2010 foi aplicado a cerca de 11% dos domicílios brasileiros. A fração amostral nos setores variou de acordo com total o populacional do município.

A amostra do Censo Demográfico forneceu as seguintes quantidades:

M ..... Número de setores em uma área qualquer;

N_i ..... Número de domicílios no i-ésimo setor da área de interesse , $i \in \{1,2,...,M\}$**;**

n_i ..... Número de domicílios selecionados para a amostra do Censo no i-ésimo setor **;**

y_{ij} ..... Valor da variável de interesse y observado no j-ésimo domicílio da amostra do i-ésimo setor da área de interesse, $j \in a_i$ e $i \in \{1, 2,..., M\}$

Supondo que a amostra de domicílios dentro dos setores foi obtida por Amostragem Aleatória Simples sem reposição temos que:

$\bar{y}_i = \frac{1}{n_i} \sum_{j \in a_i} y_{ij}$ é um estimador não viciado para $\bar{Y}_i$ e

$s_i^2 = \frac{1}{n_i - 1} \sum_{j \in a_i} \left(y_{ij} - \bar{y}_i\right)^2$ é um estimador não viciado para S_i^2

Como a média amostral $\bar{y}_i$ é não viciada para estimar a média populacional $\bar{Y}_i$ na UPA i e s_i^2 é não viciado para estimar a variância populacional S_i^2 na UPA i e as expressões abaixo são funções dessas medidas, temos que essas expressões (16), (18), (19) e (20) fornecem estimadores consistentes para os EPAs definidos em (11), (13), (14) e (15), respectivamente.

Estimação do EPA para a estratégia amostral E1

Para a estratégia amostral E1, o efeito do plano amostral estimado através dos microdados da amostra do Censo é o seguinte:

$$\text{epa}\left(\text{E}1, \hat{Y}_{HT}\right) = \frac{S_a^2}{S_b^2}\left[1 + \hat{p}_o\,(n-1)\right] \qquad (16)$$

Onde:

$$s_a^2 = \frac{2_e^2}{\overline{N}^2} + \frac{\overline{N}-1}{\overline{N}}\frac{M}{N}S_d^2 \quad , \quad s_e^2 = \frac{1}{M}\sum_{i\in U}\left(N_i\,\overline{y}_i - \hat{\overline{Y}}_C\right)^2 - \frac{M-1}{M^2}\sum_{i\in U}\frac{N_i\,(N_i-n_i)}{n_i}S_i^2,$$

$$\hat{Y}_C = \frac{\hat{Y}}{M} = \frac{\sum_{i\in s}\frac{N_i}{n_i}\sum_{j\in a_i} y_{ij}}{M} = \frac{\sum_{i\in s} N_i\,\overline{Y}_i}{M}, s_d^2 = \frac{1}{M}\sum_{i\in U} N_i\,S_i^2 \ ,$$

$$S_b^2 = \frac{1}{N-1}\sum_{i\in U}(N_i-1)s_i^2 + \frac{N}{N-1}S_{e,3}^2 \qquad (17)$$

$$S_{e,3}^2 = \frac{1}{N}\sum_{i\in U} N_i(\overline{y}_i - \overline{y})^2 - \frac{1}{N}\sum_{i\in U}\left(1-\frac{N_i}{N}\right)(N_i-n_i)\frac{s_i^2}{n_i} \quad \text{onde} \quad \overline{y} = \frac{y}{n} = \frac{\sum_{i\in s} Y_i}{n}$$

$$\hat{\rho}_e = \frac{S_e^2 - S_d^2}{S_e^2 + (\overline{N}-1)S_d^2}$$

Estimação do EPA para a estratégia amostral E2

$$\text{epa}\left(E2, \hat{Y}_R\right) = \frac{S_C^2}{S_b^2}\left[1 + \hat{\rho}_f\,(\overline{n}-1)\right] \qquad (18)$$

Onde:

$$S_c^2 = \frac{S_{e,2}^2}{\overline{N}^2} + \frac{\overline{N}-1}{\overline{N}}\,\frac{M}{N}S_d^2$$

$$\hat{\rho}_f = \frac{S_{e,2}^2 - S_d^2}{S_{e,2}^2 + (\overline{N}-1)S_d^2}$$

$$S_{e,2}^2 = \frac{1}{M}\left[\sum_{i\in U} N_i^2 (\bar{y}_i - \bar{\bar{y}})^2 - \sum_{i \in U} N_i (N_i - n_i)\frac{s_i^2}{n_i}\left(1 - \frac{2N_i}{N} + \frac{1}{N^2}\sum_{K\in U} N_K^2\right)\right]$$

$$\bar{\bar{y}} = \frac{1}{N}\sum_{i \in U} N_i \bar{y}_i$$

Estimação do EPA para a estratégia amostral E3

$$epa\left(E3, \hat{Y}\right) = 1 + (\bar{n} - 1)\hat{\rho}_C \qquad (19)$$

$$\hat{\rho}_C = \frac{S_{e,3}^2 - \frac{1}{N}\sum_{i\in U} S_i^2}{S_{e,3}^2 + \frac{1}{N}\sum_{i\in U}(N_i - 1)S_i^2}$$

Estimação do EPA para a estratégia amostral E4

$$epa\left(E4, \hat{Y}_{HT}\right) = \hat{V}_{E4}\left(\hat{Y}_{HT}\right) \;/\; \hat{V}_{AAS}\left(\hat{Y}_{HT}\right) \qquad (20)$$

A estimação do efeito do plano amostral é feita considerando a estimação de parâmetros desconhecidos nas expressões de cada uma das variâncias em (15) (ver apêndice A.3).

$\hat{V}_{AAS}(\hat{Y}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\hat{S}_y^2$ onde $\hat{S}_y^2 = \frac{N-1}{N}s_b^2$ e s_b^2 é definido em (17).

6. Resultados

Os resultados deste trabalho são estimativas de Efeitos do Plano Amostral para as quatro estratégias amostrais através dos microdados de Minas Gerais, Região Metropolitana de Belo Horizonte e Interior, considerando $\bar{n} = 14$. Para as estratégias E1, E2 e E3 também foram estimados coeficientes de correlação intraclasse.

Essas estimativas podem ser usadas para calcular tamanhos amostrais dos planos conglomerados em dois estágios considerados.

6.1 Resultados para variáveis do questionário básico

Na quadro 1 estão 5 variáveis selecionadas entre as contidas no questionário básico do Censo Demográfico 2010. Arantes (2012) fornece estimativas de EPAs e coeficientes de correlação intraclasse para outras 15 variáveis, além das variáveis consideradas no quadro 1. As variáveis foram selecionadas de acordo com a importância socioeconômica e demográfica.

Quadro 1: Variáveis selecionadas do questionário básico do Censo Demográfico 2010

Variável	Descrição
B2	Total de moradores responsáveis analfabetos com 10 anos ou mais de idade em domicílio particular permanente (DPP)
B9	Total de mulheres responsáveis pelo domicílio em DPP
B17	Número de domicílios particulares permanentes com energia elétrica
B18	Número de domicílios particulares permanentes unipessoais
B20	Total de moradores com renda per capita abaixo de meio salário mínimo em DPP

As tabelas 1, 2 e 3 mostram os coeficientes de correlação intraclasse e efeitos do plano amostral calculados para Minas Gerais, para a RM de Belo Horizonte e para o interior de Minas Gerais, respectivamente. Para as três áreas consideradas e para todas as variáveis observamos a seguinte tendência para os coeficientes de correlação intraclasse obtidos: $\rho_c < \rho_f < \rho_e$. Nota-se também a tendência correspondente para os efeitos do plano amostral: $EPA_{E4} < EPA_{E3} < EPA_{E2} \ll EPA_{E1}$ como era esperado, pois quanto maior é o coeficiente de correlação intraclasse (no caso de E1, E2 e E3) maior é o efeito do plano amostral. A exceção a esta "regra" é a variável B17, onde o EPA para E2 é o menor de todos. Note que aqui o símbolo $\ll$ é usado para indicar que os valores de EPA_{E2} são muito menores que os de EPA_{E1}.

Esses resultados sugerem a superioridade da estratégia E4 sobre as demais estratégias consideradas. Silva e Moura (1990) haviam constatado a superioridade de E3 sobre E1 e E2, sendo este resultado confirmado aqui. Os resultados confirmam a vantagem de planos com sorteio PPT de setores sobre os que praticam sorteio AAS dos setores. Entre os planos PPT, a vantagem foi para a estratégia E4 em que o plano amostral seleciona os setores sem reposição. Planos amostrais sem reposição possuem menores variâncias dos estimadores que planos com reposição de tamanhos equivalentes.

Os setores mais homogêneos representam conglomerados com elementos mais parecidos e não contribuem para captar a variabilidade populacional. Portanto quanto mais homogêneos forem, mais setores serão necessários na amostra.

Tabela 1: Coeficientes de Correlação Intraclasse e Efeitos do Plano Amostral por estratégia amostral para variáveis do questionário básico do Censo Demográfico 2010, Minas Gerais

Variável	E1		E2		E3		E4
	ρ_e	EPA_{E1}	ρ_f	EPA_{E2}	ρ_c	EPA_{E3}	EPA_{E4}
B2	0,13	2,66	0,12	2,63	0,12	2,52	2,46
B9	0,21	4,48	0,08	2,05	0,07	1,94	1,90
B17	0,98	509,81	0,09	2,05	0,13	2,63	2,57
B18	0,05	1,73	0,02	1,28	0,02	1,23	1,21
B20	0,24	4,52	0,19	3,55	0,16	3,14	3,05

Tabela 2: Coeficientes de Correlação Intraclasse e Efeitos do Plano Amostral por estratégia amostral para variáveis do questionário básico do Censo Demográfico 2010, RM de BH

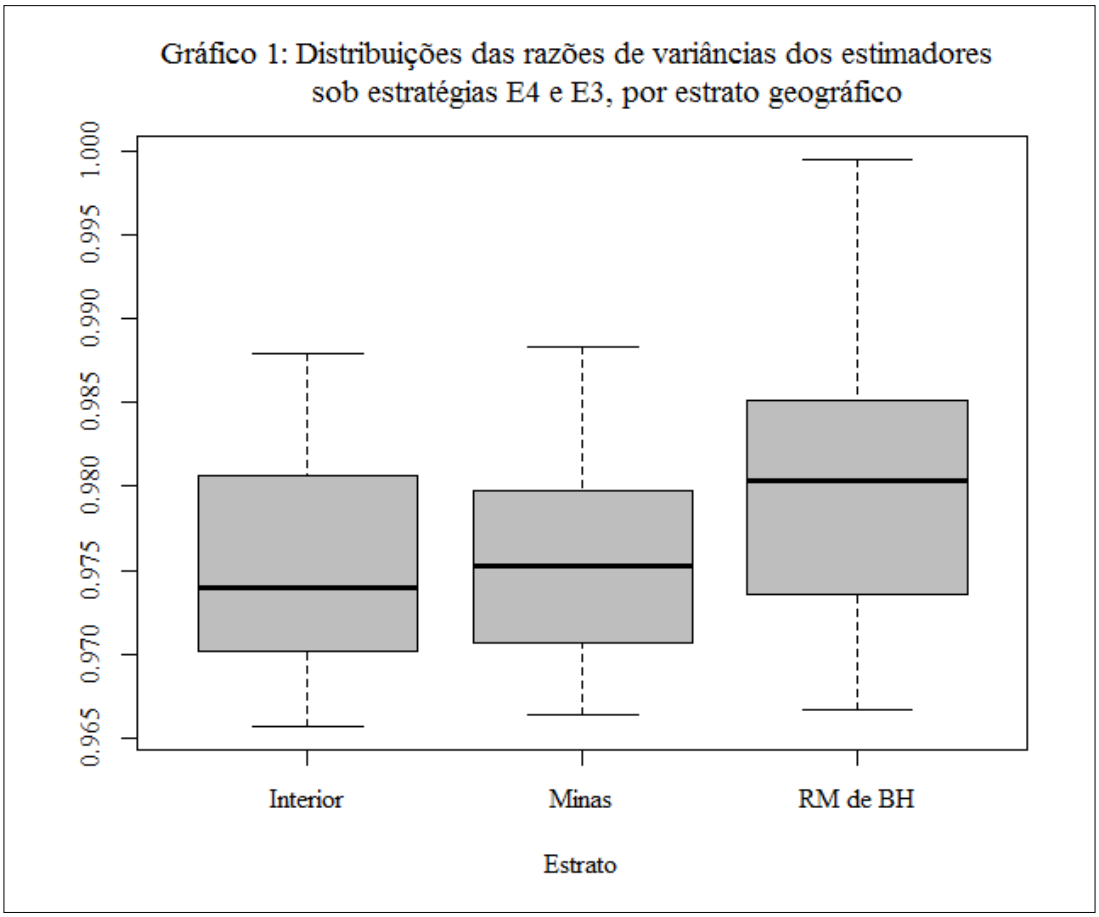
Variável	E1		E2		E3		E4
	ρ_e	EPA_{E1}	ρ_f	EPA_{E2}	ρ_c	EPA_{E3}	EPA_{E4}
B2	0,05	1,72	0,05	1,62	0,04	1,56	1,53
B9	0,19	4,05	0,06	1,86	0,05	1,71	1,68
B17	0,99	2598,84	0,02	1,26	0,04	1,53	1,58
B18	0,05	1,75	0,03	1,34	0,02	1,28	1,27
B20	0,20	3,95	0,15	3,01	0,13	2,65	2,58

Tabela 3: Coeficientes de Correlação Intraclasse e Efeitos do Plano Amostral por estratégia amostral para variáveis do questionário básico do Censo Demográfico 2010, Interior de MG

Variável	E1		E2		E3		E4
	ρ_e	EPA_{E1}	ρ_f	EPA_{E2}	ρ_c	EPA_{E3}	EPA_{E4}
B2	0,13	2,70	0,13	2,66	0,12	2,50	2,44
B9	0,22	4,54	0,08	2,08	0,08	1,98	1,93
B17	0,97	401,05	0,09	2,10	0,13	2,67	2,61
B18	0,05	1,72	0,02	1,25	0,02	1,20	1,19
B20	0,25	4,65	0,19	3,59	0,16	3,14	3,04

O gráfico 1 apresenta a distribuição da razão dos EPAs das estratégias amostrais E4 e E3 observadas para 16 variáveis do questionário básico cujas razões são menores que 1, calculados por Arantes (2012). Esta razão de EPAs pode ser interpretada também como a razão de variâncias entre as duas estratégias, pois:

$$\frac{EPA\left(E4,\hat{Y}_{HT}\right)}{EPA\left(E3,\hat{Y}\right)} = \frac{V_{E4}\left(\hat{Y}_{HT}\right)/V_{AAS}\left(\hat{Y}\right)}{V_{E3}\left(\hat{Y}\right)/V_{AAS}\left(\hat{Y}\right)} = \frac{V_{E4}\left(\hat{Y}_{HT}\right)}{V_{E3}\left(\hat{Y}\right)}$$



Neste gráfico observamos que a mediana das razões da região metropolitana de Belo Horizonte é maior que a mediana das razões de Minas Gerais, que por sua vez é maior que a mediana do interior. Isso significa que a estratégia amostral E4 é mais eficiente que a estratégia amostral E3 principalmente no interior de Minas. Observa-se também que os ganhos de E4 sobre E3 são modestos, como esperado.

6.2 Resultados para variáveis do questionário da Amostra

No quadro 2 estão 4 variáveis selecionadas do questionário da amostra do Censo Demográfico 2010. Arantes (2012) estimou EPAs e coeficientes de correlação intraclasse para outras 11 variáveis, além das variáveis consideradas no quadro 2. As variáveis da amostra também foram selecionadas de acordo com a importância socioeconômica e demográfica.

Quadro 2: Variáveis selecionadas do questionário da amostra do Censo Demográfico 2010

Variável	Descrição
A5	Número de moradores não naturais da UF em domicílio particular permanente
A10	Número de filhos nascidos vivos de mulheres de 15 anos ou mais em DPP
A12	Rendimento mensal total de moradores em DPP
A13	Domicílios particulares permanentes que possuem geladeira

As tabelas 4, 5 e 6 contém as estimativas dos efeitos do plano amostral e coeficientes de correlação intraclasse para as estratégias consideradas calculadas através dos microdados da amostra para Minas Gerais, RM de Belo Horizonte e interior de Minas.

Observando essas tabelas nota-se que a tendência $\rho_c < \rho_f < \rho_e$ também se verifica com as estimativas dos EPAs estimados para as variáveis do questionário da amostra. Observa-se também que $EPA_{E4} < EPA_{E3} < EPA_{E2} << EPA_{E1}$.

Algumas variáveis possuem um grau de homogeneidade maior considerando a estratégia 1, como por exemplo, a variável A13, gerando efeito de conglomeração bastante alto.

Em relação às diferenças entre as estimativas das tabelas 4, 5 e 6 nota-se que na região metropolitana de Belo Horizonte a variável A13 (DPP que possui geladeira) apresenta um aumento na homogeneidade $(\rho_e = 0{,}91)$ gerando um aumento significativo no efeito do plano amostral $(EPA_{E1} = 142)$, considerando a estratégia 1.

Tabela 4: Coeficientes de Correlação Intraclasse e Efeitos do Plano Amostral por estratégia amostral, para variáveis do questionário da amostra, Censo Demográfico 2010, Minas Gerais

Variável	E1		E2		E3		E4
	ρ_e	epa_{E1}	ρ_f	epa_{E2}	ρ_c	epa_{E3}	epa_{E4}
A5	0,32	6,14	0,22	3,96	0,18	3,38	3,27
A10	0,20	4,20	0,06	1,78	0,05	1,69	1,67
A12	0,20	3,93	0,16	3,21	0,14	2,82	2,74
A13	0,84	65,33	0,11	2,36	0,11	2,45	2,40

Tabela 5: Coeficientes de Correlação Intraclasse e Efeitos do Plano Amostral por estratégia amostral, para variáveis do questionário da amostra, Censo Demografico 2010, RM de BH

Variável	E1		E2		E3		E4
	ρ_e	epa_{E1}	ρ_f	epa_{E2}	ρ_c	epa_{E3}	epa_{E4}
A5	0,35	6,78	0,25	4,37	0,21	3,76	3,64
A10	0,18	3,90	0,05	1,68	0,05	1,59	1,56
A12	0,21	3,95	0,19	3,48	0,17	3,25	3,16
A13	0,91	141,93	0,02	1,25	0,02	1,26	1,25

Tabela 6: Coeficientes de Correlação Intraclasse e Efeitos do Plano Amostral por estratégia amostral, para variáveis do questionário da amostra, Censo Demografico 2010, Interior de MG

Variável	E1		E2		E3		E4
	ρ_e	epa_{E1}	ρ_f	epa_{E2}	ρ_c	epa_{E3}	epa_{E4}
A5	0,24	4,78	0,14	2,82	0,12	2,53	2,47
A10	0,20	4,32	0,06	1,76	0,05	1,67	1,64
A12	0,16	3,34	0,11	2,43	0,08	2,08	2,02
A13	0,82	56,17	0,11	2,43	0,11	2,47	2,42

7. Cálculo de tamanho de amostra

Nos desenhos amostrais complexos a obtenção de tamanhos amostrais não é tão trivial e em muitos casos é impossível de se obter. Como esses desenhos em geral envolvem um ou mais estágios de seleção e probabilidades desiguais de seleção, também demandam mais informações para o cálculo dos tamanhos amostrais (Silva & Moura, 1990).

Segundo Silva (2002), para determinar o tamanho amostral de uma pesquisa que emprega um plano AC2 basta seguir o seguinte roteiro:

1) Escolher um parâmetro que se deseja estimar (total, média, proporção);

2) Fixar a margem de erro máximo para estimação do parâmetro de interesse;

3) Calcular o tamanho amostral sob amostragem aleatória simples (n_{AAS});

4) Estimar o efeito do plano amostral para um valor de $\bar{n}$ fixado com base em alguma pesquisa anterior, como o Censo Demográfico, por exemplo;

5) Calcular o tamanho de amostra sob um plano amostral complexo utilizando a relação dada em (10): $n_{AC2} \approx EPA.n_{AAS}$;

6) Escolher o valor de $\bar{n}$ e consequentemente de m em função de custos e precisão.

7) Calcular o tamanho amostral de UPAs utilizando a seguinte relação $m_{AC2} = n_{AC2} / \bar{n}$.

Exemplo de cálculo de tamanho de amostra: Suponha que estamos interessados em calcular o tamanho de amostra para estimar o rendimento mensal total de moradores em domicílios particulares permanentes (DPPs) em Minas Gerais. Para isto planejamos um desenho amostral semelhante à estratégia E3 (ver p. 3). Utilizaremos a relação dada em (10): $n_{AC2} \approx EPA.n_{AAS}$. Observe que o termo que queremos encontrar **é** n_{AC2}. Suponha que encontramos o termo desconhecido n_{AAS} e o valor 1000 para este tamanho amostral com um certo erro de estimação pré-determinado. Para encontrarmos o tamanho amostral do plano referente à estratégia E3, basta multiplicarmos o EPA correspondente (tabela 4) pelo tamanho amostral que utiliza amostra aleatória simples, logo $n_{AC2} \approx EPA.n_{AAS} \approx 2{,}82x1000 = 2820$ domicílios. Como o EPA utilizado considera que

o tamanho amostral em cada UPA é 14, temos que o tamanho amostral de UPAs será 2820/14 que é aproximadamente 202 UPAs em Minas Gerais.

Para o cálculo de tamanhos de amostras estratificadas com uma precisão global fixada, podemos encontrar o tamanho de amostra global através da aproximação no item 5 do roteiro anterior. Sejam 1, ... , H os estratos que juntos formam a população de interesse. A variância do estimador do parâmetro de interesse populacional global (o total Y por exemplo) denotaremos por $V_{AC2E}\left(\hat{Y}\right)$. Assim o efeito do plano amostral é dado por $EPA = \frac{V_{AC2E}\left(\hat{Y}\right)}{V_{AAS}\left(\hat{Y}\right)}$. Como sabemos que o tamanho amostral global pode ser obtido da aproximação descrita no item 5 do roteiro anterior, temos a seguinte relação:

$$\frac{n_{AC2}}{n_{AAS}} \approx \frac{V_{AC2E}\left(\hat{Y}\right)}{V_{AAS}\left(\hat{Y}\right)} \quad \rightarrow \quad n_{AC2} \approx \frac{V_{AC2E}\left(\hat{Y}\right)}{V_{AAS}\left(\hat{Y}\right)} n_{AAS} \rightarrow \sum_{i=1}^{H} n_h \approx \frac{V_{AC2E}\left(\hat{Y}\right)}{V_{AAS}\left(\hat{Y}\right)} n_{AAS}.$$

A expressão $V_{AC2E}\left(\hat{Y}\right)$ e a alocação da amostra nos estratos $\left(n_h's\right)$ dependem do plano adotado.

8. Considerações Finais

Os resultados dos EPAs estimados mostram que existem diferenças importantes de eficiência entre as estratégias amostrais consideradas. As estratégias que utilizam sorteio de UPAs com probabilidade proporcional ao tamanho apresentaram melhor desempenho para a maioria das variáveis. Além disso a estratégia considerando Amostragem PPT de Pareto (sem reposição) apresentou efeitos do plano amostral geralmente menores que as demais estratégias consideradas.

Os resultados obtidos dão suporte para usuários que pretendem trabalhar com algum dos planos mencionados e até mesmo com planos alternativos, pois oferece suporte na criação de variáveis e descrição da metodologia de cálculos de variância e tamanhos amostrais. O cálculo de tamanhos de amostra em planos amostrais complexos é facilitado quando é utilizado o conceito de efeito do plano amostral.

O conceito do efeito do plano amostral se aplica a outros tipos de estratégias amostrais conglomeradas em dois ou mais estágios e pode ser utilizado para planejamentos amostrais que utilizem pesquisas anteriores como os censos.

Referências bibliográficas

ALBIERI, S. A., BIANCHINI, Z. M. Principais aspectos da amostragem das pesquisas domiciliares do IBGE – Revisão 2002. Texto para discussão número 8. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2002.

ALBIERI, S. A. Unidade de Metodologia e a Evolução do Uso de Amostragem no IBGE. Texto para discussão número 12. Diretoria de Pesquisas, Rio de Janeiro: IBGE, 2003.

ALBIERI, S. A. FREITAS, M. P. S. Censo Demográfico 2010: Plano amostral para coleta dos dados. Revisão 2012. Rio de Janeiro, IBGE, 2012.

BUSSAB, W. O , BOLFARINE, H. Elementos de Amostragem, São Paulo, Ed. Blucher, 2004.

COCHRAN, W.G. Sampling Techniques(3rd ed.). New York: Wiley, 1977.

COSTA, G.T.L. Coordenação de amostras PPT em pesquisas repetidas, utilizando o método de amostragem de Pareto. Dissertação (Mestrado), Rio de Janeiro, Escola Nacional de Ciências Estatísticas, 2007.

FREITAS, M. P. S. et. al. Amostra Mestra para o Sistema Integrado de Pesquisas Domiciliares. Texto para discussão número 23. Diretoria de Pesquisas, Rio de Janeiro: IBGE, 2007.

HAGGARD, E.A. Intraclass correlation and the analysis of variance. New York, Dryden Press, 171p., 1958.

IBGE. Pesquisa Mensal de Emprego – PME, Série Relatórios Metodológicos, vol. 23, Rio de Janeiro, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, 2007a.

IBGE. Sistema Integrado de Pesquisas Domiciliares. Texto para discussão número 24. Diretoria de Pesquisas, Rio de Janeiro, 2007b.

IBGE. Síntese da PNAD 2009, Coordenação de Trabalho e Rendimento, Diretoria de Pesquisas. Rio de Janeiro, 2010. Sinopse do Censo Demográfico, Rio de Janeiro, 2011a.

IBGE. Base de informações do Censo Demográfico 2010: resultados da Sinopse por setor censitário, Rio de Janeiro, 2011b.

IBGE. Base de informações do Censo Demográfico 2010: Resultados do Universo por setor censitário, 127 p., Rio de Janeiro, 2011c.

IBGE. Pesquisa de Orçamentos Familiares 2008-2009, Diretoria de Pesquisas: Análise do Consumo Alimentar Pessoal no Brasil. Rio de Janeiro, 2011d.

IBGE. Pesquisa Nacional por Amostra de Domicílios Contínua. 11º Fórum, Sistema Integrado de Pesquisas Domiciliares. Rio de Janeiro, 2011e. Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/sipd/decimo_primeiro_forum/ implantacao_da_PNAD_continua.pdf. Acesso em 12/11/2012.

IBGE. Censo Demográfico 2010. Notas Metodológicas. Rio de Janeiro, 2012a.

IBGE. Censo Demográfico 2010: Resultados Gerais da Amostra, Rio de Janeiro, 2012b.

KISH , L. Survey sampling. New York: John Wiley & Sons, 1965.

Lila, M. F. Estimação de Variâncias em Pesquisas Amostrais Domiciliares. Dissertação (Mestrado) – Escola Nacional de Ciências Estatísticas, Rio de Janeiro, 2004.

LILA, M. F. e FREITAS, M. P. S. Uma medida de homogeneidade da migração nos municípios brasileiros, Diretoria de Pesquisas, Rio de Janeiro, 2006.

ROSÉN, B. On sampling with probability proportional to size, Journal of Statistical Planning an Inference, nº 62, 159-191, 1997.

Särndal, C. E., Swensson B. & Wretman, J. H. Model assisted survey sampling. Springer-Verlag, Nova York, 1992.

Silva, P. L. Determinação do tamanho da amostra da PNDS-2002. Relatório Técnico, Rio de Janeiro, 2002.

SILVA, P. L. N e BIANCHINI, Z. M. Determinação do tamanho de amostra para pesquisa domiciliar conglomerada em dois estágios considerando o efeito do plano amostral, IBGE, Rio de janeiro, 2002.

SILVA, P. L. N. e MOURA, F. A. S. Efeito de conglomeração da malha setorial do Censo Demográfico 1980, IBGE, Rio de Janeiro, Texto para discussão n.32, 1990.

Abstract

Complex sample designs are used in most of the household sample surveys conducted by the Brazilian Institute of Geography and Statistics. One of the key steps of sample design comprises the sample size determination, possibly for several variables of interest. This paper presents an approach for calculating sample sizes using design effects (DEFF) and intraclass correlation coefficients estimated for four different sampling strategies using data from the census 2010 enumeration area frame. All four sampling strategies comprise two stage cluster sampling, where in the first stage census tracts are the primary sampling units (PSUs) and households are selected in the second stage. The results indicate that PPS designs for sampling the PSUs are generally more efficient (i.e. lead to smaller sample sizes) because they take advantage of the correlation between target variables and the size of the PSUs. Pareto PPS sampling design (without replacement sampling of PSUs) is more efficient than with replacement PPS sampling of PSUs.

Key Words: design effect, Household sample survey, Pareto sampling, sample size

Apêndice

A.1 Amostragem de Pareto

A amostragem de Pareto é um dos métodos de amostragem por ordenação que utiliza probabilidade proporcional ao tamanho (PPT) (Rosén, 1997). Pode ser utilizada em desenhos que necessitem de coordenação de amostras.[1]

Neste método é possível prefixar o tamanho de amostra e estimar variâncias com boas propriedades (Costa, 2007). Para o sorteio das unidades é utilizada a distribuição de Pareto para geração das variáveis de ordenação que auxiliam na seleção da amostra.

A amostragem por ordenação consiste na associação de um número aleatório a cada unidade da população, e em seguida na seleção das unidades associadas aos menores números aleatórios. A Amostragem PPT de Pareto foi considerada para seleção das unidades primárias de amostragem (UPAs).

O método de seleção da Amostragem de Pareto PPT aplicado à seleção de UPAs é descrito nos passos a seguir.

Passos para a seleção de UPAs pelo método PPT de Pareto:

1. Estabelecer o tamanho m da amostra de UPAs a ser retirada da população U, cujas medidas de tamanho são $N_1, N_2, ..., N_M$.
2. Calcular as probabilidades de inclusão nominais $\pi_i = \frac{mN_i}{N}$ para cada uma das UPAs. Caso $m N_i > N$ para algum i, inclua esta unidade com certeza na amostra, recalcule m e N, e refaça os cálculos dos π_i até que todos os valores sejam menores que 1.
3. Gerar, de forma independente, para cada UPA uma determinação de variável aleatória $q_i = \frac{Z_i/(1-Z_i)}{\pi_i/(1-\pi_i)}$ onde $Z_i \sim \text{Unif}(0,1)$.

1 A coordenação de amostras consiste no controle da sobreposição amostral de unidades elementares em períodos consecutivos. Em geral é fixado um percentual que se pretende sobrepor, como por exemplo, na Pesquisa Nacional por Amostra de Domicílios Contínua, onde há sobreposição amostral de 80% dos domicílios em trimestres consecutivos.

4. Ordenar as unidades populacionais em ordem crescente de acordo com as determinações das variáveis de ordenação q_i .
5. Incluir na amostra as m UPAs com os menores valores das variáveis de ordenação q_i .

A.2 Variância do estimador de total para a amostragem conglomerada em dois estágios com seleção PPT de pareto no primeiro estágio e AAS no segundo estágio

Considerando que as probabilidades de inclusão nominais λ_i dos conglomerados no primeiro estágio correspondem às probabilidades de inclusão efetivas π_i temos que $\pi_i = m\frac{N_i}{N}$.

O estimador de Horvitz-Thompson do total sob a estratégia E4 pode ser escrito como:

$$\hat{Y}_{E4} = \sum_{i\in s}\frac{\hat{Y}_i}{\pi_i} = \sum_{i\in U}\delta_i\frac{\hat{Y}_i}{\pi_i} \quad \text{onde} \quad \delta_i = \begin{cases} 1 & \text{se } i\in s \\ 0 & \text{caso contrário} \end{cases}$$

Considerando que a UPA i foi selecionada para a amostra, o estimador de Horvitz-Thompson para o total populacional da UPA i é dado por :

$$\hat{Y}_i = \sum_{j\in a_i}\frac{y_{ij}}{\pi_{j|i}} \qquad \text{onde} \qquad \pi_{j|i} = \frac{n_i}{N_i}$$

Para obter a variância do estimador de total sob a estratégia 4, note que:

$$V\left(\hat{Y}_{E4}\right) = V\left(\sum_{i\in U}\delta_i\frac{\hat{Y}_i}{\pi_i}\right) = E_1\left[V_2\left(\sum_{i\in U}\delta_i\frac{\hat{Y}_i}{\pi_i}|s\right)\right] + V_1\left[E_2\left(\sum_{i\in U}\delta_i\frac{\hat{Y}_i}{\pi_i}|s\right)\right] \text{ (Cochran, 1977)}$$

Onde E_1, V_1 denotam esperança e variância referentes ao sorteio de UPAs no primeiro estágio, e E_2, V_2 denotam esperança e variância referentes ao sorteio de USAs no segundo estágio. Segue-se então que:

$$V_1\left[E_2\left(\sum_{i\in U}\delta_i\frac{\hat{Y}_i}{\pi_i}|s\right)\right] = V_1\left[\sum_{i\in U}\delta_i\frac{1}{\pi_i}E_2\left(\hat{Y}_i\right)\right] = V_1\left[\sum_{i\in U}\delta_i\frac{1}{\pi_i}Y_i\right] = V_1\left[\sum_{i\in s}\frac{Y_i}{\pi_i}\right]$$

Para chegar ao resultado acima utilizamos a propriedade de que a esperança de um estimador Horvitz-Thompson $\left(\hat{Y}_i \quad \text{no} \quad \text{caso}\right)$ é o valor do parâmetro $\left(Y_i\right)$. Esta parcela corresponde à variância para o estimador de um total com seleção de PPT Pareto apresentada em Rosén (1997, eq. 4.2). A outra parcela é obtida a seguir:

$$E_1\left[V_2\left(\sum_{i\in U}\delta_i\frac{\hat{Y}_i}{\pi_i}\mid s\right)\right]=E_1\left[\sum_{i\in U}\delta_i\frac{1}{\pi_i^2}V_2\left(\hat{Y}_i\right)\right]=E_1\left[\sum_{i\in U}\delta_i\frac{1}{\pi_i^2}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2\right]$$
$$=\sum_{i\in U}\frac{N_i^2}{\pi_i}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2$$

Assim temos que a variância do estimador de total sob um plano amostral conglomerado em dois estágios com seleção PPT de Pareto no primeiro estágio e AAS no segundo estágio é a seguinte:

$$V\left(\hat{Y}_{E4}\right)=\frac{M}{M-1}\left\{\sum_{i\in U}\frac{Y_i^2}{\pi_i}\left(1-\pi_i\right)-\frac{\left[\sum_{i\in U}Y_i\left(1-\pi_i\right)\right]^2}{m-\sum_{i\in U}\pi_i^2}\right\}+\sum_{i\in U}\frac{1}{\pi_i}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2$$ que

corresponde a (8).

A.3 Estimação da variância do estimador de total para a amostragem conglomerada em dois estágios com seleção PPT de Pareto no primeiro estágio e AAS no segundo estágio, usando a a amostra do Censo Demográfico 2010

Para estimar a variância $V\left(\hat{Y}_{E4}\right)$ com base na amostra do Censo Demográfico 2010 é necessária a estimação dos parâmetros desconhecidos Y_i, Y_i^2 e S_i^2. Os estimadores para estas quantidades desconhecidas serão descritos a seguir:

Para estimar Y_i, empregou-se o estimador $\hat{Y}_i=\sum_{i\in a_i}\frac{y_{ij}}{\pi_{j|i}}=\frac{N_i}{n_i}\sum_{j\in a_i}y_{ij}$.

Para estimar Y_i^2 note que a variância de uma variável aleatória X é dada por:

$$V(X) = E\left(X^2\right) - [E(X)]^2 \rightarrow E\left(X^2\right) = V(X) + [E(X)]^2$$

Consequentemente para a variável aleatória $\hat{Y}_i$ segue que:

$E\left[\left(\hat{Y}_i^2\right)\right] = V\left[\hat{Y}_i\right] + \left[E\left(\hat{Y}_i\right)\right]^2 = Y_i^2 + V\left(\hat{Y}_i\right)$ o que indica que $\hat{Y}_i^2$ é um estimador viciado para Y_i^2 **.**

Para obter um estimador não viciado (ENV) de Y_i^2 **,** basta subtrair de $\hat{Y}_i^2$ um ENV de $V\left(\hat{Y}_i\right)$ obtendo então o seguinte ENV para Y_i^2 :

$$\hat{Y}_i^2 = \hat{Y}_i^2 - \hat{V}\left(\hat{Y}_i\right).$$

Para estimar S_i^2 **,** utiliza-se:

$$\hat{S}_i^2 = s_i^2 = \frac{1}{(n_i - 1)} \sum_{j \in a_i} \left(y_{ij} - \overline{y}_i\right)^2.$$

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 demográfico e econométrico.

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 a 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 para o site http://rbes.submitcentral.com.br/login.php

Secretaria da RBEs

Revista Brasileira de Estatística – RBEs

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - IBGE

Rua André Cavalcanti, 106, sala 503-A

Centro, Rio de Janeiro – RJ

CEP: 20031-050

Tels.: 55 21 2142-3596 (Marilene Pereira Piau Câmara – Secretária)

55 21 2142-4957 (Pedro Luis do Nascimento Silva – Editor–Executivo)

Fax: 55 21 2142-0501

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais enviados para publicação devem obedecer às normas seguintes:

1. Podem ser submetidos originais processados pelo editor de texto *Word for Windows* ou originais processados em LaTeX (ou equivalente) desde que estes últimos sejam encaminhados e 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, se for o caso de constarem no documento, também devem figurar nesta página;

3. No caso de a 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 a 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 concisos e apropriados. 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 sequencialmente (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;

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;

10. Não serão permitidos itens que identifiquem os autores do artigo dentro do texto, tais como: número de projetos de órgãos de fomento, endereço, *e-mail*, etc. Caso ocorra, a responsabilidade será inteiramente dos autores; e

11. No caso de o artigo ser aceito para a publicação após a avaliação dos pareceristas, serão encaminhadas as sugestões/comentários aos autores sem a sua identificação. Uma vez nesta condição, é de responsabilidade única dos autores fazer o *download* da formatação padrão da revista (em doc ou em LaTeX) para o envio da versão corrigida.