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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 71 número 234 janeiro/dezembro 2010

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 71, n. 234, p. 1-151, jan./dez. 2010

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

© IBGE. 2011

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

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

Editor Responsável
Francisco Louzada (USP - São Carlos)

Editor-Executivo
Pedro Luis do Nascimento Silva (ENCE/IBGE)

Editor de Metodologias
Fernando Moura (UFRJ)

Editor de Estatísticas Oficiais
Denise Britz do Nascimento Silva (DPE/IBGE)

Editores Associados
Dalton Francisco de Andrade (UFSC)
José André de Moura Brito (ENCE/IBGE)
Viviana Giampaoli (IME-USP)
Beatriz Vaz de Melo Mendes (UFRJ)
Thelma Sáfadi (UFLA)
Paulo Justiniano Ribeiro Junior (UFP)
Josmar Mazucheli (UEM)
Luis A Milan (UFSCar)
Cristiano Ferraz (UFPE)
Gleici Castro Perdoná (FMRP-USP)
Ana Maria Nogales Vasconcelos (UNB)
Ronaldo Dias (UNICAMP)
Mário de Castro (ICMC-USP)
Nuno Duarte Bittencourt (ENCE/IBGE)
Solange Trindade Corrêa (DPE/IBGE)

Editoração
Sandra Cavalcanti de Barros - ENCE/IBGE
Arnoldo Furtado de Sá - ENCE/IBGE
Letícia Baptista de Paula Barros – ENCE/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2011.

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

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

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

Trimestral (1940-1986), semestral (1987-).
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).
Co-edição com a Associação Brasileira de Estatística a partir do v.58.
ISSN 0034-7175 = Revista brasileira de estatística.

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

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

Impresso no Brasil/*Printed in Brazil*

Nota do Editor

Este volume da RBEs do ano de 2010 é composto por cinco artigos. O primeiro artigo, de autoria de Francisco Louzada, Anderson Ara, Cleyton Z. Oliveira e Cláudio V. Gonçalves, apresenta um diagnóstico do ensino da ciência estatística nas universidades públicas brasileiras. O segundo artigo, de autoria de Flávio H. M. A. Freire, Maria H. C. Spyrides, Moisés A. C Aguirre e Kátia L. Souza, apresenta um diagnóstico da dinâmica matrimonial no nordeste do Brasil. O terceiro artigo, de autoria de José A. M. Brito, Luiz S. Ochi, Luciana R. Brito e Flávia M. T. Montenegro, apresenta um algoritmo para o agrupamento baseado em K-Medoids. O quarto artigo, de autoria de Rejane C. Rocha e Thelma Sáfadi, analisa dos dados de doação de sangue do Núcleo Regional de São João Del Rei da Fundação Homominas. O quinto artigo, de autoria de Sumala A. Latif e Pedro A. Morettin apresenta uma introdução a cópulas com aplicações na avaliação do desempenho de empresas.

Aproveito a oportunidade para agradecer a colaboração de Pedro Luis do Nascimento Silva (Editor Executivo), todos os Editores Associados, revisores do periódico, autores, IBGE e ABE.

Uma excelente leitura.

Francisco Louzada
Editor Responsável

Sumário

Diagnóstico do Ensino da Estatística nas Universidades Públicas Brasileiras: Uma Descrição de Algumas das Dimensões que Compõem o Perfil do seu Corpo Docente

Francisco Louzada[1]
Anderson Ara[2]
Cleyton Z. Oliveira[2]
Claudio V. Gonçalves[2]

Resumo

Contexto: O ensino da ciência estatística é obrigatório em praticamente todos os cursos de graduação das universidades brasileiras. Além disso, vários são cursos de Graduação em Estatística, distribuídos pelas várias universidades nacionais. Entretanto, apesar da importância desta ciência, não existe na literatura nacional estudos sistemáticos direcionados à caracterização dos docentes responsáveis pelo ensino da ciência estatística no país.

Objetivo: Neste contexto, apresentamos neste artigo uma descrição de tais docentes, particularmente, no que tange aos cursos de Graduação em Estatística.

Método: Esta descrição foi realizada por meio de um levantamento descritivo, relacionado aos aspectos de sua formação e produção científica. E finalizada com a apresentação da previsão de demanda de Doutores em Estatística necessários para suprir as vagas em aberto a partir das ocorrências das aposentadorias dos docentes das Graduações em Estatística.

Coleta de Informação: Considerando a listagem dos cursos de graduação em estatística de universidades públicas reconhecidos pelo MEC, pesquisamos os currículos Lattes disponíveis dos 648 docentes vinculados a essas instituições durante a segunda quinzena de julho de 2009.

Conclusões alcançadas: O levantamento revela que, de uma forma geral, os docentes vinculados às Graduações em Estatística, têm, em sua maioria, mestrado e/ou doutorado. No

[1] Universidade de São Paulo [2] Universidade Federal de São Carlos

entanto, considerando graduação, mestrado e doutorado, a proporção com formação formal em estatística como área de concentração não é grande, metade dos docentes possui Graduação em Estatística, aproximadamente 1/3 dos mestres são mestres em Estatística e 1/3 dos doutores são doutores em Estatística. Em termos da produção científica nos anos 2006 a 2008, praticamente, a cada 10 docentes, 4 não publicaram, 4 publicaram entre 1 e 4 artigos e 2 publicaram mais do que 5 artigos completos no período. O estudo também revela que existem grandes diferenças entre as regiões geográficas do país, bem como entre os docentes com diferentes tempos de docência e contemplados com bolsa de produtividade científica do CNPq. Como esperado, os docentes que são bolsistas de produtividade possuem maior de produção científica, em média, em comparação com os docentes não bolsistas. Além disso, existe uma interação positiva entre tempo de docência e bolsa produtividade, com um aumento gradativo no ritmo da produção científica dos docentes com o aumento do tempo de docência, o qual é intensificado pela presença de bolsa de produtividade. Também detectamos que existe déficit de Doutores em Estatística necessários para suprir as vagas em aberto a partir das ocorrências das aposentadorias dos docentes das 24 Graduações em Estatística estudadas. Considerando um cenário pessimista, até o ano de 2017 teríamos ainda déficit de doutores para preencher as vagas de docentes. Em última instância, assumindo que 20% dos doutores egressos são absorvidos pelo mercado a cada ano ainda em 2030 o déficit de doutores deverá perdurar.

Palavras Chave: Perfil do Profissional Docente; Graduação em Estatística.

1. Introdução

A ciência estatística tem sido contribuinte significativa em praticamente todas as áreas do conhecimento, dentre as quais podemos citar desde áreas básicas como a física, a química e a biologia, às mais tecnológicas, como as engenharias, passando, pela agronomia, astronomia, criminologia, demografia, saúde e psicologia. Consequentemente, todos os setores de atividades, primário, secundário e terciário, têm se beneficiado dos avanços desta ciência, inclusive, a administração pública. Assim, dada sua natureza de ciência do significado e do uso dos dados, nenhuma disciplina tem interagido tanto com as demais disciplinas em suas atividades do que a ciência estatística (Memória, 2004), sendo o ensino desta ciência, atualmente, obrigatório em quase todos os cursos de graduação das mais diversas áreas, com pouquíssimas exceções (Lopes, 1998).

No Brasil, a ciência estatística tem sua história associada ao Instituto Brasileiro de Geografia e Estatística (IBGE), cujas raízes foram fincadas ainda durante o Império. Sendo que, segundo Araujo (1993) e Dantas (2002), o primeiro curso de ¨Inferência Estatística¨ foi ministrado em 1947, baseado no livro de Harald Cramer "Mathematical Methods of Statistics". Porém, somente em 1953 duas escolas iniciaram o ensino da ciência estatística no Brasil: a Escola Nacional de Ciências Estatísticas - ENCE, criada pelo IBGE e a Escola de Estatística da Bahia, fundada e mantida pela Fundação Visconde de Cairú (Silva, 1989; IBGE,1987).

Nas décadas seguintes, dado o avanço da ciência e a pluralização da educação, vários cursos de Graduação em Estatística foram criados em todo o país. Atualmente, de acordo com consulta realizada em 31/10/2009 (Fonte: http://emec.mec.gov.br/), existem cursos de Graduação em Bacharelado e Licenciatura em Estatística, reconhecidos pelo Ministério da Educação (MEC), distribuídos por 33 universidades (30 universidades públicas e 3 privadas) espalhadas por todo o território nacional com um total de mais de 600 docentes diretamente envolvidos no ensino desta ciência.

Neste contexto, uma questão de importância consiste em caracterizar os docentes responsáveis pelo ensino da ciência estatística no Brasil no que se relaciona aos aspectos de sua formação e atividades inerentes a sua profissão. Particularmente, esta

caracterização é fundamental para os cursos de Graduação em Estatística, núcleos básicos para a formação dos profissionais estatísticos, sendo que estes devem ser formados com o perfil exigido para a sua atuação, em instituições privadas e governamentais, bem como para a sua capacitação continuada e possível atuação em instituições de ensino superior. Este perfil evidentemente é balizado pelas necessidades do mercado de trabalho e pelas características formativas e profissionais dos seus formadores.

Entretanto, apesar da importância do tema, depois de uma busca exaustiva não encontramos na literatura estudos sistemáticos voltados à caracterização destes profissionais. Este é o objetivo principal deste artigo: caracterizar os docentes responsáveis pelo ensino dentro das Graduações em Estatística do país, segundo algumas das dimensões que compõem o perfil desta classe de profissionais. Particularmente, trata-se de um levantamento descritivo, com ênfase dada ao tempo de docência do docente e à sua formação no que se relaciona às características de sua graduação, mestrado e doutorado. Também focamos a produção científica docente, representada aqui pela quantidade de artigos completos publicados no período de 2006 a 2008, sendo esta, possivelmente, dentre todas as atividades relacionadas à sua profissão, a responsável pela manutenção do princípio da indissociabilidade acadêmica quanto ao binômio pesquisa-ensino. As atividades de extensão, pertinentes a este contexto, não foram consideradas no presente estudo por falta deste tipo de informação na base de dados utilizada.

Nosso interesse é apresentar subsídios para que possam ser respondidas, dentre outras, questões como: Quantos são os docentes responsáveis pelo ensino dentro das graduações em estatística do país? Esta quantidade é proporcionalmente balanceada entre as regiões geográficas? Quais as porcentagens de homens e mulheres? Qual é o tempo médio de docência? Quais as proporções de docentes com doutorado, mestrado e somente graduação? As mesmas proporções se apresentam nas diferentes regiões? Como se dá a formação destes docentes em termos de área de concentração em estatística? Qual é a produção científica dos docentes? Como se comporta a distribuição desta produção nas cinco regiões geográficas do país? Existem docentes que se distanciam dos demais quanto a sua produção científica? Qual a relação entre tempo de docência e produção científica? Quantos são os bolsistas do CNPq? Em quais

regiões estão alocados? Estes apresentam produção científica diferenciada? Quantos doutores serão necessários para repor as vagas em aberto a partir das ocorrências das aposentadorias dos docentes das Graduações em Estatística estudadas,

Muitas interpretações e inferências lógicas, gerais e específicas podem ser tecidas a partir do levantamento aqui apresentado. Entretanto, apesar de termos consciência do fato de que a simples escolha de certas variáveis a serem estudadas, com categorizações específicas, já ser indicativo da predileção dos autores por algumas visões referentes à caracterização dos docentes aqui estudados em detrimento de outras, e, mesmo que indiretamente, propiciar a postulação de conjecturas específicas, dentro do possível, preferimos omitir nossas percepções advindas da análise. Assim, somente apresentamos os resultados de forma descritiva, acreditando na importância de isenção de opinião e de concorrência entre interpretações quando da realização que uma pesquisa envolvendo a caracterização de uma classe de profissionais, e que os leitores se encarregarão desta importante tarefa que poderá ser instrumentada de forma adequada pelo levantamento aqui apresentado.

O artigo é organizado como segue. A Seção 2 apresenta as especificações gerais da metodologia, descrevendo quais os docentes que foram incluídos no levantamento, a forma de coleta dos dados e as variáveis estudadas. A Seção 3 apresenta os resultados obtidos, incluindo uma descrição geral, seguida de descrições por região geográfica, tempo de docência, bolsa de produtividade e a análise conjunta de tempo de docência e bolsa de produtividade. A Seção 4 apresenta alguns comentários finais, resumindo os resultados obtidos. Também, visualizando o planejamento estratégico das Graduações em Estatística, esta seção apresenta uma previsão de demanda de Doutores em Estatística necessários para suprir as vagas em aberto a partir das ocorrências das aposentadorias dos docentes das Graduações em Estatística estudadas.

2. Metodologia

Para este estudo consideramos docentes necessariamente vinculados aos Departamentos de Estatística do país que oferecem curso de Graduação em Estatística. Uma pesquisa on-line foi realizada para identificação dos departamentos que deveriam

ser considerados, tendo como referência a listagem dos Departamentos de Estatística fornecida pelo website www.redeabe.org.br, portal da Associação Brasileira de Estatística (ABE).

A home page de cada um dos Departamentos de Estatística que oferecem cursos de Graduação em Estatística foi acessada durante a segunda quinzena do mês de julho de 2009 para obtenção do nome dos docentes. A home page de alguns departamentos não foi encontrada. Neste caso obtivemos os nomes dos respectivos docentes através de e-mails enviados às instituições.

Dentre as 33 universidades que possuem Departamento de Estatística oferecendo cursos de Graduação em Estatística credenciados pelo MEC até o mês de dezembro de 2008, somente foram considerados os docentes vinculados aos Departamentos de Estatística de instituições públicas de ensino com cursos de Graduação em Estatística credenciados pelo MEC, com início das atividades anterior ao ano de 2008. Assim, não foram consideradas as seguintes universidades públicas: Universidade Federal de Goiás (UFG), Universidade Federal de Rondônia (UNIR), Universidade Federal do Mato Grosso (UFMT), Universidade Federal de Ouro Preto (UFOP), Universidade Federal de Santa Maria (UFSM) e Universidade Federal do Piauí (UFPI) devido ao fato destas terem iniciado seus cursos de Graduação em Estatística em 2008 ou após este ano. Também, não consideramos na análise os docentes vinculados às 3 universidades particulares: Escola Superior de Estatística da Bahia (ESEB), Universidade Salgado de Oliveira (UNIVERSO) e Centro Universitário Capital (UNICAPITAL). Desta forma, neste trabalho, foram pesquisados os docentes de 24 dentre os 30 Departamentos de Estatística de universidades públicas que oferecem os cursos de Graduação em Estatística reconhecidos pelo MEC, o que corresponde a 80% do total dos cursos de Graduação em Estatística oferecidos por instituições públicas do país.

Os departamentos pesquisados, bem como suas respectivas fontes de informação, são apresentados na Tabela 1. A Figura 1 apresenta a distribuição geográfica das Graduações em Estatística reconhecidas pelo MEC e com início anterior a 2008. Dos cursos analisados, 45,8% encontram-se na região Sudeste, 12,5% na região Sul, 21,2% na região Nordeste, 4,2% na região Centro-Oeste e 8,3% na região Norte. A Tabela A.1, em anexo, apresenta as características gerais das Graduações em Estatística participantes do estudo.

2.1. Fonte de informações

Com os nomes dos docentes a serem incluídos no estudo em mãos, durante a segunda quinzena de julho de 2009, pesquisamos os seus currículos Lattes. A Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) consiste em uma ferramenta largamente conhecida entre os acadêmicos e significativa para o conhecimento da trajetória de docentes e pesquisadores de todas as áreas do conhecimento. Sua importância se estende às atividades operacionais de fomento à pesquisa em nível nacional. Ela está disponível no site lattes.cnpq.br.

Tabela 1. Departamentos de Estatística pesquisados organizados por estado.

ESTADO	DEPARTAMENTO	FONTE DE CONSULTA
Amazonas	Instituto de Ciências Exatas – UFAM	http://portal.ufam.edu.br/index.php/unidades-academicas/17-instituto-de-ciencias-exatas
Bahia	Departamento de Estatística – UFBA	http://www.est.ufba.br/
Brasília	Departamento de Estatística – UnB	http://e-groups.unb.br/ie/est/
Ceará	Departamento de Estatística e Matemática Aplicada – UFC	http://www.estatistica.ufc.br/
Espírito Santo	Departamento de Estatística – UFES	solicitação via e-mail
Minas Gerais	Departamento de Estatística – UFMG	http://www.est.ufmg.br/portal/
	Departamento de Estatística – UFJF	http://www.ufjf.br/estatistica/
Pará	Departamento de Estatística – UFPA	http://www.ufpa.br/ccen/est/
Paraíba	Departamento de Estatística – UEPB	e-mail enviado à coordenação
	Departamento de Estatística – UFPB	http://www.de.ufpb.br/
Paraná	Departamento de Estatística – UEM	http://www.des.uem.br
	Departamento de Estatística – UFPR	http://www.est.ufpr.br/
Pernambuco	Departamento de Estatística – UFPE	http://www.de.ufpe.br/
Rio Grande do Norte	Departamento de Estatística - UFRN	http://www.estatistica.ccet.ufrn.br/
Rio Grande do Sul	Departamento de Estatística – UFRGS	http://www.mat.ufrgs.br/
Rio de Janeiro	Escola Nacional de Ciências Estatísticas – ENCE	http://www.ence.ibge.gov.br/
	Departamento de Estatística – UFF	http://www.proac.uff.br/depestatistica/
	Departamento de Métodos Estatísticos- UFRJ	http://www.dme.im.ufrj.br/
	Instituto de Matemática e Estatística – UERJ	http://www.ime.uerj.br/
Sergipe	Departamento de Estatística – UFS	solicitação via e-mail
São Paulo	Instituto de Matemática, Estatística e Computação Científica – UNICAMP	http://www.ime.unicamp.br/
	Departamento de Estatística – UFSCar	http://www.ufscar.br/~des
	Departamento de Estatística – USP Campus São Paulo	http://www.ime.usp.br/
	Departamento de Matemática, Estatística e Computação - UNESP Campus Presidente Prudente	http://www.fct.unesp.br/departamentos/mat_est_comp/index.php?menu_esq1=2

Como premissa básica, assumimos como verdadeira toda informação disponível nos currículos Lattes. Entretanto, temos consciência da possível existência de inconsistências entre o currículo Lattes e a verdadeira situação acadêmica do docente, muitas vezes ocasionadas pelo dinamismo presente em bases curriculares que podem ser permanentemente alteradas na Internet pelo usuário, que é o caso da base de dados aqui considerada. Este fato pode causar vícios e possíveis inconsistências na análise. Como providência básica, realizamos em nosso banco de dados três revisões sistemáticas de todas as informações coletadas para validação dos mesmos.

De acordo com os departamentos pesquisados (ver Tabela A.1), existem 648 docentes atuando no ensino da Estatística no Brasil nos cursos de Graduação em Estatística com início de suas atividades antes do ano de 2008 (ver Tabela 1). Neste contexto, foram realizadas 648 buscas aos currículos Lattes desses docentes.

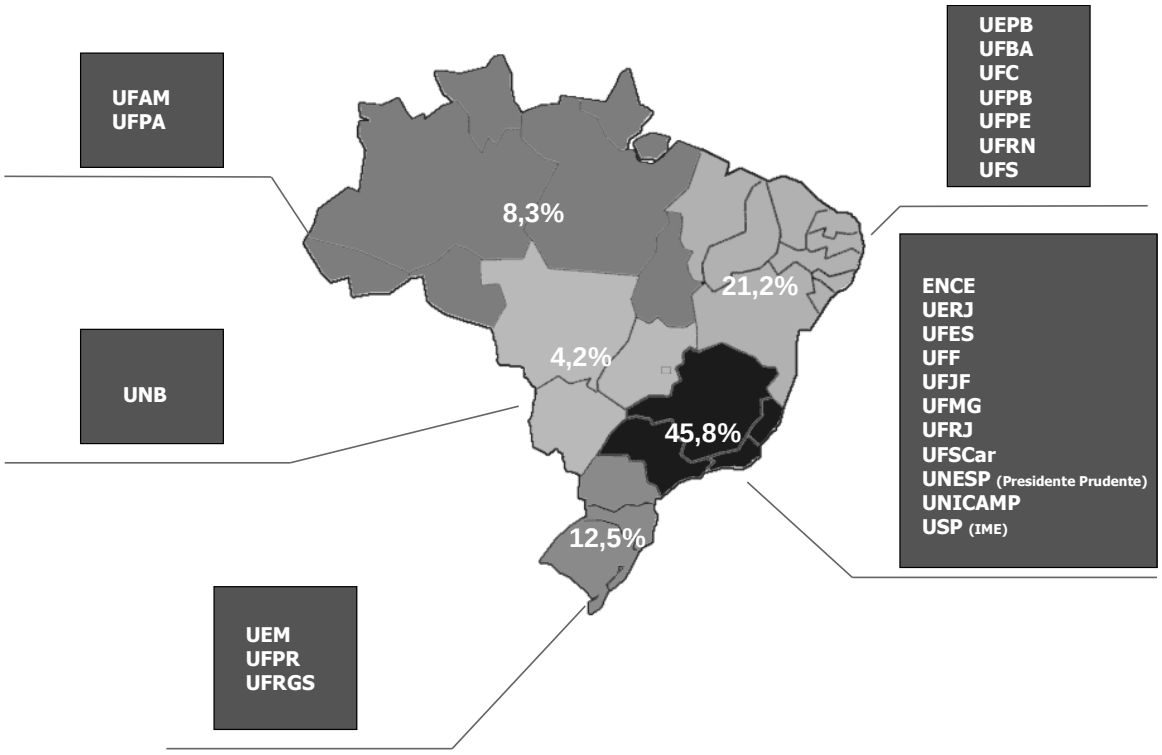


Figura 1. Distribuição dos Departamentos de Estatística pesquisados por região geográfica.

2.2. Variáveis Estudadas

A Tabela 2 apresenta as variáveis que foram abordadas no estudo. Consideramos os mestrados e doutorados nacionais em estatística registrados no portal da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), disponível em http://www.capes.gov.br. Deste modo, consideramos mestre e doutores em estatística docentes que se titularam pelos programas de pós-graduação, reconhecidos pela CAPES, como sendo de probabilidade e estatística, conforme mostra a Tabela 3. Avaliamos os mestrados e doutorados em estatística realizados no exterior considerando apenas a declaração do docente exibida no currículo Lattes, uma vez que estes programas fornecem, geralmente, titulação referenciada por ¨Mestre ou Doutor em Filosofia¨. Assim, se o docente informa que seu doutorado no exterior foi em estatística, consideramos esta informação. Com relação à variável BOLSISTA, levamos em consideração somente a bolsa de Produtividade em Pesquisa - PQ, fornecida pelo Conselho Nacional de Desenvolvimento Científico de Tecnológico (CNPq), uma vez que esta possui importantes critérios para sua concessão, tais como: produção científica do candidato, número de formação de recursos humanos em nível de Pós-Graduação, contribuição científica, tecnológica e em inovação, coordenação ou participação principal em projetos de pesquisa, participação em atividades editoriais, de gestão científica e administração de instituições e núcleos de excelência científica e tecnológica. Entretanto observamos, com pequena frequência, docentes que possuem outros tipos de bolsas: Bolsa de Apoio Técnico a Pesquisa do CNPq - Nível 1 (2 docentes), Bolsa de Desenvolvimento Tecnológico Industrial do CNPq - Nível 1 (2 docentes), e Bolsa de Produtividade Desenvolvimento Tecnológico e Extensão Inovadora do CNPq - Nível 2 (1 docente), as quais, foram desconsideradas na análise.

Tabela 2. Informações coletadas através da pesquisa de currículos acadêmicos.

DESCRIÇÃO	VARIÁVEL	CODIFICAÇÃO
Se o docente possui, ou não, currículo lattes	PLATTES	{0}=não; {1}=sim
Sexo do docente	SEXO	{Masc}; {Fem}
Se o docente possui, ou não, graduação em estatística	GRAD.EST	{0}=não; {1}=sim
Se o docente possui ou não mestrado	MEST	{0}=não; {1}=sim
Caso mestre, Se o Mestrado é em Estatística	MEST.EST	{0}=não; {1}=sim
Se o docente possui, ou não, doutorado	DOC	{0}=não; {1}=sim
Caso Doutor, Se o Doutorado é em Estatística	DOC.EST	{0}=não; {1}=sim
Caso doutor, se o doutorado foi realizado no Exterior	DOC.EXT	{0}=não; {1}=sim
Tempo de docência, em anos (2009 – Data do primeiro vínculo +1)	TEMPO.D	{0,1,2 ...}
Número de artigos completos publicados em 2006	AC.2006	{0,1,2 ...}
Número de artigos completos publicados em 2007	AC.2007	{0,1,2 ...}
Número de artigos completos publicados em 2008	AC.2008	{0,1,2 ...}
Número de artigos completos de 2006 a 2008	AC.TOTAL	{0} = 0; {1} = 1 ou 2; {2} = 3 ou 4; {3} = de 5 a 10; {4} = mais que 10
Se o docente possui bolsa, ou não, de Produtividade em Pesquisa fornecida pelo CNPq	BOLSISTA	{0}=não; {1}=sim

Tabela 3. Mestrados e doutorados em Estatística nacionais considerados.

PROGRAMA	UNIVERSIDADE	MESTRADO	DOUTORADO
Estatística	UNB	x	
Estatística	UFMG	x	x
Estatística	UFPE	x	x
Estatística	UFRJ	x	x
Estatística	UFSCar	x	x
Estatística	USP	x	x
Estatística	UNICAMP	x	x
Matemática Aplicada e Estatística	UFRN	x	

3. Resultados

Dos 648 docentes pesquisados, encontramos 561 (87%) currículos Lattes, dos quais apresentamos uma descrição geral das características dos docentes de acordo com as variáveis coletadas, seguida das descrições por Região Geográfica, Tempo de Docência, Bolsa de Produtividade em Pesquisa, e uma descrição cruzada por Tempo de Docência e Bolsa de Produtividade em Pesquisa. Devemos ressaltar que mesmo nos

currículos encontrados, alguns docentes não possuíam todas as informações consideradas neste artigo.

3.1. Descrição Geral

A Tabela 4 apresenta a descrição geral das informações coletadas. Observamos que 87% dos docentes possuem currículo Lattes, 63% são do sexo masculino, 54% não se graduaram em estatística, 91% concluíram o mestrado, dos docentes que concluíram o mestrado, 38% deles realizaram em Estatística, 70% dos docentes possuem doutorado, 25% destes realizaram seu doutorado no exterior. Também, podemos notar que existem docentes que iniciaram suas atividades letivas com vínculo há quase meio século (49 anos), bem como docentes que iniciaram suas atividades letivas em 2009, sendo que o tempo médio de docência é de aproximadamente 18 anos. A distribuição do número de artigos completos publicados nos 3 anos considerados é assimétrica a direita, sendo que a maior frequência se encontra em nenhum artigo publicado (42%), seguida por: 1 ou 2 artigos (30%), 3 a 4 artigos (10%), 5 a 10 artigos (12%) e mais do que 10 artigos (6%) publicados. Ainda, o número de artigos completos publicados anualmente, em média, é semelhante em todos os anos analisados, com uma pequena variação, de 0,7 e 1,0 artigos. Porém de 2006 a 2008, notamos que existe um leve aumento na média de publicação, bem como na amplitude. Além disso, 11% dos docentes possuem bolsa de produtividade científica.

Tabela 4. Descrição geral das Informações coletadas.

VARIÁVEL	RESULTADO (n=648)
PLATTES: sim(%) / não(%)	87 / 13
SEXO: masculino(%) / feminino(%)	63 / 37
GRAD.EST: sim(%) / não(%)	46 / 54
MEST: sim(%) / não(%)	91 / 9
MEST.EST: sim(%) / não(%)	38 / 62
DOC: sim(%) / não(%)	70 / 30
DOC.EST: sim(%) / não(%)	32 / 68
DOC.EXT: sim(%) / não(%)	25 / 75
TEMPO.D (em anos): Média; DP [Min, Max]	18,4; 10,3 [1, 49]
AC.2006: Média; DP [Min, Max]	0,9; 1,7 [0, 12]
AC.2007: Média; DP [Min, Max]	0,9; 1,6 [0, 13]
AC.2008: Média; DP [Min, Max]	1,0; 2,0 [0, 16]
AC.TOTAL: 0(%) / 1-2(%) / 3-4 (%)/ 5-10(%) / +10(%)	42 / 30 / 10 / 12 / 6
BOLSISTA: sim(%)/não(%)	11 / 89

DP= desvio padrão. As demais siglas das variáveis são listadas na Tabela 3.

3.2. Descrição Por Região Geográfica

A Tabela 5 apresenta os resultados do estudo por região geográfica. Notamos que na região Norte existe uma alta proporção de docentes que possuem Graduação em Estatística (82%). Ainda, 42% dos docentes têm doutorado e destes 27% possuem Doutorado em Estatística, todos realizados no Brasil. A distribuição do número de artigos completos publicados de 2006 a 2008 é visualmente assimétrica, com 56% dos docentes que não publicaram no período, 25% que publicaram 1 ou 2 artigos e 14% que publicaram entre 3 e 4 artigos e 5% que publicaram entre 5 e 10 artigos. O número de artigos completos produzidos anualmente, em média, decresceu de 0,5 artigo em 2006 para 0,2 em 2008. Nesta região, não existem docentes contemplados com bolsa de produtividade científica.

Na região Nordeste, observamos que 16% dos docentes não possuem currículo Lattes, Ainda, 89% dos docentes detém o título de mestre, dentre eles 49% possuem Mestrado em Estatística. Do mesmo modo, temos que 58% dos docentes possuem o título de doutor, 31% destes possuem Doutorado em Estatística. Notamos, também, que 29% dos doutorados foram realizados no exterior. A distribuição de artigos

completos de 2006 a 2008 é visualmente assimétrica e com 51% dos docentes que não publicaram no período, 29% que publicaram 1 ou 2 artigos, 10% que publicaram entre 3 e 4 artigos, 8% que publicaram entre 5 e 10 artigos e 2% que publicaram mais de 10 artigos no período estudado. A média de artigos publicados anualmente variou entre 0,5 e 0,7, com um máximo de 11 de artigos publicados no ano de 2008. Nesta região, 8% dos docentes possuem bolsa de produtividade científica.

Com relação à região Centro-Oeste, no presente estudo sendo representada por uma única universidade (UnB), observamos que metade dos docentes possui Graduação em Estatística (53%) e todos possuem o título de mestre, sendo que 56% destes possuem Mestrado em Estatística. Ainda, 86% dos docentes possuem o título de doutor, destes 44% o realizaram em estatística e 67% foram realizados no exterior. Além disso, o tempo médio de docência é de 24,6 anos, sendo este o maior dentre as cinco regiões geográficas. As maiores frequências de publicação se encontram entre 0 artigos (33%) e 1 ou 2 artigos publicados no período (33%).Temos ainda que 19% dos docentes publicaram entre 5 e 10 artigos e que 5% publicaram mais do que 10 artigos no período. A quantia de 5% dos docentes possui bolsa de produtividade científica.

Na região Sudeste, os resultados obtidos são similares aos resultados gerais apresentados na Tabela 4, exceto pelo fato de que esta região possui um baixo grau de docentes com titulações em estatística (graduação, mestrado ou doutorado). Além disso, apresenta uma proporção ligeiramente acima da proporção nacional de docentes com bolsa de produtividade científica (13%).

Na região Sul, somente 3% dos docentes não possui currículo Lattes, 27% dos docentes são graduados em estatística, sendo esta a região apresenta a menor proporção de graduados em estatística entre as regiões geográficas. Além disso, 87% possuem o título de mestre, mas apenas 24% dos mestrados são em estatística. Ainda, temos que 74% dos docentes possuem doutorado, destes somente 8% realizados em estatística e 15% foram realizados no exterior. A distribuição do número de artigos completos publicados no período estudo é a única dentre todas as regiões geográficas que não apresenta assimetria visual acentuada, com uma alta proporção de docentes que publicaram mais de 5 artigos no período (39%). O número médio de artigos publicados anualmente está próximo de 2 artigos. Comparando a região Sul com as

demais regiões do país, temos que esta região apresenta a maior produção científica média.

Com relação ao tempo de docência, as regiões Sudeste e Sul possuem seus tempos médios próximos da média nacional (19 anos), enquanto as regiões Norte e Nordeste apresentam tempos médios de docência próximos de 15 anos e a região Centro-Oeste tempo médio próximo de 25 anos. Para todas as regiões, os tempos máximos de docência variam, aproximadamente, de 40 a 50 anos.

Graficamente, algumas informações relevantes e contidas na Tabela 5 são exibidas nas Figuras 2, 3 e 4.

Tabela 5. Descrição de todas as Informações coletadas por Região.

VARIÁVEL	REGIÕES DO BRASIL				
	NORTE (n=40)	NORDESTE (n=141)	CENTRO-OESTE (n=24)	SUDESTE (n=372)	SUL (n=70)
PLATTES: sim(%)/não(%)	90 / 10	84 / 16	88 / 12	85 / 15	97 / 3
SEXO: masc.(%)/ fem.(%)	55 / 45	60 / 40	62 / 38	65 / 35	57 / 43
GRAD.EST: sim(%)/não(%)	82 / 18	70 / 30	53 / 47	37 / 63	27 / 73
MEST: sim(%)/não(%)	69 / 31	89 / 11	100 / 0	95 / 5	87 / 13
MEST.EST: sim(%)/não(%)	48 / 52	49 / 51	56 / 44	35 / 65	24 / 76
DOC: sim(%)/não(%)	42 / 58	58 / 42	86 / 14	77 / 23	74 / 26
DOC.EST: sim(%)/não(%)	27 / 73	31 / 69	44 / 56	37 / 63	08 / 92
DOC.EXT: sim(%)/não(%)	0 / 100	29 / 71	67 / 33	24 / 76	15 / 85
TEMPO.D (em anos): Média; DP [Min, Max]	14,5; 9,9 [1, 44]	15,2; 10,2 [1, 40]	24,6; 10,7 [7, 49]	19,6; 10,1 [1, 45]	19; 9,1 [1, 40]
AC.2006: Média; DP [Min, Max]	0,5; 0,9 [0, 3]	0,5; 1,0 [0, 7]	0,9; 1,8 [0, 7]	0,8; 1,7 [0, 11]	1,8; 2,6 [0, 12]
AC.2007: Média; DP [Min, Max]	0,3; 0,6 [0, 2]	0,5; 1,2 [0, 10]	1,6; 2,5 [0, 9]	0,8; 1,5 [0, 10]	1,8; 2,5 [0, 13]
AC.2008: Média; DP [Min, Max]	0,2; 0,6 [0, 3]	0,7; 1,7 [0, 11]	0,9; 1,5 [0, 6]	0,9; 1,7 [0, 10]	2,3; 3,6 [0, 16]
AC.TOTAL(%): 0 1-2 3-4 5-10 +10	56 25 14 5 0	51 29 10 8 2	33 33 10 19 5	40 32 10 13 5	29 25 7 18 21
BOLSISTA: sim(%)/não(%)	0 / 100	8 / 92	5 / 95	13 / 87	13 / 87

A Figura 2 resume as proporções referentes aos títulos acadêmicos de mestrado e doutorado e de bolsistas de produtividade científica, bem como indica o número de docentes por região. A região Sudeste apresenta o maior número de docentes, enquanto a região Centro-Oeste o menor, possivelmente por estar sendo avaliada por meio de apenas uma única universidade. Ainda, notamos que as regiões Sudeste e Centro-Oeste

possuem alta frequência de mestres e doutores. A região Norte, comparada às demais regiões, possui frequência reduzida de mestres, doutores e bolsistas como docentes. A região Nordeste possui, também, uma baixa frequência de doutores quando comparada às demais regiões.

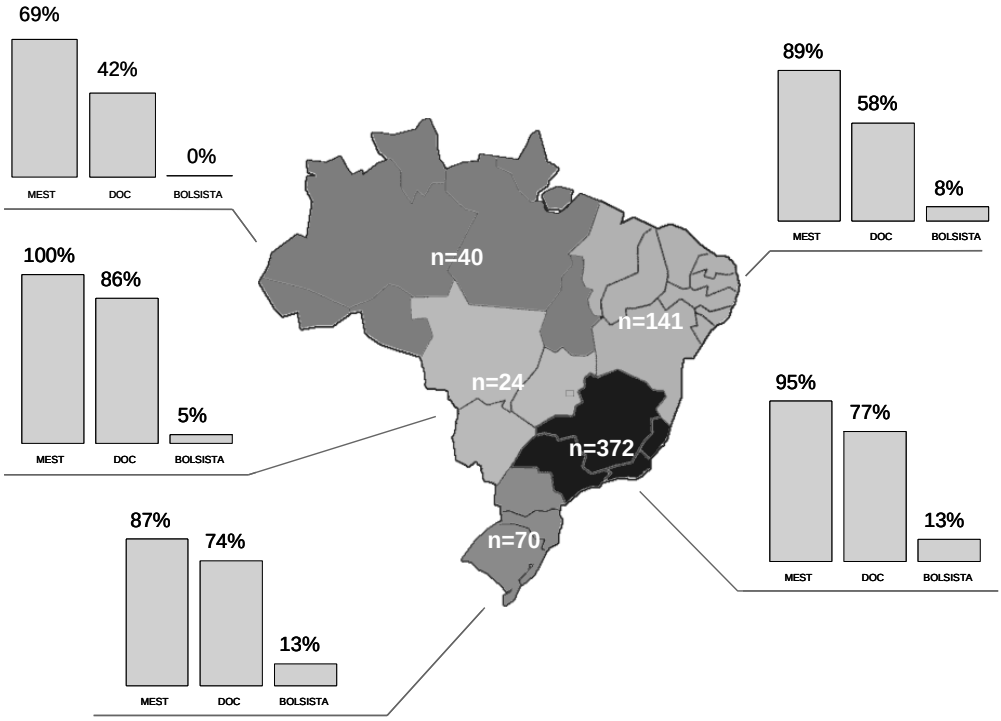


Figura 2. Titulação e bolsas de produtividade em pesquisa dos docentes por região do país.

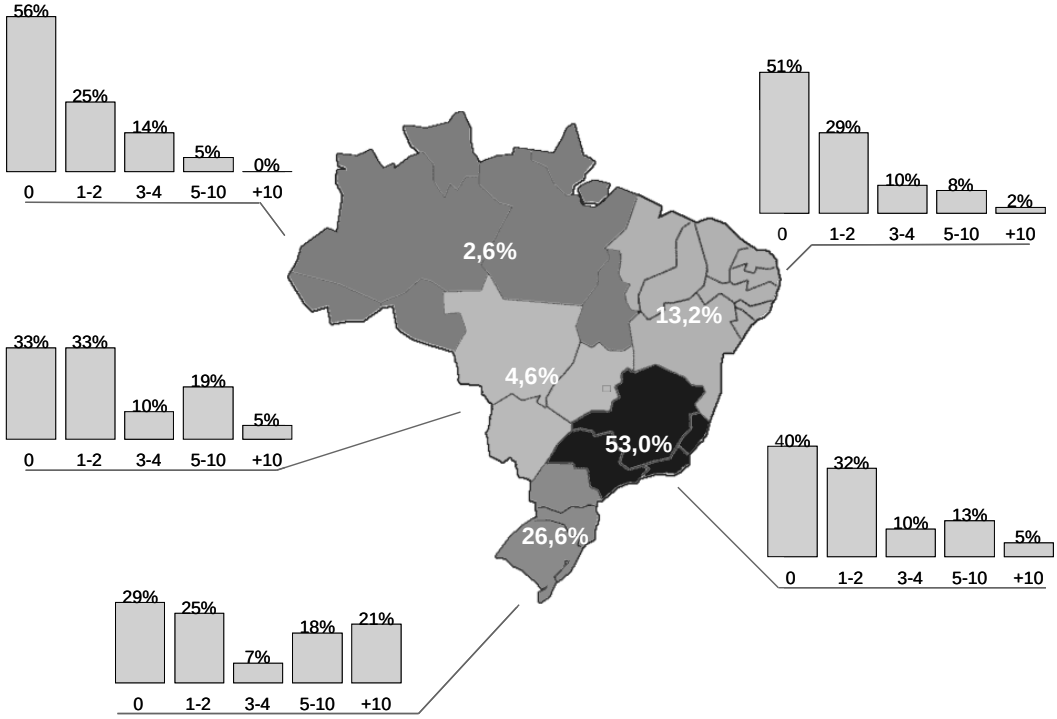


Figura 3. Número de artigos completos publicados entre 2006 e 2008 e contribuição total por região geográfica do país.

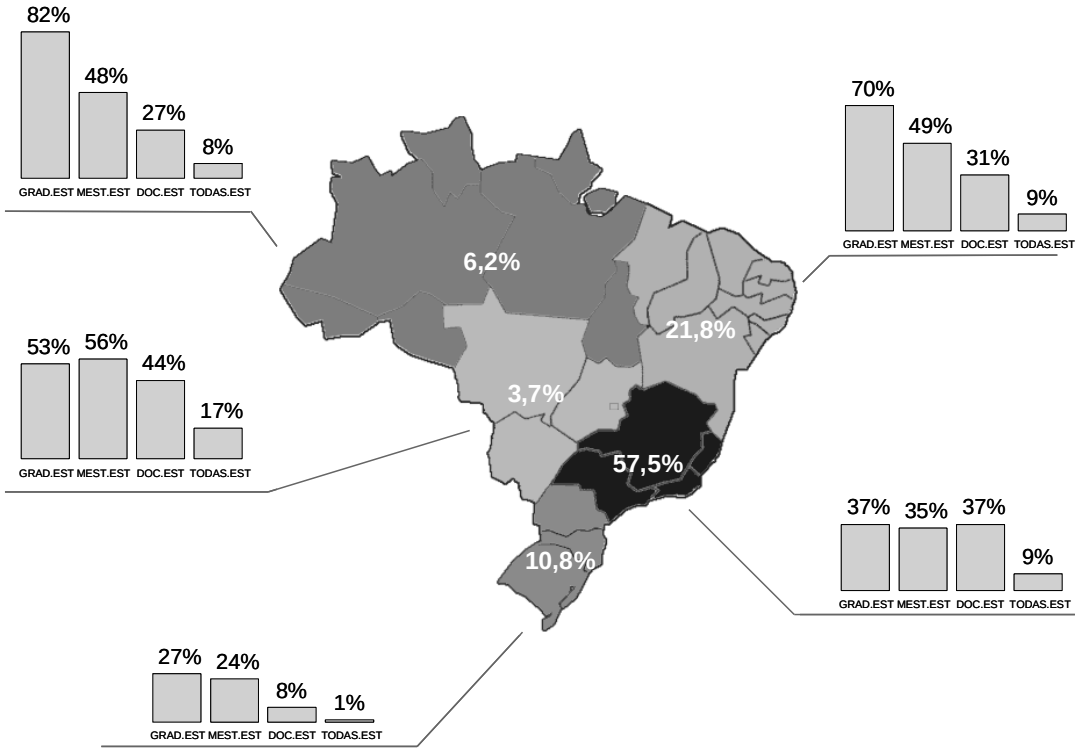


Figura 4. Titulação em estatística dos docentes e a proporção geral de docentes por região geográfica do país.

A Figura 3 exibe o número de artigos completos publicados entre os anos de 2006 e 2008, além disso, exibe a contribuição total de cada região do país. Desta forma, entre os anos avaliados, a publicação geral de artigos completos totaliza 1505, ou seja, aproximadamente 502 artigos completos por ano. Assim, a região sudeste é responsável por 53% destas publicações, seguida da região sul, responsável por 26,6%. A região norte é responsável por 2,6% das publicações nos anos analisados.

A Figura 4 apresenta as proporções de titulações obtidas na área de estatística (classificadas segundo a Tabela 3) por região geográfica, bem como exibe a proporção geral de docentes por região. Dentre as titulações, é baixa a proporção de docentes que possuem Graduação, Mestrado e Doutorado em Estatística, ou seja, docentes que se graduaram, realizaram seu mestrado e doutorado na área de estatística. Notamos também que a região Sul possui menor frequência de docentes titulados em estatística como área de concentração. A maior frequência de docentes que possuem Graduação em Estatística se encontra no Norte (82%), analogamente, a maior frequência de docentes com Mestrado em Estatística está na região Centro-Oeste, representada por

uma única universidade, (56%) e a maior frequência de docentes com Doutorado em Estatística na região Sudeste (37%). A frequência de docentes com todos os títulos em Estatística é similar entre as regiões Norte (8%), Nordeste (9%) e Sudeste (9%), mas na região Sul, somente 1% dos docentes têm todos os títulos em Estatística.

Ainda, avaliando a titulação dos docentes, consideramos uma nova avaliação referente à última titulação de cada docente. A Tabela 6 apresenta a distribuição para cada região geográfica, bem como a distribuição geral. Em todas as regiões existe a presença, em sua maioria, de docentes com titulação máxima de doutor, em especial as regiões Sul, Sudeste e Centro-Oeste, nas quais esse índice supera os 70%. A maior presença de docentes apenas graduados ocorre nas regiões Norte (29%) e Nordeste (11%). Com relação aos docentes que possuem apenas o título de mestre, as maiores frequências estão nas regiões Nordeste (32%) e Norte (29%).

Tabela 6. Ultima titulação dos docentes.

%		ULTIMA TITULAÇÃO		
		Graduação	Mestrado	Doutorado
Região	Norte (n=36)	28,5	28,5	43
	Sul (n=68)	7	21	72
	Sudeste (n=317)	2	22	76
	Centro-Oeste (n=21)	0	14	86
	Nordeste n=(119)	11	32	57
Geral		6	23	66

3.3. Descrição Por Tempo de Docência

A Tabela 7 apresenta a descrição das informações por tempo de docência, em três categorias: menos do que 5 anos, de 5 a 10 anos e mais do que 10 anos de docência. Dos docentes com tempo de docência menor que 5 anos, observamos a predominância masculina com 67%. Do mesmo modo, podemos observar que 72% dos docentes possuem graduação em Estatística, 72% possuem mestrado, destes 40% são Mestrado em Estatística. Ainda, 28% dos docentes não possuem doutorado e destes 56% são em Estatística como área de concentração. Além disso, 17% dos doutorados foram realizados no exterior. O número médio anual de publicações esteve entre 0,2 e 0,3, com um máximo de 4 artigos publicados.

No grupo de docentes com tempo de docência entre 5 e 10 anos, 43% dos docentes possuem Graduação em Estatística e 92% deles possuem mestrado, sendo que 33% possuem Mestrado em Estatística. Observamos que 66% dos docentes possuem doutorado e 27% destes o realizaram em Estatística, e 10% dos doutorados foram concluídos no exterior. O número médio de artigos publicados anualmente variou entre 0,6 e 0,7, atingindo um valor máximo de 10 artigos em 2008.

Para o grupo de docentes com mais de 10 anos de docência, 41% são graduados em Estatística e 95% possuem mestrado, sendo 39% destes mestrados são em Estatística. Observamos também que 80% dos docentes possuem doutorado, 32% destes foram realizados na área de Estatística e 28% foram concluídos no exterior. O número médio de artigos publicados anualmente ficou entre 0,9 e 1,2, com um máximo de 10 artigos publicados em 2008.

Tabela 7. Descrição das informações por tempo de docência

VARIÁVEL	TEMPO DE DOCÊNCIA		
	< 5 anos (n=67)	**entre 5 a 10 (n=79)**	**> 10 anos (n=383)**
SEXO: masculino(%)/ feminino(%)	67 / 33	60 / 40	61 / 39
GRAD.EST: sim(%)/não(%)	72 / 28	43 / 57	41 / 59
MEST: sim(%)/não(%)	72 / 28	92 / 8	95 / 5
MEST.EST: sim(%)/não(%)	40 / 60	33 / 67	39 / 61
DOC: sim(%)/não(%)	28 / 72	66 / 34	80 / 20
DOC.EST: sim(%)/não(%)	56 / 44	27 / 73	32 / 68
DOC.EXT: sim(%)/não(%)	17 / 83	10/90	28 / 72
AC.2006: Média; DP [Min, Max]	0,2; 0,4 [0, 1]	0,6; 1,3 [0, 8]	1,1; 1,9 [0, 12]
AC.2007: Média; DP [Min, Max]	0,3; 0,6 [0, 4]	0,6; 1,2 [0, 7]	1,1; 1,8 [0, 13]
AC.2008: Média; DP [Min, Max]	0,2; 0,5 [0, 2]	0,7; 1,5 [0, 10]	1,2; 2,3 [0, 16]
AC.TOTAL(%): 0 1-2 3-4 5-10 +10	60 33 6 1 0	45 32 10 11 2	34 31 12 15 8
BOLSISTA: sim(%)/não(%)	3 / 97	11 / 89	13 / 87

Comparando as 3 categorias de tempo de docência, observamos que a distribuição do número de artigos completos publicados entre 2006 a 2008 vai perdendo a assimetria com o aumento do tempo de docência. Dos docentes com tempo de docência menor do que 5 anos 60% não publicaram no período analisado. Entretanto, essa proporção cai para 34% para os docentes com mais de 10 anos de docência. Por outro

lado, a quantidade de docentes que publicaram mais do que 5 artigo no período passa de 1% para os docentes com tempo de docência menor do que 5 anos para 23% para os docentes com mais de 10 anos de docência. Aumento também é observado na proporção de docentes com bolsa de produtividade científica que passa de 3% para os docentes com tempo de docência menor do que 5 anos para 13% para os docentes com mais de 10 anos de docência.

3.4. Descrição Por Bolsa de Produtividade Científica

Na Tabela 8 apresentamos as descrições das variáveis estudadas categorizadas de acordo com a variável Bolsa de Produtividade Científica CNPq. Observamos alguns contrastes entre as categorias: 33% dos docentes bolsistas possuem Graduação em Estatística, enquanto que esse percentual sobe para 48% se analisarmos os docentes que não possuem bolsa. No grupo bolsista, 52% realizaram seu Doutorado em Estatística contra 28% do grupo de não bolsistas, 48% dos bolsistas realizaram o seu doutorado no exterior contra 21% dos não bolsistas. As distribuições do número de artigos completos de ambos os grupos são claramente distintas, com assimetrias direcionadas a lados opostos. Enquanto 72% dos bolsistas produziram mais do que 5 artigos no período e 33% produziram mais do que 10 artigos no período, 11% dos não bolsistas produziram mais do que 5 artigos e 2% produziram mais do que 10 artigos no período.

Tabela 8. Descrição de todas as Informações por ser ou não bolsista

VARIÁVEL	DOCENTE	
	BOLSISTA (n=61)	NÃO BOLSISTA (n=500)
SEXO: masculino(%)/ feminino(%)	68 / 32	61 / 39
GRAD.EST: sim(%)/não(%)	33 / 67	48 / 52
MEST: sim(%)/não(%)	93 / 7	91 / 9
MEST.EST: sim(%)/não(%)	44 / 56	37 / 63
DOC: sim(%)/não(%)	100 / 0	67 / 33
DOC.EST: sim(%)/não(%)	52 / 48	28 / 72
DOC.EXT: sim(%)/não(%)	48 / 52	21 / 79
TEMPO.D (em anos): média ± DP (variação)	20,1; 8,9 [3, 41]	18,2; 10,4 [1, 49]
AC.2006: Média; DP [Min, Max]	3,4; 3,0 [0, 12]	0,5; 1,1 [0, 9]
AC.2007: Média; DP [Min, Max]	2,9; 2,8 [0, 13]	0,6; 1,2 [0, 10]
AC.2008: Média; DP [Min, Max]	3,8; 3,7 [0, 16]	0,6; 1,4 [0, 11]
AC.TOTAL(%): 0	0	47
1-2	12	33
3-4	16	9
5-10	39	9
+10	33	2

Dando prosseguimento à análise, na Tabela 9 apresentamos os resultados considerando 3 categorias do número total de artigos produzidos pelo docente entre 2006 e 2008: docentes que não publicaram no período estudado (42%), docentes que publicaram entre 1 e 4 artigos no período (44%) e docentes que publicaram mais do que 4 artigos (18%). Dentre os docentes que não publicaram artigos completos entre 2006 a 2008 existe uma alta proporção de docentes que não publicaram e possuem Graduação em Estatística. Enquanto, 14% não possuem mestrado e 52% não possuem doutorado. A proporção de docentes que possuem Graduação em Estatística vai diminuindo para as outras categorias, sendo que 63% dos docentes que produziram mais de 4 artigos não são graduados em estatística. Estes docentes apresentam maiores proporções de mestres (47%) e de doutores (40%) em estatística, com também as maiores proporções de doutores formados no exterior (42%). Os docentes que produziram mais que 4 artigos completos possuem um tempo médio de docência maior do que as outras duas categorias (aproximadamente 4 anos de diferença). Dentre os docentes que produziram mais de 4 artigos no período estudado, 43% são bolsistas de produtividade científica do CNPq. Ainda, notamos que as regiões se distribuem de forma semelhante dentro das categorias de número de artigos publicados, porém destacamos

o aumento da presença de docentes da região Sul dentre as categorias, iniciando em 8% na categoria dos docentes que não publicaram no período, e chegando a 25% na categoria de docentes com mais do que 4 artigos. Além disso, destacamos a diminuição concomitante da presença dos docentes das regiões Norte e Nordeste com o aumento da quantidade de artigos publicados, iniciando com 9% e 26%, e chegando a 2% e 12%, respectivamente.

3.5. Descrição Conjunta Por Tempo de Docência e Bolsa de Produtividade

Nesta seção apresentamos os resultados referentes à descrição dos docentes, de acordo com as variáveis estudadas, por tempo de docência e bolsa de produtividade científica. Os resultados são apresentados na Tabela 10. A maioria dos docentes bolsistas não possui Graduação em Estatística, mas a proporção de graduados em estatística aumenta com o aumento do tempo de docência para o mesmo grupo, chegando a 36% para os docentes com mais tempo de docência. Com relação aos docentes não bolsistas, temos que a maioria possui essa titulação, porém a proporção de graduados em estatística apresenta movimento contrário, diminuindo com relação ao tempo de docência, iniciando com 75% para os mais novos, decrescendo para 42% para os mais antigos. Efeito semelhante é observado ao se analisar a variável referente ao Doutorado em Estatística para este grupo. Além disso, notamos um aumento gradativo no ritmo de publicação dos docentes com relação ao seu tempo de docência, sendo este intensificado para os docentes que possuem bolsa de produtividade científica do CNPq.

Tabela 9. Descrição de todas as Informações pelo número total de artigos.

VARIÁVEL	ARTIGOS PUBLICADOS DE 2006 A 2008		
	Igual a 0 (n=233)	Entre 1 e 4 (n=226)	Mais que 4 (n=102)
SEXO: masculino(%)/ feminino(%)	60 / 40	62 / 38	62 / 38
GRAD.EST: sim(%)/não(%)	53 / 47	44 / 56	37 / 63
MEST: sim(%)/não(%)	86 / 14	96 / 4	94 / 6
MEST.EST: sim(%)/não(%)	32 / 68	39 / 61	47 / 53
DOC: sim(%)/não(%)	48 / 52	81 / 19	96 / 4
DOC.EST: sim(%)/não(%)	27 / 73	31 / 69	40 / 60
DOC.EXT: sim(%)/não(%)	19 / 81	19 / 81	42 / 58
DI.VINCULO (anos): Média; DP [Min, Max]	17,4; 11,1 [1, 49]	17,7; 9,6 [1, 44]	22,0; 9,1 [3, 47]
BOLSISTA: sim(%)/não(%)	0 / 100	8 / 92	43 / 57
REGIÃO: norte(%) nordeste(%) centro-oeste(%) sudeste(%) sul(%)	9 26 3 54 8	6 21 4 59 10	2 12 5 56 25

Por fim, considerando apenas o ritmo de produção científica balizado pelo total de artigos completos publicados entre 2006 e 2008, apresentamos, na Figura 4, uma comparação geral entre as diferentes categorias. Existem docentes que apresentam uma quantidade de publicação muito acima dos demais. Deste modo, a Tabela 11 apresenta uma descrição geral das características dos docentes categorizados entre os que publicaram até 10 artigos nos 4 anos analisados e os que publicaram mais que 10 artigos. Comparativamente, temos que os docentes que apresentam um nível maior de publicação, em sua maioria, não têm Graduação em Estatística (71%), possuem doutorado (96%), grande parte destes no exterior (47%), tem mais tempo de docência, 6 anos (em média) a mais que os docentes que publicaram até 10 artigos. Além disso, 55% possuem bolsa de produtividade científica do CNPq. Por outro lado, no grupo de docentes que publicaram até 10 artigos no período analisado, 48% tem Graduação em Estatística, 32% não tem doutorado, dos docentes com doutorado, somente 22% o fizeram no exterior.

4. Comentários Finais

Sem pretensão de esgotar o tema, o presente levantamento teve por finalidade caracterizar os docentes dos cursos de Graduação em Estatística do país, com relação aos aspectos de sua formação e em termos da sua produção científica. O estudo aponta questões e diferenças importantes, dentre as quais, muitas já esperadas. Dos docentes analisados, 87% possuem currículo Lattes, ou seja, 13% dos docentes não são cadastrados na Plataforma Lattes. Caracterizando os docentes vinculados às Graduações em Estatística, de acordo com as variáveis estudadas, de uma forma geral, temos que estes são, em sua maioria homens (63%), com mestrado (91%) e doutorado (70%). No entanto, a formação estatística, como área de concentração, da maioria dos docentes não é grande. Apenas 46% dos docentes possuem Graduação em Estatística, dentre os que possuem mestrado somente 38% são mestres em estatística e 32% dos doutores possuem Doutorado em Estatística. Em termos de quantidade de publicações, temos que, enquanto 39% dos docentes não publicaram artigos no período, 39% publicaram entre 1 e 4 artigos e 22% dos mesmos publicaram mais do que 5 artigos.

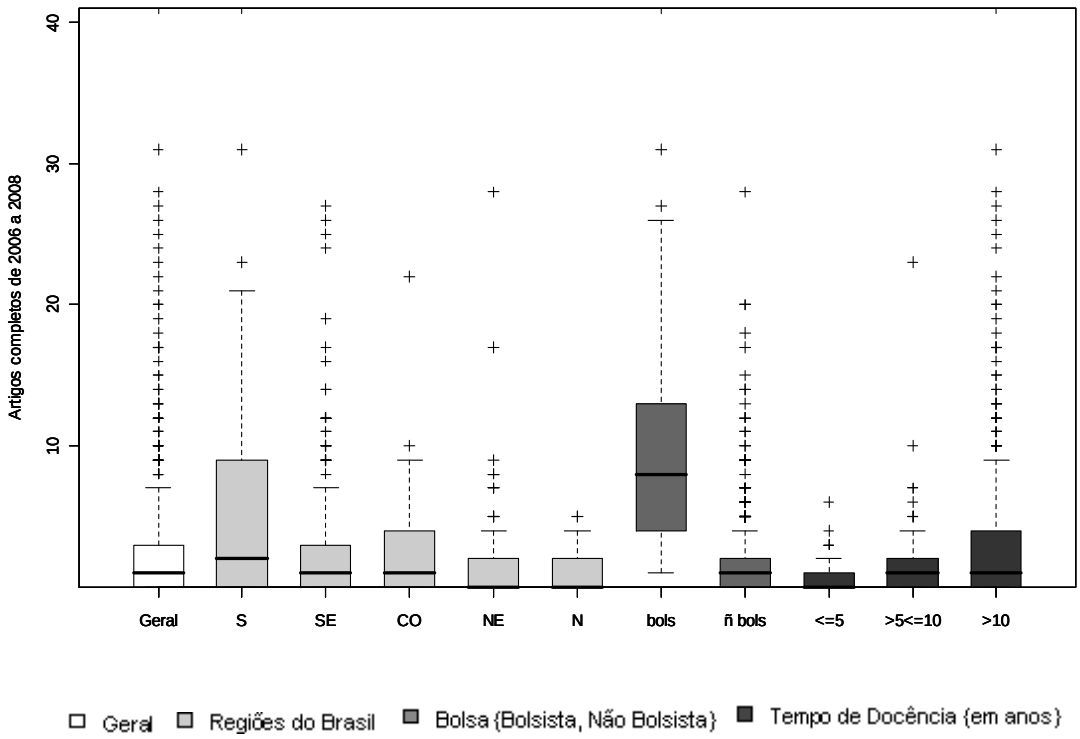


Figura 5. Comparação geral do número total de artigos completos de 2006 a 2008.

Tabela 10. Descrição dos docentes por tempo de docência e por bolsa de produtividade

	VARIÁVEL	Tempo de Docência		
		até 5 anos	de 5 a 10 anos	mais que 10 anos
Não Bolsista	n	65	71	332
	SEXO: masculino(%)/ feminino(%)	66 / 34	60 / 40	60 / 40
	GRAD.EST: sim(%)/não(%)	75 / 25	47 / 53	42 / 58
	MEST: sim(%)/não(%)	72 / 28	93 / 7	95 / 5
	MEST.EST: sim(%)/não(%)	38 / 62	35 / 65	38 / 62
	DOC: sim(%)/não(%)	25 / 75	62 / 38	77 / 23
	DOC.EST: sim(%)/não(%)	56 / 44	24 / 76	27 / 73
	DOC.EXT: sim(%)/não(%)	12 / 88	6 / 94	24 / 76
	AC.2006: Média; DP [Min, Max]	0,2; 0,4 [0, 1]	0,3; 0,7 [0, 3]	0,7; 1,2 [0, 9]
	AC.2007: Média; DP [Min, Max]	0,3; 0,6 [0, 4]	0,4; 0,7 [0, 4]	0,8; 1,4 [0, 10]
	AC.2008: Média; DP [Min, Max]	0,2; 0,5 [0, 2]	0,5; 0,8 [0, 4]	0,8; 1,6 [0, 11]
	AC.TOTAL(%): 0 1-2 3-4 5-10 +10	61 31 6 2 0	51 33 9 7 0	39 35 11 11 4
Bolsista	n	2	8	51
	SEXO: masculino(%)/ feminino(%)	100 / 0	57 / 43	67 / 33
	GRAD.EST: sim(%)/não(%)	0 / 100	14 / 86	36 / 64
	MEST: sim(%)/não(%)	50 / 50	83 / 17	46 / 54
	MEST.EST: sim(%)/não(%)	100 / 0	20 / 80	45 / 55
	DOC: sim(%)/não(%)	100 / 0	100 / 0	100 / 0
	DOC.EST: sim(%)/não(%)	50 / 50	43 / 57	54 / 46
	DOC.EXT: sim(%)/não(%)	50 / 50	29 / 71	50 / 50
	AC.2006: Média; DP [Min, Max]	1,0; 0,0 [1, 1]	2,6; 2,6 [0, 8]	3,7; 3,0 [0, 12]
	AC.2007: Média; DP [Min, Max]	0,5; 0,7 [0, 1]	2,1; 2,8 [0, 7]	3,2; 2,9 [0, 13]
	AC.2008: Média; DP [Min, Max]	0,5; 0,7 [0, 1]	2,4; 3,6 [0, 10]	4,2; 3,7 [0, 16]
	AC.TOTAL(%): 0 1-2 3-4 5-10 +10	0 100 0 0 0	0 29 14 43 14	0 6 17 40 37

Tabela 11. Descrição geral dos docentes divididos entre os que publicaram até 10 artigos e os que publicaram mais que 10 artigos.

VARIÁVEL	Até 10 Artigos (n=527)	Mais que 10 Artigos (n=34)
SEXO: masculino(%)/ feminino(%)	61 / 39	59 / 41
GRAD.EST: sim(%)/não(%)	48 / 52	21 / 79
MEST: sim(%)/não(%)	91 / 9	94 / 6
MEST.EST: sim(%)/não(%)	38 / 62	38 / 62
DOC: sim(%)/não(%)	69 / 31	97 / 3
DOC.EST: sim(%)/não(%)	31 / 69	42 / 58
DOC.EXT: sim(%)/não(%)	23 / 77	45 / 55
TEMPO.D (em anos): Média; DP [Min, Max]	17,9; 10,2 [1, 49]	25,0; 9,2 [10, 45]
AC.2006: Média; DP [Min, Max]	0,5; 0,9 [0, 6]	5,8; 2,7 [2, 12]
AC.2007: Média; DP [Min, Max]	0,5; 0,9 [0, 7]	5,6; 2,6 [1, 13]
AC.2008: Média; DP [Min, Max]	0,6; 1,2 [0, 7]	6,6; 3,7 [1, 16]
AC.TOTAL(%): 0 1-2 3-4 5-10 +10	44 32 11 13 0	0 0 0 0 100
BOLSISTA: sim(%)/não(%)	8 / 92	59 / 41

Considerando as regiões geográficas do país, observamos que as regiões Norte, Nordeste e Centro-Oeste possuem as maiores proporções de docentes graduados em estatística, respectivamente 82%, 70% e 53%, com a ressalva de que a região Centro-Oeste, nesta pesquisa, é representada somente por uma universidade. As regiões Norte e Nordeste possuem as menores quantidades de docentes com doutorado em comparação com as demais, respectivamente 42% e 58%, a região Sul é a que possui menor proporção de doutores em estatística do Brasil, apenas 15%. Na região Norte não existem docentes que realizaram seu doutorado no exterior. Dentre as regiões geográficas, considerando a quantidade de artigos completos publicados no período analisado, dentro do contexto considerado, a região Sul apresenta a maior quantidade de artigos publicados, enquanto, o Norte e o Nordeste são as regiões com as menores quantidades.

Com relação ao tempo de docência em instituição de ensino superior, temos que os docentes mais novos (até 5 anos de docência) possuem, em sua maioria, Graduação em Estatística (72%), o que não ocorre com as categorias, de 5 a 10 anos e mais de 10 anos de vínculo empregatício, respectivamente, 43% e 41%. Para os docentes na categoria 5 a 10 anos de docência, existe a menor frequência de docentes com

Mestrado em Estatística (33%) e de docentes com Doutorado em Estatística (28%), isto possivelmente devido ao fato de que os mestrados e doutorados em estatística são mais recentes. A maior presença de doutorados no exterior é observada entre os docentes que com vínculo há mais de 10 anos (28%). Além disso, a maior quantidade de artigos publicados no período estudado é apresentada pelos docentes mais antigos.

Com relação às bolsas de produtividade científica, como esperado, os docentes bolsistas possuem maior de produção científica, em média, em comparação com os docentes não bolsistas, aproximadamente 3 e 0,5 artigo anuais, respectivamente, sendo que 33% dos bolsistas publicaram mais do que 10 artigos no período estudado. Esta quantidade ser reduz a 2% dos docentes que não possuem bolsa.

A análise conjunta do tempo de docência e bolsa de produtividade científica revela, como esperada, a existência de interação positiva entre tempo de docência e bolsa de produtividade, com um aumento gradativo no ritmo da produção científica dos docentes com o aumento do tempo de docência, o qual é intensificado pela presença da bolsa de produtividade.

Os docentes que publicaram mais do que 10 artigos no período estudado, em sua maioria, não têm Graduação em Estatística, possuem doutorado, quase a metade no exterior, tem mais tempo de docência e com mais da metade sendo bolsista de produtividade quando comparados com os docentes que publicaram até 10 artigos no período.

Como mencionado inicialmente, este trabalho considera algumas dentre as possíveis informações contidas no currículo Lattes dos docentes, direcionando a descrição para algumas visualizações do contexto acadêmico estatístico do país. Porém outras informações podem ser levantadas através deste, como por exemplo, o número de alunos orientados pelo docente (em iniciação científica, trabalho de graduação, mestrado, doutorado e pós-doutorado), a quantidade de livros publicados, o número de disciplinas ministradas, dentre outras, que com certeza, colaborariam para um entendimento ainda maior das várias dimensões que compõem a caracterização do desempenho de função dos profissionais aqui descritos. Também, não pesquisamos as atividades de extensão do docente, mesmo sendo estas de extrema importância na formatação do trinômio ensino-pesquisa-extensão. O aspecto qualitativo da produção científica do docente não foi considerado, pois para uma análise qualitativa precisamos

qualificar cerca de 1500 artigos currículo a currículo, o que esta fora do alcance do presente artigo. Entretanto, entendemos que a qualificação dos artigos publicados é importante e necessária e deverá ser considerada em um estudo futuro. Uma possibilidade é o enriquecimento das informações contidas no currículo Lattes com informações de outras bases de dados referentes, dentre outras, a quantidade de citações dos artigos do docente e o fator de impacto dos periódicos em que seus artigos foram publicados.

Além disso, claramente, as informações coletadas a partir do currículo Lattes, enriquecidas por informações advindas de outras bases de dados, podem ser úteis em termos de planejamento estratégico. Com exemplo, é essencial visualizar a necessidade de material humano estatístico dentro das Graduações de Estatística do país. Esta visualização deve ter como base a situação atual das Pós-Graduações em Estatística do país, em termos de oferta de doutores egressos, em contra posição da quantidade necessária de doutores para suprir as vagas de docentes dentro das Graduações em Estatística, assumindo, como premissa, que estas vagas serão preenchidas somente por doutores egressos dos Programas de Doutorado em Estatística.

Em termos de demanda de doutores, considerando a variável TEMPO.D (tempo de docência), sabemos quanto tempo um determinado docente está vinculado à academia, e consequentemente, assumindo um tempo médio de 30 anos de serviço para aposentadoria, conseguimos saber em que ano um determinado docente irá se aposentar. Estes valores são aproximados, pois dentre os docentes vinculados aos Departamentos de Estatística podem existir docentes aposentados, bem como docentes que trabalharam em outras instituições antes de terem iniciado suas carreiras acadêmicas. A Tabela 12 apresenta as previsões do número necessário de doutores egressos necessários para suprir as vaga em aberto a partir das ocorrências das aposentadorias dos docentes das 24 Graduações em Estatística estudadas.

Por outro lado, em termos de oferta de doutores egressos, a Figura 6 apresenta a quantidade de egressos das Pós-Graduações em Estatística nos últimos 10 anos (1998-2008) em termos de mestres e doutores formados (informação disponível no site da CAPES, http://www.capes.gov.br). Claramente observamos uma tendência de crescimento do número de egressos, com 72 mestres e 20 doutores, em média, sendo formados a cada ano a partir de 2005. Em termos de previsão de oferta de doutores

egressos, acreditamos que a quantidade de doutores egressos deve aumentar de 50% a 60% a partir de 2010, uma vez que, neste momento, várias das Pós-Graduações apresentadas na Tabela 3 deverão ter formado seus primeiros doutores. Como premissa, assumimos que, a partir de 2010, os programas de doutorado das instituições, UFMG, UFPE, UFRJ, UFSCar, USP e UNICAMP deverão formar 4, 4, 4, 4, 12 e 4 doutores, respectivamente, perfazendo a quantidade de 32 doutores egressos anualmente. Também, em termos de apresentação da demanda de doutores, assumimos 3 possíveis cenários: Pessimista, Base e Otimista, sendo que o Cenário Base corresponde a oferta atual de 32 doutores egressos, enquanto os Cenários Pessimista e Otimista correspondem a oferta de 30% a menos e a mais de doutores egressos, isto é, 22 e 42 doutores egressos anualmente. Além disso, contemplamos a possibilidade de 20% dos doutores egressos não permanecerem na academia, sendo absorvidos pelo mercado.

Tabela 12. Número de aposentadorias e previsão de demanda de Doutores em Estatística.

Ano	Número de Aposentados	Previsão de Demanda
2010	92	92
2011	6	98
2012	13	111
2013	9	120
2014	5	125
2015	19	144
2016	14	158
2017	20	178
2018	11	189
2019	23	212
2020	14	226
2021	21	247
2022	18	265
2023	27	292
2024	23	315
2025	21	336
2026	15	351
2027	15	366
2028	17	383
2029	17	400
2030	13	413

A Tabela 13 apresenta os déficit de Doutores em Estatística necessários para suprir as vagas em aberto a partir das ocorrências das aposentadorias dos docentes das

24 Graduações em Estatística estudadas, de acordo com a Tabela 12, considerando os 3 cenários definidos acima, bem como a possibilidade de absorção pelo mercado de 20% dos doutores egressos. Considerando um cenário em que os Programas de Doutorado em Estatística do país formem, em média, 32 doutores, até o ano de 2012 ainda devemos ter déficit de doutores para preencher as vagas de docentes. O problema se agrava se o número médio de doutores egressos diminui. Considerando que todos os doutores egressos ficarão na academia, no Cenário Pessimista, até o ano de 2017 teríamos ainda déficit de doutores para preencher as vagas de docentes. Em última instância, assumindo que 20% dos doutores egressos são absorvidos pelo mercado a cada ano, no Cenário Pessimista, ainda em 2030 o déficit de doutores deverá perdurar.

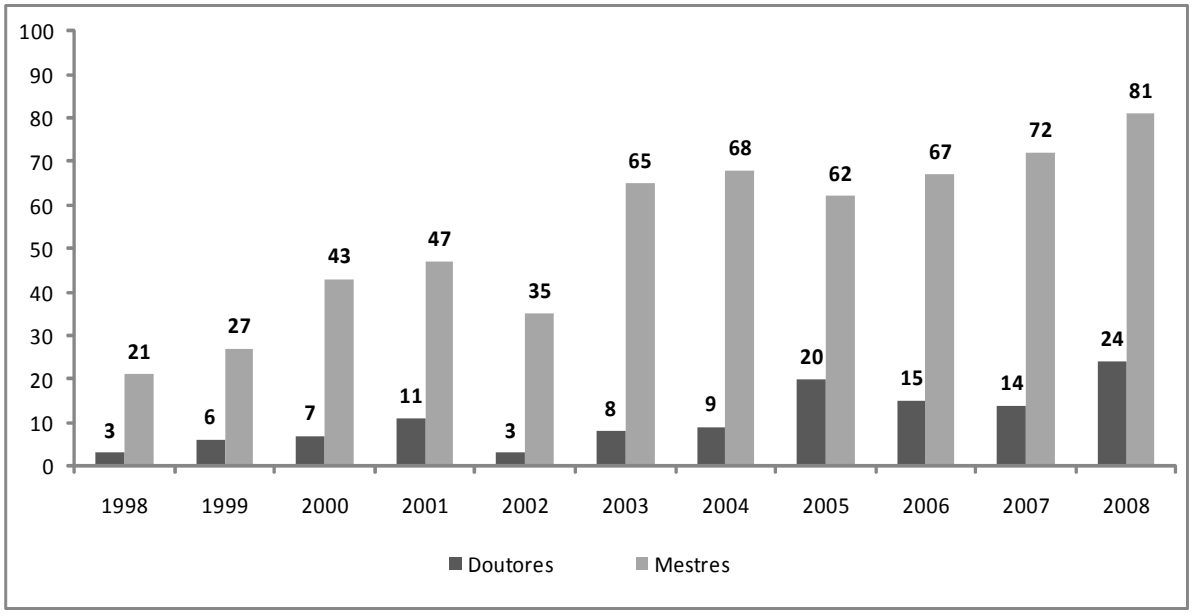


Figura 6. Número de Egressos das Pós-Graduações em Estatística do país de 1998 a 2008.

Tabela 13. Déficit de Doutores em Estatística necessários para suprir as vagas em aberto a partir das ocorrências das aposentadorias dos docentes das Graduações em Estatística.

Ano	Todos Egressos Permanecem na Academia			20% dos Egressos Absorvidos Pelo Mercado		
	Otimista	Base	Pessimista	Otimista	Base	Pessimista
2010	50	60	70	58	66	74
2011	14	34	54	30	46	62
2012		15	45	9	33	57
2013			32		16	48
2014			15			35
2015			12			36
2016			4			32
2017			2			24
2018						27
2019						32
2020						28
2021						31
2022						31
2023						40
2024						45
2025						48
2026						45
2027						42
2028						41
2029						40
2030						35

Agradecimentos

Os autores agradecem o Editor Executivo e os quatro revisores pelos excelentes comentários. Também agradecem a Airlane Alencar e Renato Assunção pelos **comentários ao artigo que são publicados juntamente com o mesmo,** a Benedito Galvão Benzé pela leitura cuidadosa do material, a Carlos de Bragança Pereira, Dani Gamerman e Gleici da Silva Castro Perdoná por sugestões e comentários referentes ao tema, e aos Coordenadores dos cursos de Graduação em Estatística que forneceram os nomes de seus docentes quando a homepage de seus departamentos não foi encontrada. Este trabalho teve o financiamento parcial das instituições governamentais: CNPq e CAPES.

Referências Bibliográficas

Araujo, M. E. A história da estatística. In: Semana do Estatístico - 93. Departamento de Estatística da UFRN, Natal, 1993.

Dantas, C. A. B. Desenvolvimento da Estatística na Universidade de São Paulo, In: Associação Brasileira de Estatística, História da Estatística no Brasil. Disponível em http://www.redeabe.org.br/historia.htm. Acesso em 23 de dezembro de 2009. ABE biênio 32000/2002.

Cramér, Harald. Mathematical Methods of Statistics. Princeton, New Jersey. Princeton University Press, 1946.e-MEC. Disponível em http://emec.mec.gov.br. Acesso em 31 de outubro de 2009.

IBGE. Fundação Instituto Brasileiro de Geografia e Estatística. Calendário comemorativo dos 50 anos de fundação do IBGE. Rio de Janeiro, 1987.

Lopes, C.E. "Probabilidade E A Estatística No Ensino Fundamental: Uma Análise Curricular" – Tese de Mestrado – Universidade Estadual de Campinas, 1998.

Memória, J. M. P. Breve História da Estatística. Texto para Discussão 21. Brasília, DF: Embrapa Informação Tecnológica, 2004.

Plataforma Lattes. Disponível em http//:lattes.cnpq.br. Acesso de Julho a Novembro de 2009.

Silva, L. F. A Fundação Instituto Brasileiro de Geografia e Estatística - IBGE e a produção das estatísticas. In: Revista Brasileira de Estatística - RBEs. Rio de Janeiro, Ano 50, n.193, 1989.

Apêndice

Neste apêndice apresentamos na Tabela A.1. as características gerais dos bacharelados em estudo. As informações foram retiradas do portal e-MEC (http://emec.mec.gov.br/).

Tabela A.1. Detalhes dos cursos de Graduação em Estatística considerados nesta pesquisa.

Univer-sidade	Grau	Município/UF	Data de Início do Curso	Per.*	Total de Vagas	Créd.**	Duração créditos (anos)	Carga horária (horas)	Número de Doc. ***
ENCE	Bach.	Rio de Janeiro/RJ	06/03/53	M e N	120	S	4,0	3024	44
UNB	Bach.	Brasília/DF	01/03/74	I	48	S	3,5	2760	24
USP	Bach.	São Paulo/SP	01/01/72	I	30	S	4,0	2055	47
UERJ	Bach.	Rio de Janeiro/RJ	26/10/69	M e N	90	S	4,0	2670	98
UEPB	Bach.	Campina Grande/PB	03/03/80	M e N	80	A	5,0	2574	11
UNICAMP	Bach.	Campinas/SP	01/03/69	I	70	S	4,0	4605	22
UEM	Bach.	Maringá/PR	20/03/00	I	40	A	4,0	3271	22
UNESP	Bach.	Presidente Prudente/SP	27/12/83	M	30	A	4,0	2820	35
UFBA	Bach.	Salvador/BA	03/03/84	I	40	S	4,0	3060	28
UFPB	Bach.	João Pessoa/PB	16/10/00	M	25	S	4,0	2760	19
UFJF	Bach.	Juiz de Fora/MG	01/03/07	M	25	S	4,0	2350	15
UFMG	Bach.	Belo Horizonte/MG	01/03/78	M	45	S	3,5	3000	27
UFPE	Bach.	Recife/PE	01/03/68	I	30	S	4,0	2760	18
UFSCar	Bach.	São Carlos/SP	18/07/76	I	30	S	4,0	2820	19
UFS	Bach.	São Cristóvão/SE	27/03/00	N	50	S	4,5	2805	15
UFAM	Bach.	Manaus/AM	01/07/76	I	48	S	5,0	2715	24
UFC	Bach.	Fortaleza/CE	01/01/65	I	60	S	4,0	2700	25
UFES	Bach.	Vitória/ES	05/01/87	M	40	S	4,5	3015	17
UFPA	Bach.	Belém/PA	28/02/75	M	50	S	4,0	2700	16
UFPR	Bach.	Curitiba/PR	01/01/74	N	66	S	4,5	2700	21
UFRJ	Bach.	Rio de Janeiro/RJ	18/06/46	I	15	S	4,0	2655	24
UFRN	Bach.	Natal/RN	01/03/74	M	50	S	3,5	2945	26
UFRGS	Bach.	Porto Alegre/RS	01/03/97	I	20	S	4,0	2790	27
	Bach./ Lic. Plena	Porto Alegre/RS	01/03/97	I	40	S	4,0	2790	
UFF	Bach.	Niterói/RJ	01/3/07	M	60	S	4,0	3260	24

* Período do curso: M = matutino, N = noturno e I = integral; ** Periodicidade em que os créditos são integralizados: S = Semestralmente, A = Anualmente; *** Número de docentes vinculados ao departamento de Estatística de cada universidade, dados coletados através desta pesquisa.

Comentários ao artigo Caracterização dos docentes de Louzada et all.

Renato Assunção

Os autores fizeram um bom trabalho descritivo da situação dos docentes em departamentos que oferecem cursos de graduação em estatística no Brasil. Devido à carência de dados, este esforço é meritório e devemos agradecer aos autores pelo seu esforço. Meu comentário talvez seja reduzido a apenas um, desdobrado de várias formas. Eu tive bastante dificuldade em interpretar os resultados apresentados e pensar nas conseqüências de longo prazo que o quadro delineado implica. A falta de uma visão histórica, do desenvolvimento ao longo do tempo, e das possíveis decomposições multivariadas de algumas estatísticas representaram, para mim, a maior dificuldade de usar os resultados do artigo. Isto não é uma crítica aos autores mas a constatação de que precisamos de mais artigos como este para que este quadro mais completo possa emergir com o tempo.

Ao fazer uma análise descritiva, é útil termos uma estrutura de referência com a qual as estatísticas obtidas possam ser comparadas. No caso atual, esta referência poderia ser a mesma situação em alguma época passada (30 anos atrás?) ou algumas outras áreas da ciência relevantes para a estatística tais como a computação, engenharia e a matemática. Sem estas âncoras de referência, fica difícil interpretar resultados sobre, por exemplo, a dispersão geográfica, a composição por sexo ou a taxa de publicação.

A falta de referência e de informações adicionais dificulta a análise dos resultados. Por exemplo, verifica-se que a região centro-oeste possui 100% de seus docentes com mestrado, uma porcentagem maior que a de todas as outras regiões. O que concluir disso? Como interpretar este resultado? Será que os departamentos do centro-oeste são universidades mais jovens, que fizeram contratações mais recentemente? Talvez por isto tenham contratado pessoas com formação acadêmica mais completa que universidades mais antigas, com professores mais idosos e contratados há muito tempo atrás. Ou a estatística reflete uma política de atração de quadros mais qualificados? Ou existe uma política mais agressiva de formação de pessoal já qualificado? Todas estas possíveis explicações implicam em situações completamente diferentes para o futuro e não podemos optar por nenhuma delas.

O professor Gauss Cordeiro comentou em mensagem veiculada na ABE-L um dos achados do artigo. Dentre os docentes dos Cursos de Graduação em Estatística com produção acadêmica de mais de 4 artigos publicados no triênio 2006-2008, 63% não fizeram graduação em Estatística. Este fato parece indicar que a graduação em estatística não prepara os docentes para a atividade de pesquisa. No entanto, como observou o professor Paulo Justiniano Ribeiro também na ABE-L, esta estatística pode refletir simplesmente uma lenta mudança histórica. No passado, o mais comum é que os programas pós-graduação em estatística recebessem, em sua maioria, pessoas

com formações em áreas diferentes da estatística, principalmente em matemática e engenharia. Isto vem mudando nos últimos anos onde a maioria dos estudantes de pós-graduação são formados em estatística. Sem saber qual era a composição de pesquisadores produtivos no passado (seriam 100% deles sem formação estatística?), fica difícil concluir algo acerca da sua divisão atual em 63% e 37%.

Eu acredito que o atual artigo possui duas importantes contribuições. A primeira delas é tirar um retrato da situação nos departamentos de estatística. Devemos complementar este retrato com outros, seja no futuro, seja no presente, para que políticas de formação profissional possam ser mais bem direcionadas.

A segunda contribuição é o quadro preocupante da demanda por doutores em estatística no Brasil apenas para suprir a necessidade de substituir os atuais professores quando eles se aposentarem. O número projetado é muito maior que a capacidade instalada dos atuais programas de pós-graduação em estatística e não existe nenhum sinal de que isto vá mudar de imediato. As conseqüências disso para a estatística brasileira podem ser desastrosas e pedem um esforço de pensar em políticas de formação que procurem diminuir o impacto que possa vir da contratação de pessoas sem formação adequada para os postos que terão de ser preenchidos.

Comentários ao artigo Caracterização dos docentes de Louzada et all.

Airlane Alencar

O artigo "Caracterização dos Docentes dos Cursos Públicos de Graduação em Estatística no Brasil: Descrição de Algumas das Dimensões que Compõem o Perfil Desta Classe de Profissionais", apresentado nesse exemplar da Revista Brasileira de Estatística, é pioneiro e fundamental. É pioneiro por não haver até o momento nenhum levantamento desse tipo e é fundamental, pois o recente crescimento do número de cursos universitários no país requer esse tipo de trabalho para orientar a criação de cursos necessários e de qualidade de acordo com a demanda de profissionais de cada área.

Além de apresentar o perfil dos docentes em departamentos de estatística das universidades públicas, esse artigo apresenta uma previsão da demanda de Doutores em Estatística somente para preencher as vagas de professores doutores nas universidades públicas. É importante ressaltar que a demanda por profissionais de estatística, tanto para graduados, quanto para pós graduados, é muito maior. Somente na lista Stat-Math [1] de discussão na internet são oferecidos em média umas 2 vagas por dia (não fiz os cálculos, somente me baseei na quantidade de mensagens que recebo). O conselho regional de estatística (CONRE) também divulga muitas vagas. Isso significa que a demanda por estatísticos é muito grande.

É usual ler na imprensa que há uma escassez de engenheiros e de outros profissionais. Recentemente, foi citado na lista da Associação Brasileira de Estatística o discurso do documentarista João Moreira Salles em evento da Academia Brasileira de Ciências [2]. Entre outros ricos comentários, ele apresenta os seguintes dados: "Em 2008, segundo o Inep [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira], o país formou 1.114 físicos, 1.972 matemáticos e 2.066 modistas."; "A Rússia forma 190 mil engenheiros por ano, a Índia, 220 mil e a China, 650 mil, diz o relatório. Nós formamos 47 mil." Isso torna evidente que o país apresenta déficit de alguns profissionais e o crescimento do país depende da formação de bons profissionais em todas as áreas, sem desmerecer nenhuma profissão.

Outro aspecto importante sobre a demanda de estatísticos é que nossa profissão é bastante bem avaliada no ranking de profissões como apresentado regularmente em pesquisa divulgada no Wall Street Journal [3]. Essa pesquisa leva em considerações aspectos como estresse, remuneração, desgaste físico e empregabilidade e classifica a profissão de estatístico como a terceira melhor em 2009 e a oitava melhor em 2010.

Não há dúvida de que a profissão de estatística é muito requisitada, mas é claro que há diferenças regionais, tanto com relação à oferta de emprego, quanto com relação à remuneração. O artigo apresentado também caracteriza os docentes por região do país. Seria fundamental observar que talvez a demanda por estatísticos apresente nuances, sendo que, por exemplo, em algumas áreas sejam mais requisitados profissionais para trabalhar com finanças, enquanto em outras regiões sejam necessários profissionais de controle estatístico de qualidade.

Quanto à excelência dos cursos de estatística no país, essa preocupação deve ocorrer tanto para os cursos de graduação, quanto nos cursos de pós. Além da qualidade do ensino e do incentivo à pesquisa feita pelos alunos, é importante observar que ao longo do tempo são necessárias atualizações nos cursos. Por exemplo, algumas disciplinas que eram somente optativas na graduação, já deveriam ser obrigatórias (são chamadas de optatórias), como por exemplo, modelos lineares generalizados. Esse é somente um exemplo localizado, mas há muitas matérias necessárias para a formação de um bom estatístico. Outras questões como o desenvolvimento computacional levam à necessidade do estatístico trabalhar com grandes bases de dados e atualmente a maioria das vagas divulgadas exigem conhecimento de ferramentas para lidar com eles. O estatístico pode até trabalhar em conjunto com profissionais especializados em banco de dados, entretanto deve saber o mínimo para o trabalho em conjunto. Outro aspecto relevante é a interdisciplinaridade já que a estatística auxiliar a análise de dados em todas as áreas.

A caracterização dos docentes dos cursos de graduação em estatística apresentou que menos da metade dos docentes tem graduação em estatística e que dentre mestres e doutores,somente um terço apresentam a correspondente pós graduação em Estatística. Isso é até esperado, pois os

cursos de estatística são recentes, principalmente em algumas regiões do país. Eu mesma já tive excelentes aulas com professores que não tem graduação em estatística.

O importante é aumentar a quantidade, mas somente se for mantido o alto nível dos cursos de graduação, mestrado e doutorado. O crescimento da importância da profissão estatística depende somente do bom desempenho dos profissionais, tanto acadêmicos, quanto no mercado de trabalho em geral.

Referências

[1] STAT-MATH. http://br.dir.groups.yahoo.com/group/STAT-MATH/

[2] Folha de S.Paulo, em 6 de junho de 2010.

[3] Neddleman, S. E. Doing the Math to Find the Good Jobs .Disponível em http://online.wsj.com/article/SB123119236117055127.html em 24/junho/2010

ENCONTROS E REENCONTROS: UM DIAGNÓSTICO DA DINÂMICA MATRIMONIAL NO NORDESTE DO BRASIL

Flávio Henrique Miranda de Araújo Freire[1]
Maria Helena Constantino Spyrides
Moisés Alberto Calle Aguirre
Kátia Lucianny de Souza

Resumo

Objetivo: Analisar a dinâmica dos estados nupciais no Nordeste do Brasil, segundo sexo e idade para o ano 2000. Método: Utilizaram-se as informações do Registro Civil e do Censo Demográfico (IBGE), de 2000. Para transformar as taxas em probabilidades de transição, utilizaram-se conceitos de tabelas de sobrevivência, mais precisamente tabelas multiestados. Resultados: A proporção de solteiros tem aumentado sistematicamente em ambos os sexos, já a de casados experimentou tendência inversa. Verificaram-se 7% de mulheres viúvas, enquanto para os homens apenas 2%. Conclusão: A probabilidade de casar pela primeira vez é maior nas mulheres com relação aos homens, já o re-casamento é mais provável por parte do homem do que entre as mulheres.

[1] UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE - Centro de Ciências Exatas e da Terra - Departamento de Estatística – DEST - Programa de Pós-graduação em Ciências Climáticas - PPGCC

1. INTRODUÇÃO

Falar da nupcialidade no sentido de sua dinâmica é falar de sujeitos que experimentam encontros e desencontros. Os encontros são gerados no plano da sensibilidade humana e nas disposições que projetam os sujeitos a um processo de construção das relações, cujo núcleo descansa na racionalidade da reciprocidade e essencialmente da solidariedade. É nestas dimensões onde se concretizam os encontros nupciais, os quais projetam os sujeitos a fortalecer e gerar novas aspirações e novos objetivos em seu projeto de vida. Já os desencontros têm uma trajetória inversa, eles são gerados essencialmente pelo conflito entre sujeitos que leva ao processo de desgaste das relações e conseqüentemente ao enfraquecimento, principalmente, das reciprocidades, levando inexoravelmente à concretização dos desencontros nupciais. Esses desencontros na sociedade moderna têm uma diversidade de desprendimentos e significados, um deles aponta e se transforma na janela de novas oportunidades na busca de reencontros, a sua concretização lhe dá dinâmica a nupcialidade. Assim, encontros, desencontros e reencontros são as dimensões que lhe dão dinâmica à nupcialidade e deles trata as análises do presente artigo.

Essas singularidades (encontros, desencontros e reencontros) da nupcialidade transformam-se em uma das componentes demográficas de maior relevância na atualidade, não apenas pela sua associação direta com os padrões de formação e dissolução de famílias, mas também, por que elas determinam o mercado matrimonial.

A dinâmica dos estados da nupcialidade no Brasil, mais especificamente da região Nordeste, possivelmente até os anos 60 do século passado, em geral, estava regida por normas e valores de uma sociedade ainda tradicional, onde a transição do estado de solteiro para casado era presumivelmente produto de arranjos familiares. Ao passo que as uniões consensuais e o trânsito para o estado de divórcio quase não existiam, ainda, o re-casamento era permitido apenas em situação de viuvez. Esta radiografia da nupcialidade vai começar a mudar a partir da década dos 70.

Essas transformações observadas na estrutura da nupcialidade estariam se dando como conseqüência de mudanças no espaço macro e espaço micro da realidade brasileira, isto é: i) na visão macro, a análise se sustenta na intensidade das transformações de ordem econômica, social e cultural que o Brasil experimentou nos

últimos quarenta anos, e ii) na visão micro, o motor transformador se alude à mulher, a suas mudanças de comportamento - fruto do ambiente macro - que têm e estão levando a construir uma sociedade cada vez mais moderna no sentido de gerar novos valores e hábitos. Isso é o reflexo de reivindicações importantes alcançadas na sua posição como sujeitos na sociedade, não só na conquista de seus direitos de cidadã, mas também, na conquista de espaços na esfera pública. Prova disso, é a sua inserção com trajetória crescente na educação e no mercado de trabalho formal, fatos que permitiram avançar na transformação da sociedade para que comportamentos antes não permitidos sejam agora aceitos pelo consenso social, como a união consensual, o divórcio, o re-casamento, fenômenos que deram nova dinâmica à nupcialidade. Essas singularidades estariam re-definindo um novo modelo nupcial no Brasil.

Assim, a nova dinâmica da nupcialidade estaria levando à conformação de estruturas novas de famílias. Como Medeiros & Osório[1] mostram, o tipo de arranjo familiar que mais cresceu no Brasil, no período de 1978 a 1998, foi o arranjo de núcleo simples feminino, caracterizado por domicílios chefiados por mulher sem cônjuge, seja com ou sem filho. Outro tipo de arranjo familiar que também experimentou crescimento nos últimos 20 anos corresponde ao formado por um casal e domicílios chefiados por homens sem cônjuge.

Essa nova tendência de configuração nos arranjos familiares brasileiros é o reflexo não apenas das mudanças sociais e culturais experimentadas pela sociedade brasileira, é o reflexo também, das mudanças entre os estados conjugais da população. Certamente, o Nordeste está inserido nesse processo de transformação e no qual se deve entender a nupcialidade e seu conseqüente desenho do mercado matrimonial.

A luz dessas reflexões, este artigo tem como objetivo analisar a dinâmica dos estados nupciais no Nordeste do Brasil, segundo sexo e idade para o ano 2000, *vis a vis* o comportamento destes estados observados no Brasil em geral, centrando atenção no casamento e no re-casamento; e, no espaço do mercado matrimonial.

2. METODOLOGIA

Aqui são descritas as fontes de dados utilizadas, os ajustes efetuados nos dados e a formalização do modelo da tábua de vida multiestado, expondo os avanços que esta técnica tem experimentado entre as décadas dos 80 e 90 do século passado. Este modelo viabiliza o cálculo das taxas e probabilidades de transição entre os estados conjugais, segundo sexo e idade.

2.1 Fontes de informação

As informações básicas para o estudo dos estados conjugais provem de duas fontes: i) Estatísticas do Registro Civil[2] e ii) Censo Demográfico[3] para os anos 1991 e 2000., respectivamente, realizadas pelo Ministério do Planejamento e Orçamento e o Instituto Brasileiro de Geografia e Estatístico (IBGE).

A informação que provem do Censo Demográfico refere-se ao volume da população segundo o estado conjugal, isto é, pessoas em situação de solteiro (a), casado (a), viúvo (a), separado judicialmente (a) e divorciado (a). Esta informação tem cobertura de todo o território nacional e determina o número de pessoas que na data da pesquisa (1/09/1991 e 01/08/2000) encontrava-se em determinado estado conjugal, sendo, portanto a informação de estoque. Esses dados de estado conjugal são obtidos a partir do questionário da amostra. No entanto, como o nível geográfico de análise é bem abrangente, Nordeste e Brasil como um todo, não há problema algum nestas informações, uma vez que, por exemplo, no Censo 2000 a amostra é representativa para áreas menores do que municípios, áreas estas que se constituem em aglomerados de setores censitários, tanto em áreas urbanas, quanto em áreas rurais.

No caso do Registro Civil é usada a informação de registros oficiais de casamento, separação judicial e divórcio, obtido de forma continua ao longo do ano, razão pela que esta informação é considerada de período.

Ficam fora da análise os casamentos realizados apenas no âmbito religioso, as uniões consensuais e as separações não judiciais.

É importante destacar no Brasil a existência de duas formas de separação: a separação judicial e o divorcio, cada um deles registrado separadamente, fato que segundo GOMES[4] estaria acarretando erros na avaliação dos montantes de casais separados, dado que o divórcio é uma confirmação da separação judicial já realizada. Por isso, analisa-se separadamente a separação judicial e o divorcio.

Nesse quadro, para estimar a transição de casado para divorciado, foi levado em conta apenas àqueles divórcios que viam somente do casamento, razão pela qual a legislação reconhece como o divórcio direto. O divórcio indireto, então, é aquele onde a dissolução do casamento passa primeiro pela separação judicial, para depois se concretizar o divórcio. Portanto, outra transição é realizada. Neste trabalho estas transições foram medidas separadamente.

2.2 Ajustes de dados e limitações

O modelo utilizado é baseado nas taxas de transição dos eventos que são calculadas através do quociente entre o número de eventos contabilizados no numerador e a população exposta no denominador. Por exemplo, a taxa de transição de solteiro para casado é calculada através do número de casamentos de pessoas solteiras (primeiro casamento), obtido no Registro Civil, dividido pela população solteira do ano em estudo coletada pelo Censo Demográfico. Aqui não se considera também a população em união consensual no denominador, pois muitas vezes quem se diz nesta situação, oficialmente, é alguém que dissolveu uma primeira união, seja por divórcio direto ou indireto, ou viuvez.

Desta forma, na ausência de uma forma precisa de inferir quanto da união consensual é composta por solteiros, divorciados ou viúvos, eliminamos esta população do denominador. Obviamente, isto implica uma limitação do modelo aplicado aos dados brasileiros. Contudo, como a idéia é analisar de maneira comparativa a dinâmica dos estados conjugais entre homens e mulheres e entre o Brasil e o Nordeste, esta limitação pode ser minimizada, pressupondo-se que a distribuição daqueles que vivem em união consensual entre solteiros, divorciados e viúvos é igual segundo o sexo e segundo os recortes regionais da análise.

Para o cálculo destas taxas foi preciso ajustar a população censitária por idade, sexo e estado conjugal para 1o de julho de cada ano em estudo. Isto se fez necessário dado que no numerador utilizam-se registros contínuos ao longo do ano, então, a população do denominador precisa levar em consideração o tempo vivido de cada habitante ao longo do ano (pessoas-ano). Neste sentido, a população em 1o de julho reflete o número médio de habitantes no ano.

Um outro ajuste foi realizado nos eventos referentes à transição de casado para viúvo. Não existe registro de viuvez, ou seja, quando uma pessoa fica viúva não é contabilizada em nenhum banco de dados. Entretanto, para calcular a taxa de transição entre casado e viúvo é necessário conhecer o número de pessoas que ficaram viúvas durante o ano. Para solucionar este problema utilizou-se o fato de que se uma pessoa fica viúva é porque alguém do sexo oposto faleceu, isto é, fez a transição de casado para morte. Então, para estimar o número de viuvez no ano em determinado sexo, aplica-se à estrutura etária de pessoas viúvas deste sexo, obtida no Censo, ao total de mortes entre casados do sexo oposto.

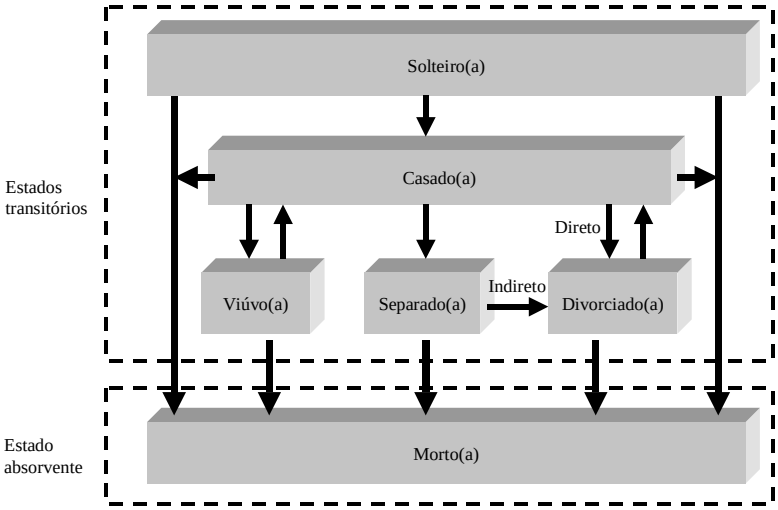
Quanto à qualidade dos dados, não se fez nenhuma correção nos dados referentes aos estados conjugais, uma vez que se propõe trabalhar com dados oficiais do Registro Civil. Neste sentido, deve-se chamar a atenção mais uma vez para o fato de que a análise deste trabalho fica no âmbito dos registros de casamento e divórcio oficiais. Já, no que tange aos registros de óbitos, utilizou-se um fator de correção de sub-registro proposto por SAWYER, et al[5].

2.3 Modelo nupcial

Willekens et al[6]; Espenshade & Eisenberg[7] desenvolvem a técnica da tábua de vida multiestado de maior precisão para observar o movimento das pessoas de uma coorte na passagem de um estado nupcial para outro. Nesse sentido, a tábua multiestado mostra que entre a vida e a morte existem estados da nupcialidade pelos quais a população transitaria, seja para: solteiro(a), casado(a), separado(a), viúvo(a) ou divorciado(a). Esta dinâmica nupcial reconhece a existência de um estado denominado absorvente (a morte) e os outros estados são considerados transitórios. Nos estados transitórios, as pessoas podem experimentar movimentos de um estado conjugal para outro, isto é, as pessoas solteiras podem se tornar casadas, as pessoas casadas podem se tornar viúvas ou divorciadas; as quais, por sua vez, podem voltar a casar.

Este modelo leva em consideração cinco estados nupciais nos quais se espera que as pessoas transitem em sua vida nupcial. Já para o caso brasileiro, este modelo sofre algumas variações, como conseqüência de normas legislativas vigentes no país. Estas variações são percebidas no caso particular da separação. Segundo a legislação reconhecem-se duas formas, isto é, a separação judicial e o divórcio. Estes dois estados funcionam de forma separada, configurando uma dinâmica diferente do que a tradicional. A nova configuração da dinâmica da nupcialidade para o caso brasileiro e aplicado à região Nordeste, pode ser apreciada no Esquema 1.

ESQUEMA 1
Modelo multiestado da nupcialidade



O Esquema 1 ilustra a dinâmica da nupcialidade, a partir de dois blocos: o primeiro corresponde aos estados transitórios e o segundo corresponde ao estado absorvente (morte).

O bloco que corresponde ao estado transitório é o mais importante do modelo, dado que ele mostra o circuito pelo qual a nupcialidade vai transitar. A trajetória da dinâmica nupcial inicia-se com o estado de solteiro (a), de onde as pessoas podem passar ao estado absorvente ou ao estado do casamento. O seguinte movimento corresponde à transição do casamento para o estado absorvente ou para o estado da viuvez ou para o estado da separação judicial ou, ainda, para o divórcio (direto). Quando estes três últimos estados nupciais não são absorvidos pela morte, eles podem voltar ao estado do casamento. Neste processo as pessoas quando se acham na situação da separação judicial, para voltar a casar passam pelo estado do divórcio (indireto).

O segundo bloco que corresponde ao estado absorvente revela o fim do movimento do modelo, aqui todos os estados denominados transitórios finalizam sua trajetória.

2.4 Formalização matemática do modelo de Tabua de Vida Multiestado aplicada à dinâmica nupcial.

As tábuas multiestado são desenvolvidas a partir da idéia Markoviana. Willekens et al[6] e SCHOEM[8] argumentam que as transições entre os vários estados de nupcialidade e a morte constituem os elementos de um processo de Markov, através do qual a probabilidade de transição entre dois estados, i e j, é tomada considerando só o estado atual do indivíduo. Isto implica numa limitação do modelo, pois seria esperado que uma pessoa que, por exemplo, era casada, separou-se e depois se casou novamente tenha probabilidade de separar desse último casamento diferente daquela que está no primeiro matrimônio. Contudo, os dados disponíveis não informam o histórico nupcial neste nível de detalhe.

As probabilidades de transição, segundo um processo Markoviano, é expressa da seguinte forma:

$$\pi_{ij}(x,t) = prob\{S(x+t) = j \setminus S(x) = i\} \qquad (1)$$

Onde πij (x, t) denota a probabilidade que uma pessoa do estado i à idade exata x esteja em j à idade exata $x+t$.

As tabelas multiestados aqui desenvolvidas baseiam-se em três condições estabelecidas pela teoria Markoviana: "tempo não homogêneo", "espaço finito", e processo de Markov de "tempo contínuo".

Segundo Schoem[8], a propriedade do tempo não homogêneo expressa que as forças de decremento podem variar dentro dos intervalos de idade. Espaço finito diz que o modelo contém $k+1$ estados onde k é um inteiro positivo maior que 1. O estado $(k+1)$ é dito absorvente (i.e., um estado onde não há nenhum decremento, por exemplo, a morte). Quanto aos outros K estados, pelo menos dois são comunicáveis, de maneira que haja incrementos (concorrentes) em pelo menos um caso. A propriedade de tempo contínuo permite usar cálculos entre intervalos de tempo discreto (ou idade) para descrever o comportamento do modelo.

Nesse quadro reflexivo, as tábuas de vida multiestado (TVME) podem ser organizadas a partir das probabilidades de transição definidas na equação (1), onde os $(k+1)$ por $(k+1)$ estados representam a matriz de probabilidades $\Pi(x, t)$. Desta forma, aplicando esta metodologia aos dados de nupcialidade utilizados, temos $K=5$ estados denominados transitórios, isto é: solteiro (a), casado (a), separado (a) judicialmente, divorciado (a) e viúvo (a). Assim, o sexto estado $(K+1)$, chamado absorvente, representa a morte. A matriz $\Pi(x, t)$ fica, portanto, expressa da seguinte forma:

$$\mathbf{\Pi(x,t)} = \begin{bmatrix} \pi_{11}(x,t)...\pi_{12}(x,t)...\pi_{13}(x,t)...\pi_{14}(x,t)...\pi_{15}(x,t)...\pi_{16}(x,t) \\ \pi_{21}(x,t)...\pi_{22}(x,t)...\pi_{23}(x,t)...\pi_{24}(x,t)...\pi_{25}(x,t)...\pi_{26}(x,t) \\ \pi_{31}(x,t)...\pi_{32}(x,t)...\pi_{33}(x,t)...\pi_{34}(x,t)...\pi_{35}(x,t)...\pi_{36}(x,t) \\ \pi_{41}(x,t)...\pi_{42}(x,t)...\pi_{43}(x,t)...\pi_{44}(x,t)...\pi_{45}(x,t)...\pi_{46}(x,t) \\ \pi_{51}(x,t)...\pi_{52}(x,t)...\pi_{53}(x,t)...\pi_{54}(x,t)...\pi_{55}(x,t)..\pi_{56}(x,t) \\ \pi_{61}(x,t)...\pi_{62}(x,t)...\pi_{63}(x,t)...\pi_{64}(x,t)...\pi_{65}(x,t)..\pi_{66}(x,t) \end{bmatrix}. \qquad (2)$$

Onde: 1 é solteiro (a), 2 casado (a), 3 separado (a) judicialmente, 4 divorciado (a), 5 viúvo, e 6 é a morte.Algumas considerações devem ser feitas sobre a matriz de probabilidades $\Pi(x,t)$. Primeiro, cada linha denota um espaço amostral condicionado ao estado inicial na idade x, ou seja, reflete todas as transições possíveis de uma pessoa do estado i à idade exata x. Assim, na primeira linha, por exemplo, tem-se todas as transições possíveis daquelas pessoas que eram solteiras à idade x. Deste modo, elas tem uma probabilidade $\pi_{11}(x,t)$ de chegarem à idade exata $x+t$ ainda solteiros(as), ou uma probabilidade $\pi_{12}(x,t)$ de estarem casados(as) aos $x+t$ anos, dado que era solteiro(a) aos x anos. Neste sentido, percebe-se que cada linha desta matriz de probabilidades deve somar 1. Além disso, a última linha, referente ao estado

absorvente, deve ser composta por 0 (zeros), exceto na última coluna, onde $\pi_{66}(x,t)$ deve ser igual a 1.

A probabilidade de transição, ou seja, a probabilidade de que uma pessoa do estado i à idade exata x chegue à idade x + t no estado j, é escrita em termos da função de sobrevivência da seguinte forma:

$$\pi_{ij}(x,t) = \frac{l_{ij}(x+t)}{l_i(x)} \qquad (3)$$

Onde:

i) li (x) representa o número de pessoas no estado i a idade exata x, e

ii) lij(x + t) representa o número de pessoas que na idade exata x estão no estado i, e na idade exata x + t pertencem ao estado j, ou seja, representa o número de pessoas que transitaram de i para j entre as idades exatas x e x + t.

iii) o lij corresponde aos fluxos brutos, pois eles representam o resultado de muitos movimentos individuais entre estados.

Com efeito, esta é a relação entre a probabilidade de transição e a função l(x) da tábua de vida, mas há um passo anterior a este, necessário para a operacionalização dos cálculos, que é escrever a probabilidade de transição como função das taxas de transição calculadas a partir dos dados. Para isto vamos lembrar de algumas funções básicas de uma tábua de vida. Primeiro, o número de pessoas à idade exata x + t - l(x + t), é encontrado da seguinte forma:

$$l(x+t) = l(x) - d(x,t) \qquad (4)$$
$$d(x,t) = l(x+t) - l(x)$$

Onde:

i) d(x, t) é o número de decrementos por morte entre as idades x e x + t numa tábua simples, e é o saldo entre os incrementos e os decrementos numa tábua multi-estado.

O número de pessoas entre as idades x e x + t é dado por:

$$L(x,t) = \int_0^t l(x+n)dn \qquad (5)$$

Que de forma numérica é calculado como segue:

$$L(x,t) = \frac{1}{2}t[l(x) + l(x+t)] \qquad (6)$$

Estas funções foram explicitadas porque são necessárias para se entender a transformação das taxas de transição para uma medida de probabilidade (π). Uma pressuposição que vamos adotar é que a taxa de transição da população real (M(x, t)[2]) é igual à taxa de transição da população estacionária da tábua multi-estado (m(x, t)), dada por:

$$m(x,t) = \frac{d(x,t)}{L(x,t)} \qquad (7)$$

Desta forma, se M(x, t) = m(x, t) e utilizando as equações (4) e (6) temos que:

$$M(x,t) = \frac{d(x,t)}{L(x,t)} = \frac{l(x) - l(x+t)}{\frac{1}{2}t[l(x) + l(x+t)]} \qquad (8)$$

Fazendo alguns exercícios algébricos em (8) chegamos facilmente à forma de cálculo para encontrar $l(x+t)$.

$$l(x+t) = l(x)\left[\frac{1 - \frac{1}{2}tM(x,t)}{1 + \frac{1}{2}tM(x,t)}\right] \qquad (9)$$

Com isso, comparando as fórmulas (3) e (9) conclui-se que a probabilidade de transição $\pi(x, t)$ é obtida, em função das taxas observadas, da seguinte forma:

$$\pi(x,t) = \left[\frac{1 - \frac{1}{2}tM(x,t)}{1 + \frac{1}{2}tM(x,t)}\right] \qquad (10)$$

Conseqüentemente, para operacionalizar os cálculos de uma TVME é necessário trabalhar de forma matricial, pois para cada grupo etário tem-se uma combinação de transições possíveis. Neste sentido, a matriz l(x) utilizada neste trabalho é da seguinte forma:

[2] A taxa de transição obtida a partir dos dados é da seguinte forma: $M_{ij}(x, t) = D_{ij}(x, t)/P_i(x, t)$, onde $D_{ij}(x, t)$ o número observado de transferência do estado *i* para o estado *j*, entre as idades x e $x+t$, e $P_i(x,t)$ representa a população no meio do ano em estudo observada no estado *i* entre as idades x e $x+t$.

$$\mathbf{l(x+t)} = \begin{bmatrix} l_{11}(x+t)...l_{12}(x+t)...l_{13}(x+t)...l_{14}(x+t)...l_{15}(x+t)...l_{16}(x+t) \\ l_{21}(x+t)...l_{22}(x+t)...l_{23}(x+t)...l_{24}(x+t)...l_{25}(x+t)...l_{26}(x+t) \\ l_{31}(x+t)...l_{32}(x+t)...l_{33}(x+t)...l_{34}(x+t)...l_{35}(x+t)...l_{36}(x+t) \\ l_{41}(x+t)...l_{42}(x+t)...l_{43}(x+t)...l_{44}(x+t)...l_{45}(x+t)...l_{46}(x+t) \\ l_{51}(x+t)...l_{52}(x+t)...l_{53}(x+t)...l_{54}(x+t)...l_{55}(x+t)...l_{56}(x+t) \\ l_{61}(x+t)...l_{62}(x+t)...l_{63}(x+t)...l_{64}(x+t)...l_{65}(x+t)..l_{66}(x+t) \end{bmatrix} \quad (11)$$

Onde, lij (x + t) representa o número de pessoas que na idade exata x estavam no estado i, e na idade exata x + t pertencem ao estado j. Os índices recebem a mesma classificação dada na matriz Π(x, t).

Com isso, teremos todas as funções da TVME de nupcialidade em forma matricial – l(x, t); L(x, t); T(x, t); e (x, t). As fórmulas matriciais são as seguintes:

$$\mathbf{l(x+t) = l(x)}\left[\mathbf{I} - \frac{1}{2}t\mathbf{M(x,t)}\right]\left[\mathbf{I} + \frac{1}{2}t\mathbf{M(x,t)}\right]^{-1} \quad (12)$$

$$\mathbf{\Pi(x,t) = l_*(x)^{-1} l(x+t)} \quad (13)$$

onde, $\mathbf{l\ (x)}$ é uma matriz diagonal cujos elementos são a soma de cada coluna de $\mathbf{l(x+t)}$, isto é, em cada diagonal enumera-se o total da população de cada estado, por exemplo, o segundo valor da diagonal de $\mathbf{l_*(x)}$ significa o número de pessoas casadas à idade exata x, independente do estado anterior (solteira, separada, divorciada ou viúva).

A probabilidade de transição em função da matriz de taxas (**M(x, t)**)[3] fica como segue:

$$\Pi(\mathbf{x},\mathbf{t}) = \left[\mathbf{I} - \frac{1}{2}t\mathbf{M}(\mathbf{x},\mathbf{t})\right]\left[\mathbf{I} + \frac{1}{2}t\mathbf{M}(\mathbf{x},\mathbf{t})\right]^{-1} \qquad (14)$$

A matriz **L(x, t)** é obtida da forma normal, fazendo a média aritmética simples entre os valores l_{ij} (x) e l_{ij} (x+t).

$$\mathbf{L}(\mathbf{x},\mathbf{t}) = \frac{1}{2}t\left[\mathbf{l}(\mathbf{x}) + \mathbf{l}(\mathbf{x}+\mathbf{t})\right] \qquad (15)$$

O valor de pessoas-ano para o intervalo de idade 85 anos e mais, pode ser estimado usando:

$$\mathbf{L}(\mathbf{85},\infty) = \mathbf{l}(\mathbf{85})\mathbf{M}^{-1}(\mathbf{85},\infty) \qquad (16)$$

Contudo, para o cálculo de (16) é necessário reduzir a matriz M(x, t) por que na sua dimensão original ela é singular e, portanto, não possui inversa. Isto ocorre, devido ao estado absorvente (morte), a última linha desta matriz é formada de zeros, o que implica que o determinante de M(x, t) é igual a zero. Assim, para operacionalizar o cálculo de $\mathbf{L(85,\infty)}$ retira-se a linha e a coluna referente a morte, com isso, a matriz M(x,t) que era 6x6 passa a ser 5x5. Isto implica que, a partir desse ponto, os cálculos das funções restantes são feitos utilizando apenas os 5 estado de nupcialidade.

---

[3] Segundo FREIRE & AGUIRRE[9] a matriz de taxas de transição observadas deve ser da forma abaixo, de modo que a soma das linhas seja zero:

$$\mathbf{M}(\mathbf{x},\mathbf{t}) = \begin{bmatrix}
\sum_{j\neq i} M_{1j}(x,t)... & -M_{12}(x,t)... & -M_{13}(x,t)... & -M_{14}(x,t)... & -M_{15}(x,t)... & -M_{16}(x,t) \\
-M_{21}(x,t)... & \sum_{j\neq i} M_{2j}(x,t)... & -M_{23}(x,t)... & -M_{24}(x,t)... & -M_{25}(x,t)... & -M_{26}(x,t) \\
-M_{31}(x,t)... & -M_{32}(x,t)... & \sum_{j\neq i} M_{3j}(x,t)... & -M_{34}(x,t)... & -M_{35}(x,t)... & -M_{36}(x,t) \\
-M_{41}(x,t)... & -M_{42}(x,t)... & -M_{43}(x,t)... & \sum_{j\neq i} M_{4j}(x,t)... & -M_{45}(x,t)... & -M_{46}(x,t) \\
-M_{51}(x,t)... & -M_{52}(x,t)... & -M_{53}(x,t)... & -M_{54}(x,t)... & \sum_{j\neq i} M_{5j}(x,t)... & -M_{56}(x,t) \\
-M_{61}(x,t)... & -M_{62}(x,t)... & -M_{63}(x,t)... & -M_{64}(x,t)... & -M_{65}(x,t)... & \sum_{j\neq i} M_{6j}(x,t)
\end{bmatrix}$$

Onde, os índices *i* e *j* seguem a mesma classificação descrita na matriz $\mathbf{\Pi(x, t)}$.

Para finalizar a tábua de vida multi-estado para nupcialidade precisa-se das funções T(x) e(x). A primeira é obtida da forma usual, ou seja, soma-se todos os valores de Lij (x, t) a partir da idade x e, portanto, não envolve operação matricial. Então, T(x) fica da seguinte forma:

$$T_{ij}(x) = \sum_{y=x}^{85} L_{ij}(y,t) \qquad (17)$$

No que tange a esperança de vida, pode-se dizer que num estudo multi-estado existe mais de uma variante desta medida. Neste trabalho, optou-se por utilizar o número médio de anos vividos por uma pessoa em cada estado conjugal. A fórmula utilizada foi a seguinte:

$$e_j(x) = \frac{T_j(x)}{l(x)} \qquad (18)$$

Onde, Tj(x) e l(x) representa a soma nas colunas das respectivas matrizes.

Há ainda uma outra forma de calcular as funções L(x), T(x) e(x). Observando as fórmulas (14), (15), (16) nota-se que o que está sendo feito é a mesma coisa de tomar separadamente as colunas da matriz l(x) e calcular uma tábua de vida simples separada para cada estado conjugal. Com isso, têm-se as funções l(x), L(x), T(x) e(x) para solteiro (a), casado (a), separado (a), divorciado (a) e viúvo (a).

Uma alternativa de cálculo, talvez até mais fácil de ser implementada, é utilizar as funções matriciais até obter a matriz l(x) e a partir dela, trabalhar de forma individual com cada estado conjugal.

3. RESULTADOS

3.1 Análise múltipla dos estados nupciais

Segundo a ótica demográfica, o mercado matrimonial, segundo Greene & Rao[10], estaria relacionado a uma maior ou menor oferta de homens ou mulheres no mercado de casamento e determinado pela escassez de um sexo ou outro na faixa etária em que, geralmente, acontecem os casamentos, fato que estaria influindo não apenas na constituição das uniões, mas também na dinâmica nupcial.

Nesse sentido, uma aproximação preliminar do mercado matrimonial a partir dos estados nupciais para o Nordeste expõe-se na Tabela 1, onde se observa a distribuição percentual da população de 15 anos e mais por sexo, segundo o estado conjugal entre 1980 e 2000. As uniões consensuais, que em 1980 representavam 8,5% nos homens e 7,8% nas mulheres, vinte anos mais tarde em 2000, estes valores sobem para 18,6% nos homens e 17,3% nas mulheres (Tabela 1). Em que pese a melhoria na qualidade da informação coletada pelo IBGE, desde o Censo de 1980 até 2000, tais avanços no percentual de uniões consensuais certamente refletem mudança estrutural na composição familiar.

Os resultados mostram que neste período, a proporção da população no estado de solteiros tem aumentado sistematicamente em ambos os sexos, já a de casados, tanto em homens como em mulheres, experimentou tendência inversa. O aumento das uniões consensuais contribui para esta tendência que, aliado ao fato do ingresso ao matrimônio ocorrer cada vez em idades mais maduras, demonstra uma mudança estrutural na dinâmica nupcial no Brasil, também acompanhada pelo Nordeste. No Brasil em 1991 a idade média do primeiro casamento era de 25,7 para mulheres e 27,9 para os homens. Em 2000, este números mudaram para 28,3 e 31,2, para mulheres e homens respectivamente. Já no caso do Nordeste a idade de ingresso ao casamento para o ano 2000 foi estimada em 31,8 para os homens e 28,9 para as mulheres.

Tabela 1 Nordeste: Distribuição percentual da população de 15 anos e mais,por sexo segundo o estado conjugal

Ano	1980		1991		2000	
Estado Conjugal	H	M	H	M	H	M
Solteiros	37,0	32,4	37,3	31,0	39,7	35,8
Casados	49,9	46,1	45,3	41,7	37,6	35,0
U.Consensual	8,5	7,8	12,5	11,5	18,6	17,3
Sep.não-judicial	1,3	3,7	2,1	5,7	-	-
Divor/Sep.Judicial	0,1	0,2	0,4	0,9	1,1	2,3
Viúvos	1,8	7,8	1,6	7,7	1,4	6,4

Fonte: Elaboração própria com base nos Censos Demográficos: 1980, 1991[11-12] e 2000, IBGE.

Na mesma Tabela, concomitantemente, pode-se verificar que as mudanças nas proporções do estado do divórcio e separação são mais marcantes nas mulheres (de 0,2%, em 1980, passa a 2,4%, em 2000), do que nos homens (que varia de 0,1% a 2,0%, no mesmo período). O estado da viuvez tem comportamento sem grandes oscilações no período considerado, seja nas mulheres, seja nos homens. Contudo, chama a atenção o grande diferencial no percentual do estado da viuvez segundo o sexo. Enquanto o percentual de mulheres no estado de viuvez gira em torno de 7%, os homens no estado de viuvez não atingiram 2% ao longo do período estudado. Uma hipótese que estaria explicando este diferencial, como se argumenta acima, pode ser imputada à sobremortalidade masculina, principalmente devido às mortes por causas externas que afetam os homens adultos jovens. Desta forma, se além do homem ter uma probabilidade de morte maior do que a mulher, ele ainda em geral é o mais velho do casal, conseqüentemente a chance dele ser viúvo será tanto mais baixa quanto menor for a mortalidade feminina com relação à masculina. Além disso, conforme apresentado em páginas seguintes, a probabilidade de re-casamento dos homens é maior do que nas mulheres. Embora tenham ocorrido estas mudanças na dinâmica nupcial, o casamento formal ainda continua a forma de união que tanto homens quanto mulheres privilegiam.

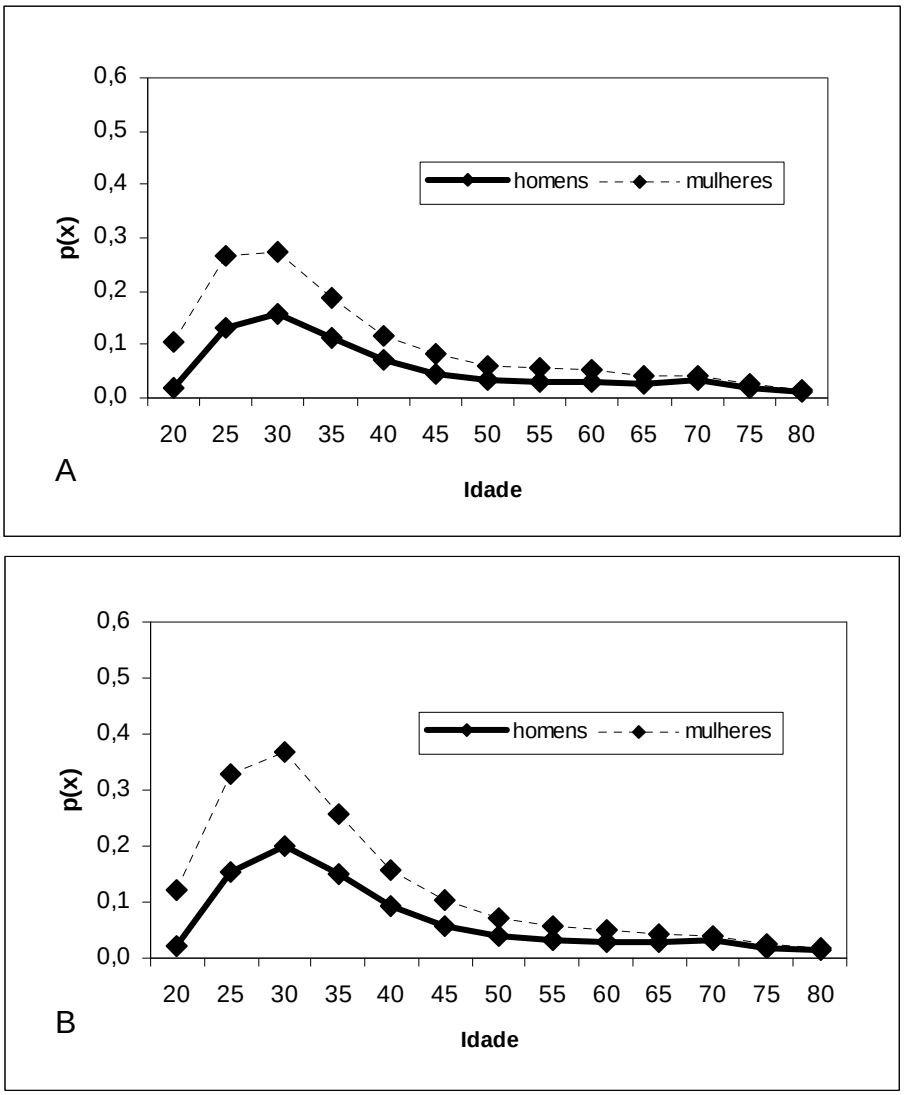
3.2 O primeiro casamento

As Figuras 1A e 1B expõem a concretização do mercado matrimonial através das probabilidades de transição de solteiro(a) para casado(a) por grupos de idade em 2000 para o Nordeste e o Brasil como um todo. Demograficamente, estes gráficos revelam para ambos os sexos, a trajetória curvilínea côncava das probabilidades do primeiro

casamento com ritmo acelerado até os 30 anos para, a partir daí, cair vertiginosamente até os 40 anos.

Observa-se também, que a probabilidade de ingressar ao primeiro matrimônio é mais alta nas mulheres do que nos homens de forma marcante até os 30 anos, sendo que o diferencial por sexo apresenta-se levemente maior no Brasil. Esse comportamento pode ser atribuído, por um lado, ao fato das mulheres ingressarem ao matrimônio a idades mais jovens em relação aos homens, por outro, à sobremortalidade masculina, particularmente aquelas relacionadas às causas de mortes violentas (homicídios, acidentes de trânsito, acidente de trabalho, etc.) a idades jovens. Lembre que na metodologia foi explicado que as probabilidades somam 1 (um) na linha da matriz de transição, ou seja, a probabilidade de uma pessoa que na idade x está solteiro chegar à idade $x+t$ casado vai concorrer com a probabilidade de permanecer solteiro ou de vir a morrer. Portanto, se nos homens a probabilidade de morte é maior que nas mulheres, isso terá influência direta na probabilidade de transição do estado de solteiro para casado.

As Figuras 1A e 1B, também, podem ser interpretadas como a distribuição etária do risco do primeiro casamento. Nesse sentido, observa-se que não há grande diferença na distribuição etária do primeiro casamento quando se compara o Nordeste com o Brasil. Por se tratar da primeira união formal, para ambos os casos as maiores probabilidades estão nas primeiras idades (abaixo dos 30 anos). Contudo, se não há diferença de padrão, há uma pequena diferença de nível, revelando que a probabilidade de contrair o primeiro casamento oficial no Nordeste é ligeiramente menor do que no Brasil, mesmo nas idades mais jovens. Este resultado ratifica os números apresentados na Tabela 1, na qual 18,6% dos homens e 17,3% das mulheres nordestinas estão em uniões livres, ditas consensuais, enquanto que no Brasil estes valores são 16,7% e 15,7%, respectivamente para homens e mulheres.



Fonte: Registro Civil/Censo-IBGE(2000)
Figura 1: Probabilidade do primeiro casamento por sexo e grupo de idade (A) Região Nordeste 2000 e (B) Brasil 2000

As Figuras 1A e 1B mostram a estrutura do ingresso ao mercado matrimonial e sua concretização estende-se muito além da condição apenas demográfica. Encontra, também, suporte analítico e explicativo nas transformações do ambiente macro, ou seja, homens e mulheres sofrem o impacto da onda de mudanças socioeconômicas que o país tem experimentado desde a década de 80. No ambiente micro, a concretização do mercado matrimonial estaria determinada no campo social e econômico, ou seja, no seio da condição socioeconômica – emprego, renda, educação - em que se encontram homens e mulheres.

3.3 Re-casamento

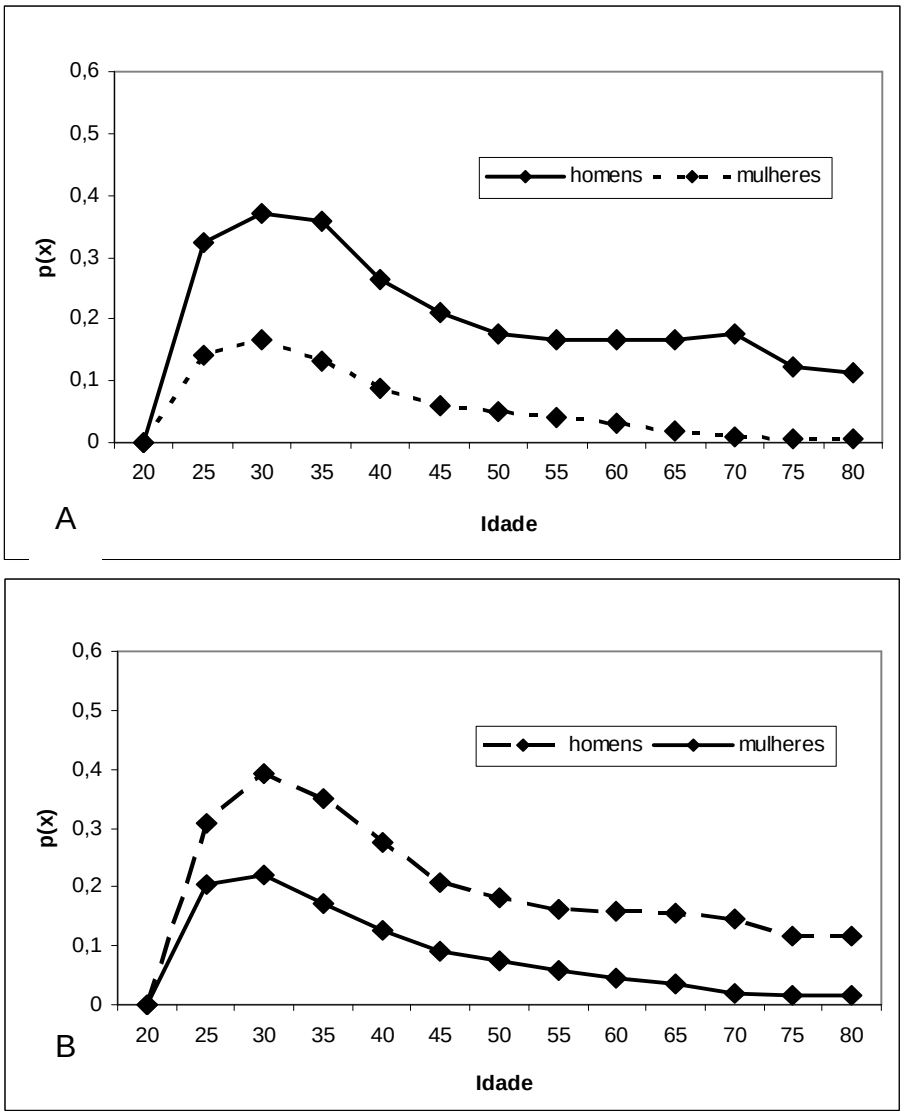
Na hipótese extrema dos casais permanecerem nas primeiras núpcias até ficarem absorvidos pela morte, a dinâmica nupcial teria apenas um movimento, o passo de solteiro(a) para casado(a). Essa hipótese é derrubada, com a ruptura das primeiras núpcias e a viúves a idades jovens. Estas singularidades da nupcialidade abrem espaço para que separados(as) e viúvos(as) não apenas voltem, mas também, expandam o mercado matrimonial tornando-se determinantes para a dinâmica do mesmo. Assim, só as pessoas que experimentaram o divórcio ou a viuvez podem transitar novamente ao estado de casamento. Nesse contexto, as Figuras 2A e 2B apresentam as probabilidades de re-casamentos de divorciados(as) por sexo e idade para o Nordeste e Brasil, em 2000.

Estas Figuras 2A e 2B revelam comportamento bastante semelhante no que se refere ao re-casamento de pessoas divorciadas. Observa-se que o diferencial por sexo e idade das probabilidades de re-casamento no Nordeste segue o mesmo padrão do Brasil, onde o reingresso de uma mulher divorciada a um novo casamento oficial é bem menos provável que do re-casamento de um homem divorciado.

Particularizando as faixas etárias de 30 a 34 anos, que possuem os valores mais altos, a probabilidade do re-casamento de divorciados no Nordeste é de 0,372 e 0,166 respectivamente, para homens e mulheres, menor do que a registrada para a média nacional nesta mesma faixa etária.

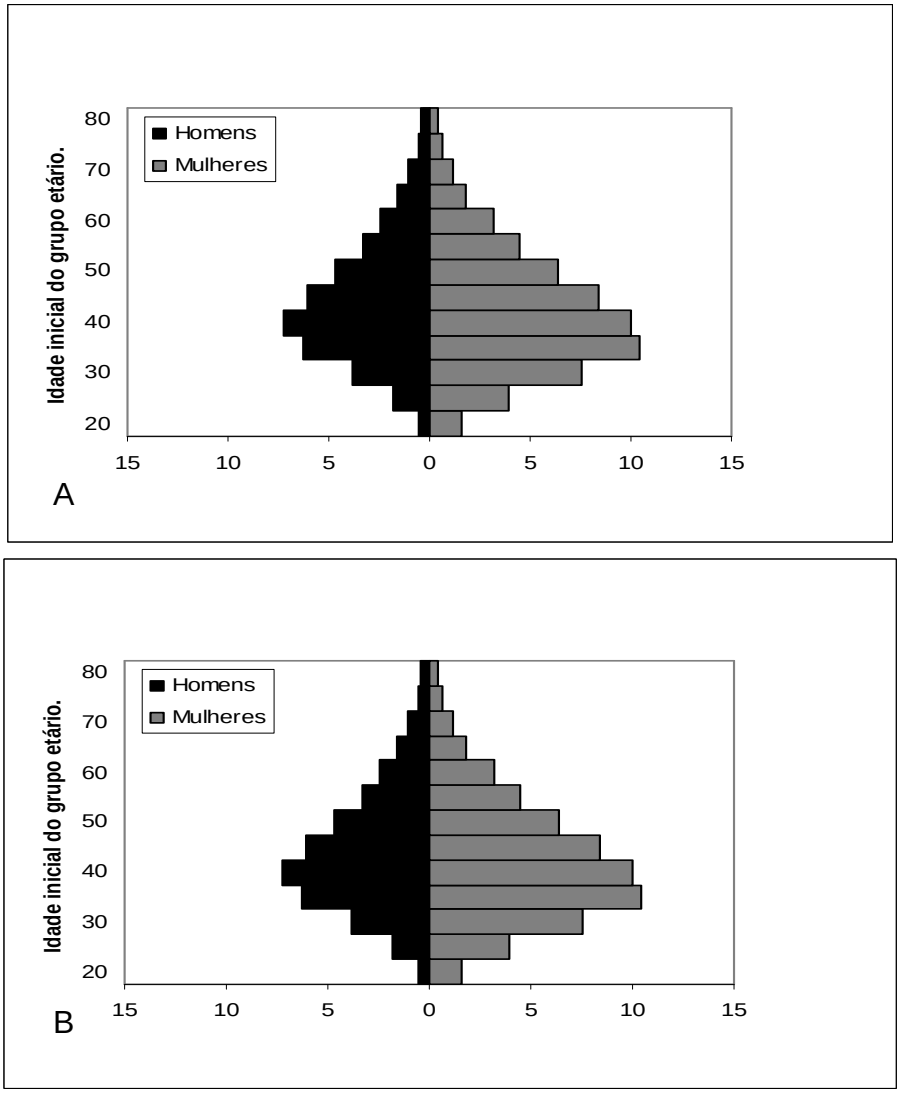
- O desequilíbrio das probabilidades entre os sexos pode ser explicado pelos obstáculos que são postos durante a busca de um novo parceiro(a), que parecem ser maiores para as mulheres do que para os homens. Segundo Medeiros & Osório[1], 14% dos arranjos familiares no Brasil, em 1998, eram do tipo chefiado por mulher, sem cônjuge e com filhos. Enquanto isso, os arranjos familiares formados apenas por homens com filho representam 2%. Os valores das probabilidades de re-casamento entre divorciados indicam que essa concentração pode ser ainda maior no Nordeste.
- Além disso, devido à sobremortalidade masculina e à maior probabilidade de re-casamento dos homens, quanto mais avançada for a idade com que

a mulher dissolve seu casamento menor será o "estoque" de homens no mercado com a mesma faixa etária.



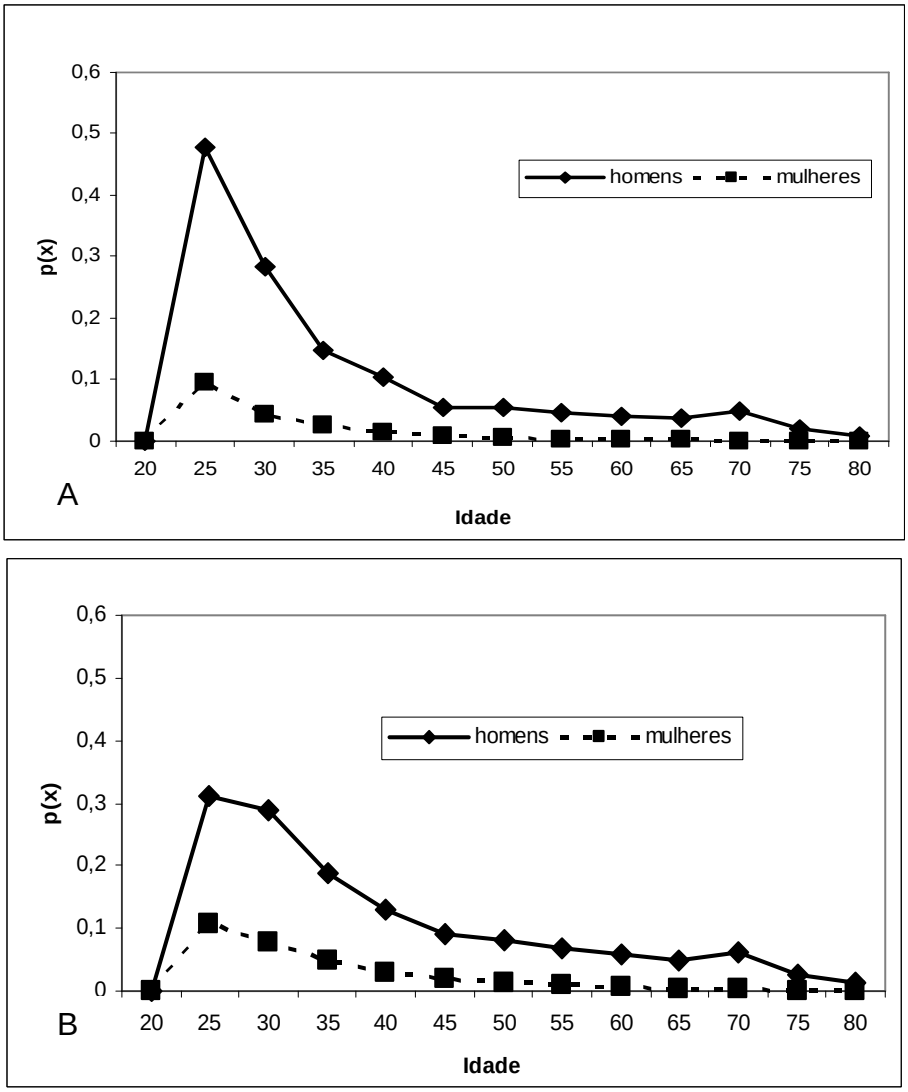
Fonte: Registro Civil/Censo-IBGE(2000)
Figura 2: Probabilidade de recasamento de divorciados por sexo e grupo de idade (A) Região Nordeste 2000 e (B) Brasil 2000

Chances escassas de re-casamento entre pessoas divorciadas acabam por manter um estoque de população divorciada, principalmente mulheres. É o que revela as pirâmides etárias da população divorciada para o Nordeste e o Brasil (Figuras 3A e 3B).



Fonte: Registro Civil/Censo-IBGE(2000)
Figura 3: Pirâmide etária da população divorciada a partir dos 20 anos (A) - Região Nordeste 2000 e (B) Brasil 2000

Além do re-casamento entre divorciados, as pessoas viúvas, por lei, podem voltar a contrair casamento civil. As Figuras 4A e 4B mostram, assim como foi visto para o caso da população divorciada, que é bem maior a chance de um viúvo voltar a casar do que uma viúva.



Fonte: Registro Civil/Censo-IBGE(2000)
Figura 4: Probabilidade de recasamento de viúvos por sexo e grupo de idade (A) Região Nordeste 2000 e (B) Brasil 2000.

As Figuras 2 e 4 mostram a estrutura do re-ingresso ao mercado matrimonial e revelam o abismo que separa as probabilidades do re-casamento entre homens e mulheres depois de completar os 25 anos. São indícios reveladores para inferir que sua concretização estende-se muito além da condição apenas demográfica, e suas explicações podem se achar no campo social e econômico em que se encontram homens e mulheres.

4. DISCUSSÃO

O propósito deste trabalho foi mostrar a dinâmica das transições dos componentes da nupcialidade (matrimônio, separação judicial, divórcio, viuvez e o re-casamento) que experimentou o Nordeste com relação ao Brasil, em 2000. Os resultados apresentados neste artigo, além de gerar conclusões, instigam a levantar hipóteses, sobre as quais se discorre mais adiante.

A dinâmica nupcial é importante como objeto de pesquisa, uma vez que a forma de formação de famílias revela, de certa maneira, as regras que regem a sociedade. Além disso, o estudo da nupcialidade é muito importante como fator interveniente que influencia diretamente outros aspectos demográficos de uma população. Um exemplo disso é a reprodução, que é influenciada pela postergação da idade de ingresso ao casamento. Conforme relatado anteriormente, a idade média do primeiro casamento aumentou no Brasil e Região do Nordeste, tanto para as mulheres quanto para os homens. Segundo Davis e Blake[13], certamente a idade ao casar é associada à reprodução, que funciona como um dos mecanismos de regulação do tamanho de família. Quando as mulheres casavam em grande proporção a idades jovens, presumia-se uma fecundidade elevada. Já quando a proporção de casamento é menor e a idades mais maduras, a fecundidade tende a ser menor. Na reflexão de Notestein[14], citado por Coale[15], as idéias sobre o tamanho da descendência foram mudando e resulta impossível ser preciso aos vários fatores causais que levaram ao novo ideal da família pequena. Todavia, segundo Lazo[16] uma diversidade de trabalhos apontam para a relação entre a idade da primeira união e variáveis sociais, econômicas, demográficas ou culturais no Brasil.

A vida urbana privou a família de muitas funções na produção, consumo, recreação e educação. Além disso, as mulheres obtiveram maior independência das obrigações domésticas e passaram a desempenhar novos papéis econômicos menos compatíveis com a procriação. Sobre múltiplas pressões, inclusive por meio de difusão dos novos padrões culturais de configuração familiar por meio da televisão, rádio e deslocamentos populacionais inter-regionais, idéias e crenças antigas começaram a enfraquecer-se, para dar passo a uma nova idéia de família com número reduzido de filhos e onde a dinâmica da nupcialidade tem desempenhado papel importante.

Na visão sociológica, Srinivasan[17] argumenta, a dinâmica da nupcialidade vai estar atrelada às transformações de ordem social e econômica como conseqüência de uma intensa modernização da sociedade ocidental. Este fato estaria levando as mulheres a se inserir cada vez com maior intensidade no mercado de trabalho e aumentar seu nível de educação formal, dimensões que levam a gerar uma nova ordem das relações sociais, dando lugar à formação de um novo paradigma sobre o tamanho de família, na qual a fecundidade passa a ser prioridade de segunda ordem e aspirações de realização pessoal são incorporadas como prioritárias. Esta visão pode ser sustentada por Beltrão[18] quando diz que as mudanças sociais ocorridas no matrimônio e na família de hoje seria o resultado do confronto entre dois tipos de família que refletem em seus traços características da família tradicional e da família de hoje (moderna).

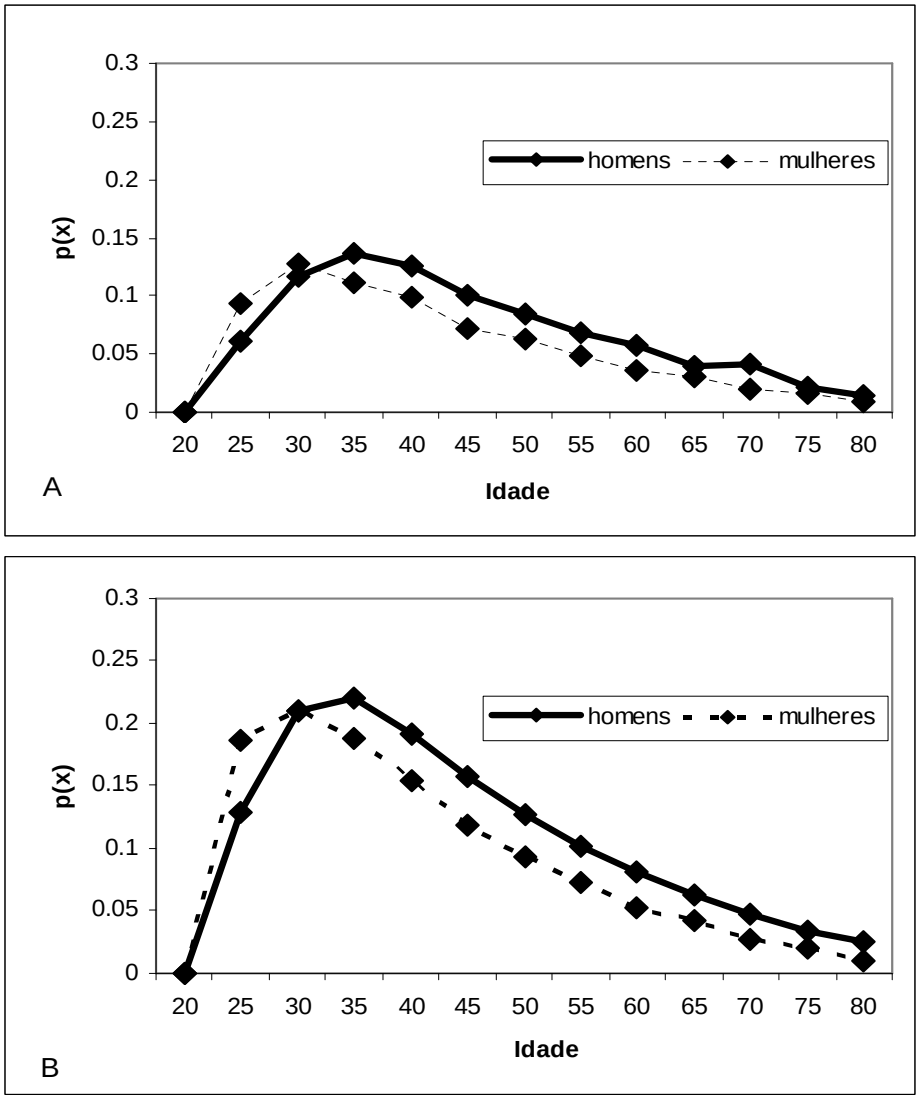
Todavia, Galloway (1988), citado por Moreira[19], argumenta que o aumento da mortalidade adulta poderia ter um efeito contrário, ou seja, diminuir as expectativas de novas núpcias seja adiando os casamentos temporariamente, ou mesmo definitivamente. Com tudo, estas duas singularidades, estariam sendo responsáveis pela transformação da dinâmica dos estados conjugais e gerando novas estruturas familiares. Assim, hoje não existe somente a família tradicional de pais e filhos, existem também as famílias compostas de mulheres sem marido com filhos, casais sem filhos e pessoas vivendo sozinhas.

A região nordeste do Brasil não ficou fora destas transformações ocorridas no padrão do tamanho da família, de 7,5 filhos por mulher em 1970 cai para 2,6 filhos por mulher no ano 2000. Esta transição vertiginosa experimentada pelo comportamento reprodutivo das mulheres nordestinas, certamente tem relação com o padrão de formação familiar na região e com a dinâmica dos estados da nupcialidade, particularmente com os re-casamentos.

A probabilidade de contrair o primeiro casamento no Nordeste é menor do que a média do Brasil, mesmo nas idades mais jovens. Este resultado pode indicar índice maior de uniões consensuais e/ou maior celibato no Nordeste do que a média do Brasil.

Em relação ao re-casamento, antes é importante tecer alguns comentários e revelar alguns resultados sobre a dissolução do casamento como, por exemplo, através do divórcio. Os resultados referentes as probabilidades de divórcio, tanto em homens

como em mulheres, mostram um padrão de crescimento acentuado a partir dos 25 anos até os 30 anos, no caso das mulheres e até 35 anos no caso dos homens. A partir dessas idades inicia-se queda acelerada sem mostrar grandes diferenças por sexo.



Fonte: Registro Civil/Censo-IBGE(2000)
Figura 5: Probabilidade de divórcio por sexo e grupo de idade
(A) Região Nordeste 2000 e (B) Brasil 2000.

Sobre este fenômeno, não existem no caso brasileiro ainda estudos em profundidade que possam orientar sobre os determinantes da separação tanto em mulheres quanto nos homens. Certamente fatores sócio-econômicos e demográficos estariam explicando com maior força os determinantes deste fenômeno não apenas no caso do Brasil, mas em particular no caso da Região do Nordeste.

Nesse sentido, os estudos que têm abordado as causas da dissolução concluem que, nos países onde o nível de divórcio é baixo, os segmentos com maior nível de educação são os que mostram taxas de separação mais elevadas. Isso seria o resultado de atitudes mais liberais e modernas em relação à instituição matrimonial, particularmente entre as mulheres mais educadas (JALOVAARA[20], CLARKE[21], HOULE et al[22], KIERNAN[23], VAZ[24]).

O ingresso das mulheres ao mercado de trabalho pode influenciar na separação matrimonial na medida em que modificou as bases da divisão sexual do trabalho. A importância da relação entre trabalho feminino, autonomia econômica e divórcio, foi adotada pela grande maioria dos trabalhos que analisam as mudanças na estabilidade dos casamentos ocorridas desde finais dos anos sessenta. No plano individual, os efeitos do trabalho feminino exprimiram resultados contraditórios por um lado, remetendo ao fato de que a mulher se torne menos dependente economicamente acaba por aumentar a probabilidade de dissolução, na medida em que se reduzem os benefícios do casamento (efeito de independência). Por outro lado, outros autores observaram que o fato de a mulher gerar rendimentos significa um alívio para a situação econômica da família e, portanto, pode contribuir para a estabilidade do casamento. (CHERLIN[25], GOODE[26], RUGGLES[27], VAZ[24])

Quanto ao re-casamento, observa-se que é mais provável um homem voltar a casar do que uma mulher e, isso é mais acentuado no Nordeste do que no Brasil como um todo.

Cortez, Lazo e Magalhães[28] encontram similares resultados, mas não avançam nas possíveis explicações que estariam por traz desse indicador, principalmente entre as faixas etárias de 25 a 40 anos e em particular das mulheres. A hipótese aqui levantada, com relação a este estado nupcial, é que fatores não apenas de ordem demográfico, mas também, de ordem socioeconômica e cultural poderiam explicar esses diferenciais. Esta hipótese esta baseada nas argumentações de Vaz[24] que diz que o re-casamento envolve implicações teóricas diferentes das que se aplicam à primeira união. Segundo Sweeney[29], o fato de ter passado por uma união anterior afeta a valoração do re-casamento ao ponto de depender sua concretização da avaliação da experiência passada. Dado que o fenômeno do re-casamento se dá com muito menor freqüência que

a primeira união, principalmente entre as mulheres, possivelmente como conseqüência de dar prioridade a suas aspirações individuais de ordem profissional, de cidadania e de processos de seleção que subjazem ao re-casamento. Finalmente, porque na medida em que o re-casamento tem lugar numa etapa mais tardia do ciclo de vida, as pessoas contam com maiores recursos econômicos, uma carreira laboral mais sólida e uma trajetória ou projeto reprodutivo consolidado. Teoricamente, esses fatores contribuem para que a decisão de entrar numa nova união dependa fortemente das experiências reprodutiva e de trabalho passadas.

Apesar do modelo de Tábuas de Vida Multiestado adotado seguir o arcabouço teórico Markoviano, onde uma transição de um estado para o outro depende apenas do estado anterior, foi possível perceber a diferença nas probabilidades de contrair o primeiro casamento com relação ao re-casamento, sobretudo o diferencial na probabilidade entre os sexos, nas idades de 20 a 35 anos. Os resultados apontaram que probabilidade de casar pela primeira vez é levemente maior nas mulheres com relação aos homens, já o re-casamento é muito mais provável por parte do homem do que entre as mulheres e este é um padrão que se apresenta no Nordeste e no Brasil como um todo.

AGRADECIMENTOS

O projeto original foi desenvolvido com o apoio do CNPq e FAPERN.

Referências Bibliográficas

1. **Medeiros, M & Osório, R.G. Mudanças na composição dos arranjos domiciliares no Brasil – 1978 a 1998. Revista Brasileira de Estudos de População. ABEP, 2000. Vol 17, nº 1.**
2. **ESTATÍSTICA DO REGISTRO CIVIL – RC. Ministerio do Planejamento e Orçamento. Fundação Instituto Brasileira de Geografia e Estatística – IBGE. Diretoria de Pesquisas Departamento de População, 2000, v. 18-2000.**
3. **CENSO DEMOGRÁFICO – CD. Ministério do Planejamento e Orçamento. Fundação Instituto Brasileira de Geografia e Estatística – IBGE. No 1. 2000.**
4. **GOMES, E. U. S. Dissolução da Sociedade e do Vínculo Conjugal.In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 8, 1992. Anais..., Brasília: ABEP, 1992 Anais, v.1, p 243-261.**
5. **SAWYER, D.O. ET. Al. Projeção populacional, por sexo e grupos de idades qüinqüenais, das unidades da federação: Brasil, 1990-2020. CEDEPLAR. Belo Horizonte, 1999.**
6. **Willekens, F.J., Shah, I., Shah, J.M., Ramachadran, P. Multi-State Analysis of Marital Status Life Tables: Theory and Application. In: Readings in Population Research Methodology. Nuptiality, Migration, Household, and Family Research., v. 4., 1993, p.13-30/13-37. Published for United Nations Population Fund by Social Development Center Chicago, Illinois.**
7. **Espenshade, T. J., Eisenberg B. R. Life Course Analysis and Multi-State Demography: An Application to Marriage, Divorce, and Remarriage. In: Readings in Population Research Methodology. Nuptiality, Migration, Household, and Family Research. v. 4., 1993, p.13-19/13-29. Published for United Nations Population Fund by Social Development Center Chicago, Illinois.**
8. **Schoem, R. The Multistate Life Table. In: Modeling multigroup populations. New York: Plenum Press, 1988, p. 63-105.**
9. **Freire, F.H.M., Aguirre, M.A.C. Dinâmica entre os estados conjugais da população brasileira: uma aplicação das tábuas de vida multiestados. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 12, Caxambu, 2000. Anais... Belo Horizonte: ABEP, 2000. (CD-ROM).**
10. **Greene, M. E. Rao, V. A compressão do mercado matrimonial e o aumento das uniões consensuais no Brasil. Revista brasileira de estudos populacionais. Campinas, v.9, n.2, 1992.**
11. **CENSO DEMOGRÁFICO – CD. Ministério do Planejamento e Orçamento. Fundação Instituto Brasileira de Geografia e Estatística – IBGE. No 1. 1980.**
12. **CENSO DEMOGRÁFICO – CD. Ministério do Planejamento e Orçamento. Fundação Instituto Brasileira de Geografia e Estatística – IBGE. No 1. 1991.**
13. **Davis, K.; Blake, J. Fatores Sociologicos de la Fecundidad. 1 ed. México: Gráfica Panamericana, 1956. 154 p.**
14. **Notestein, F. Population – The long View. In: Schultz, T. (ed) Food for the Word, Chicago, University of Chicago Press, 1945.**
15. **Coale, A. J. The demographic transition: a summary, some lessons and some observations. In: CHO, L., KATUMASA, K. (eds.). Fertility Transition of East-Asian Populations. Honolulu: University Press of Haway, 1979. Cap2: 9-23.**

16. **Lazo, A.C.G.V. Nupcialidade em São Paulo: um estudo por corte e coorte. Tese de Doutorado apresentada ao Departamento de Ciencias Sociais do Instituto de Filosofia e Ciencias Humanas, Universidade Estadual de Campinas. 1991.**
17. **Srivivasan, K. Basic demography techniques and applications: Nupcialidade, 1998. p.121-126.**
18. **Beltrão P. Ciência da População Análise e Teoria. Nupcialidade, 1973, p.85-95.**
19. **Moreira, M. R. Aspectos Teóricos dos Efeitos das Flutuações Econômicas sobre as Variáveis Demográficas.Rev.Univ.Rural, Ser. Ciênc.Humanas, Vol.23(2): 161-168 julho./dez.2001, UFRRJ. HILL, K. and Palloni, A., Demographic reponses to economic shocks: The case of Latin America. IUSSP, The Peopling of the Americas, Vera Cruz, México, v.3, 1992, p.411-438.**
20. **JALOVAARA, M. Socio-economic status and divorce in first marriages in Finland 1991-1993. Population Studies, London, v.55, n.2, p.119-133, 2001.**
21. **CLARKE, L.; BERRINGTON, A. Socio-demographic predictors of divorce. In: High divorce rates: the state of the evidence on reasons and remedies: reviews of evidence on the causes of marital breakdown and the effectiveness of policies and services intended to reduce its incidence. Lord Chancellor's Department, 1999. p.1-38.**
22. **HOULE, R. et al. Los determinantes sociodemográficos y familiares de las rupturas de uniones en España. In: SEMINARI URBÁ DIVORCIALIDAD Y DISOLUCIÓN DEL HOGAR: CAUSAS Y EFECTOS, 3., 1998, Barcelona. Anais... Barcelona: Centre d'estudis demografics, 1998.**
23. **KIERNAN, K.; MUELLER, G. The divorced and who divorces? London: Centre for Analysis of Social Exclusion, London School of Economics and Political Science, 1998. (Discussion Paper).**
24. **Vaz, W.M.C. Dissoluções e formação de novas uniões: uma análise demográfica das tendências recentes no Uruguai. Texto NEPO – UNICAMP, nº 56, Campinas, setembro de 2008.**
25. **CHERLIN, A.; KIERNAN, K.; CHASE LANDALE, L. Parental divorce in childhood and demographic outcomes in young adulthood. Demography, Chicago, v.32, n.3, p.299-318, 1995.**
26. **GOODE, W. J. World changes in divorce patterns. London: Yale University Press, 1993.**
27. **RUGGLES, S. The rise of divorce and separation in the United States, 1880-1990. Demography, Chicago, v.34, n.4, p.455-466, 1997.**
28. **Cortez, B.F., Lazo, A.V, Magalhães, M.S. A nupcialidade legal no Brasil e nas Grandes Regiões: uma análise utilizando tábuas de vida de múltiplos-estados. Trabalho apresentado no XVI Encontro Nacional de Estudos Populacionais, realizado em Caxambu- MG –Brasil, de 29 de setembro a 03 de outubro de 2008.**
29. **SWEENEY, M. Remarriage of men and women: the role of socio-economic prospects. CDE Working Paper n.95-08, University of Wisconsin-Madison, 1995. 30p.**

Abstract

Objective: To analyze the dynamics of marriage states in northeastern Brazil, by sex and age for. Method: We used information from the Civil Registry and the 2000 Population Census (IBGE), 2000. To convert the rates in the transition probabilities were used concepts of tables of survival, call multistage tables. Results: The proportion of singles has increased consistently in both sexes, meanwhile the proportion of married tried to reverse trend. There is 7% of women in the state of widowhood, while for men only 2%. Conclusion: The probability of first marriage is higher in women compared to men, since the re-marriage is more likely by men than among women.

Um Algoritmo para o Agrupamento Baseado em K-Medoids

José André de Moura Brito[*]
Luiz Satoru Ochi
Luciana Roque Brito
Flávio Marcelo Tavares Montenegro

Resumo

Descreve-se um novo algoritmo para um problema clássico de agrupamento, conhecido como problema dos *k*-medoids. Esse problema é análogo ao problema das *k*-médias, cuja resolução encontra aplicação frequente na fase de estratificação em amostragem probabilística. Entretanto, substituem-se os centroides por medoids, objetivando-se obter agrupamentos mais homogêneos e robustos. Dados *n* objetos com *p* atributos e fixado um número *k* de agrupamentos, deve-se selecionar *k* objetos (medoids) de forma a minimizar a soma das distâncias de cada um dos *(n-k)* objetos restantes ao seu medoid mais próximo. Ao se limitar o número máximo de objetos por grupo, obtemos o chamado problema dos *k*-medoids capacitado. O algoritmo proposto neste trabalho é baseado no método ILS (Iterated Local Search). A partir de dados do censo demográfico do IBGE, resultados computacionais mostram um desempenho razoável do algoritmo em termos de tempo computacional e de qualidade das soluções obtidas.

Palavras-Chave: Agrupamento, k-Medoids e ILS

[*] Endereço para correspondência: Escola Nacional de Ciências Estatísticas – ENCE . Rua André Cavalcanti,106 - sala 403, Santa Teresa, Rio de Janeiro – RJ - CEP:20231-050 E-mails: jose.m.brito@ibge.gov.br, satoru@dcc.ic.uff.br, lubritoroque@gmail.com. , flavio.montenegro@ibge.gov.br

1 Introdução

O problema de categorização ou de agrupamento pode ser descrito como o processo de agrupar os indivíduos, ou objetos, de uma base de dados de forma que os grupos obtidos representem uma possível configuração onde os objetos pertencentes a cada um dos k grupos resultantes possuam o mais alto grau de associatividade entre si (Negreiros *et al.*, 2002).

Para agrupar os indivíduos ou os objetos, é necessário determinar uma medida de similaridade ou dissimilaridade que quantifique o grau de associatividade entre eles. Essas medidas são obtidas por transformações a partir de dados quantitativos e/ou qualitativos (Kaufman e Rousseeuw, 1989).

Independentemente da medida utilizada, observa-se, ainda, que a procura por um agrupamento ótimo pode muitas vezes ser impraticável, devido à enorme quantidade de partições possíveis. Em geral, os métodos que são utilizados para a determinação dos grupos têm a capacidade de produzir soluções aproximadamente ótimas, sendo os métodos divididos em duas classes: hierárquicos e não hierárquicos.

Os métodos hierárquicos englobam procedimentos simples em que os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos (Everitt, 2001). Além disso, tais métodos não necessitam que seja definido *a priori* um número de agrupamentos e trabalham com uma matriz denominada matriz de similaridades entre agrupamentos.

Os métodos não hierárquicos são baseados na minimização de uma função de custo, onde os objetos são agrupados em um número k de agrupamentos escolhido *a priori*. Dentre as principais vantagens desses métodos, quando comparados aos métodos hierárquicos, destacamos: (i) a possibilidade de os objetos serem trocados de agrupamento durante a execução do algoritmo, (ii) operar com bases de dados maiores e (iii) produzir soluções (possíveis agrupamentos) de forma extremamente rápida.

Em particular, no presente trabalho, nos concentramos no estudo de um novo algoritmo para o problema de agrupamento com *k*-medoids, um problema similar ao das k-médias (Hartigan e Wong, 1979) e cuja resolução encontra uso frequente em várias áreas, incluindo a Estatística, a Inteligência Artificial, a Economia, etc. Observamos,

ainda, que os algoritmos baseados em k-médias ou k-medoids são classificados como não hierárquicos.

No problema dos k-medoids, dado um conjunto de n objetos com p atributos e definido um número k de agrupamentos, deve-se selecionar k objetos, chamados de medoids, de forma a minimizar a soma[1] das distâncias (função dos atributos) de cada um dos $(n-k)$ objetos restantes ao seu medoid mais próximo. Ao restringir-se o número máximo de objetos por agrupamento, define-se o chamado problema dos k-medoids capacitado, que é o objeto de estudo deste trabalho.

Uma análise mais detalhada desses dois problemas mostra que os algoritmos que trabalham com o conceito de k-medoids são mais eficientes do que os algoritmos baseados em k-médias, uma vez que tais algoritmos são mais robustos à existência de ruídos ou a *outliers*, geralmente produzem agrupamentos de alta qualidade (Kaufman e Rousseeuw, 1989; Han e Ng, 2002; Wei *et al.*, 2003; Zhang e Couloigner, 2005; Chu *et al.*, 2008; Park e Jun, 2009; Sheng e Liu, 2004; Grira e Houle, 2007; e Barioni *et al.*, 2007) e minimizam uma soma de distâncias (dissimilaridades) em vez do desvio quadrático médio (Späth, 1980).

Ademais, os medoids são de uso mais geral para a construção de agrupamentos, sendo aplicáveis em situações onde os objetos que serão agrupados não podem ser representados por atributos quantitativos ou a média desses objetos não está definida.

A solução exata (ótimo global) do problema dos k-medoids com ou sem a restrição de capacidade pode ser obtida através de uma formulação de programação matemática (Kaufman e Rousseeuw, 1989). Todavia, mesmo para um número de objetos apenas moderado, a resolução dessa formulação, ou seja, a obtenção do ótimo global, pode levar a um consumo expressivo de tempo computacional ou até mesmo a não convergência do algoritmo, resultando em uma solução que corresponde a um ótimo local.

[1] Kaufman e Rousseeuw (1989) sugerem a avaliação dos medoids considerando-se a função soma em vez da função média, uma vez que, do ponto de vista matemático, a minimização de uma dessas funções é equivalente à minimização da outra, havendo, entretanto, maior acurácia nos cálculos obtidos mediante a aplicação da função soma.

Em decorrência do aspecto combinatório desse problema, a obtenção do ótimo global resulta em uma tarefa de alta complexidade computacional (Garey e Johnson, 1979), o que significa uma exigência de tempo de processamento que cresce excessivamente tanto com o número de objetos a serem agrupados e quanto com o número *k* predefinido de grupos. Na verdade, caso seja aplicado um processo de busca exaustiva para garantir a obtenção do ótimo global, será necessário enumerar todas as soluções, ou seja, todas as possibilidades de combinação dos n objetos em k grupos. O número de possibilidades, nesse caso, está associado ao número de Stirling de segundo tipo (Johnson e Wichern, 2002), dado por $\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$. Caso se tenha, por exemplo, $n = 16$ objetos a serem alocados em 2 grupos, o número de soluções a serem consideradas é 32.767. Para o caso de 3 grupos, esse número cresce para 7.141.686 soluções possíveis. Considerando-se um número n maior de objetos, esses valores crescem exponencialmente.

Como alternativas para a resolução desse problema e de outros problemas de agrupamento (Kaufman e Rousseeuw, 1989; Späth, 1980), têm sido propostos vários algoritmos heurísticos (não exatos), com maior ou menor capacidade de produzir boas soluções (ótimos locais) em um tempo computacional muito pequeno quando comparado ao tempo consumido pelas formulações ou por procedimentos de enumeração (Nemhauser e Wolsey, 1999).

Em função de tais considerações, propomos para a resolução do problema dos k-medoids um novo algoritmo desenvolvido a partir do estudo do método ILS (*Iterated Local Search*). Este é um método relativamente recente (Lourenço *et al.*, 2002) que tem sido aplicado com sucesso em diversos problemas da área de otimização. O ILS consiste, basicamente, em um procedimento de construção de uma solução inicial e outro de melhoria dessa solução, o qual utiliza, para isso, um procedimento de busca local e de perturbação e um critério de aceitação.

O presente trabalho está dividido da seguinte forma: na seção dois, apresenta-se, em detalhes, o problema dos k-medoids. A seção três traz uma descrição dos tipos de variáveis que foram utilizadas no cálculo das distâncias entre os medoids e os objetos que compõem os grupos. Na seção quatro, temos os principais conceitos associados ao

método ILS, bem como uma descrição detalhada do algoritmo que foi proposto para a resolução do problema dos k-medoids capacitado. Finalmente, na seção cinco, apresentamos algumas análises referentes a um conjunto de resultados obtidos mediante a aplicação do algoritmo ILS, considerando-se um conjunto de problemas teste, gerado a partir de uma base de dados do Censo Demográfico de 2000.

2 O Problema dos k-Medoids

Dado um conjunto X formado por n objetos ($X = \{o_1, o_2, ..., o_n\}$) com p atributos (quantitativos e/ou qualitativos), deve-se selecionar, a partir de X, k objetos que definem um conjunto $M = \{med_1, med_{2,}..., med_k\}$ de medoids, de forma a minimizar a soma das distâncias de cada um dos $(n-k)$ objetos restantes ao seu medoid $med_i \in M, i \in \{1,..., k\}$ mais próximo. Ou seja, procura-se minimizar a soma das distâncias d_{ij} de todos os objetos $o_j \in med_i, i = 1,..., k$ aos seus respectivos medoids:

$$\textit{Minimizar} \sum_{i=1}^{k} \sum_{\forall o_j \in med_i} d_{ij} \qquad (1)$$

As distâncias da equação (1) determinam o grau de dissimilaridade entre os objetos e os seus respectivos medoids, sendo calculadas em função dos p atributos de cada um dos objetos.

Uma possível variação para esse problema consiste em restringir-se o número máximo de objetos associados a cada um dos k-medoids, obtendo-se, dessa forma, o chamado problema dos k-medoids capacitado.

Uma motivação para o estudo e a resolução dessa variante é que, por exemplo, em alguns processos de estratificação estatística em pesquisas amostrais, há a necessidade da formação de grupos com um número máximo (ou mínimo) de unidades de amostragem (Freitas *et al.*, 2007), sendo cada unidade definida por um município, setor, domicílio, etc.

O problema dos k-medoids também pode ser interpretado, em termos de análise de agrupamentos, como localizar n objetos e então substituí-los por k pontos (ou objetos representativos) no espaço n-dimensional (vide Hansen e Jaumard, 1997).

A solução exata (ótimo global) desse problema, bem como de suas variantes, pode ser obtida através de uma formulação de programação matemática (Kaufman e Rousseeuw, 1989). Todavia, mesmo para uma quantidade de objetos apenas moderada (da ordem de centenas), a resolução dessa formulação pode levar ao consumo expressivo de tempo computacional (dias, meses, anos,...) ou até mesmo a não convergência do algoritmo, resultando em um ótimo local.

Como possíveis alternativas, encontram-se, na literatura, alguns algoritmos bem conhecidos para o problema dos k-medoids. Um dos mais populares é o algoritmo proposto por Kaufman e Rousseeuw (1989), denominado PAM *(Partitioning Around Medoids*). Basicamente, o algoritmo PAM inicia o processo de formação dos agrupamentos, selecionando os k-medoids aleatoriamente dentre os n objetos de uma base de dados. Em seguida, em cada iteração, efetua-se uma troca entre um objeto que corresponde a um medoid e um objeto qualquer, de forma a reduzir o valor da função definida na equação (1).

Kaufman e Rousseeuw (1989) também desenvolveram uma versão modificada do algoritmo PAM, denominada CLARA (*Clustering Large Applications*). Tal algoritmo pode ser aplicado em bases de dados de dimensão elevada, considerando a combinação de uma técnica de amostragem e do algoritmo PAM. Em vez de determinar os k-medoids considerando toda a base de dados, o algoritmo CLARA seleciona m amostras[2] compostas por n' objetos da base de dados ($n' < n$), aplicando, em seguida, o algoritmo PAM em cada uma dessas amostras.

Mais recentemente, Han e Ng (2002) propuseram uma variação do algoritmo CLARA, denominada CLARANS. Tal algoritmo utiliza uma técnica de computação mais intensiva (Campello e Maculan, 1994) para determinação dos medoids.

Os algoritmos destacados acima tendem fortemente a convergir para pontos de ótimo local que apresentem, em geral, valores (equação 1) distantes do valor no ótimo

[2]Kaufman e Rousseeuw (1989) sugerem que se trabalhe com pelo menos cinco amostras. Além disso, propõem uma expressão para determinar o número de objetos que se deve selecionar para cada amostra.

global, ou seja, no ponto de ótimo correspondente à melhor solução obtida caso fossem verificadas todas as soluções possíveis para o problema.

Além desses, destacam-se os trabalhos de Wei *et al.* (2003), Zhang e Couloigner (2005), Chu *et al.* (2008), Park e Jun (2009), Sheng e Liu (2004), Grira e Houle (2007) e Barioni *et al.* (2007), que tratam o problema de k-medoids utilizando estratégias de busca local mais sofisticadas. Em particular, faremos uma pequena descrição dos três últimos trabalhos, pois estes apresentam algumas semelhanças conceituais com a proposta do presente artigo, embora em nenhum deles seja considerada a restrição de capacidade.

Sheng e Liu (2004) desenvolveram um algoritmo híbrido que combina um procedimento de busca local com a meta-heurística algoritmos genéticos (Linden, 2008), de forma a produzir melhores soluções. A busca local, aplicada na fase de cruzamento do algoritmo genético, é descrita, em linhas gerais, da seguinte forma: uma vez definidos os k-medoids, são selecionados os seus p pontos mais próximos, também definidos como vizinhos. Para cada um desses vizinhos, é avaliada a sua troca com o atual medoid. Havendo melhoria, atualiza-se o medoid e novamente são selecionados seus p vizinhos aplicando-se o mesmo procedimento. Tal procedimento é repetido até que não haja mais uma mudança nos medoids.

Nesse algoritmo genético, a população é composta por m cromossomos, sendo que cada cromossomo corresponde a uma solução, ou seja, um conjunto de medoids. Em seguida, calcula-se o valor da função objetivo para cada solução. A partir desses valores, considerando-se o procedimento de reprodução, as melhores soluções são promovidas para a geração seguinte com a utilização do método do torneio. Definindo-se uma probabilidade de cruzamento e de mutação para as soluções, são aplicados os procedimentos de cruzamento e mutação. Nesses procedimentos, basicamente, selecionam-se dois cromossomos e trocam-se os valores associados aos seus medoids, atribuindo-se novos medoids a essas posições. Tal troca definirá duas novas soluções, isto é, um conjunto de medoids, aplicando-se o procedimento de busca local. Os procedimentos de reprodução, cruzamento e mutação são aplicados por certo número de iterações ou até que não haja melhoria da solução após k iterações.

Grira e Houle (2007) propuseram um algoritmo híbrido que combina o método das k-médias com o método dos k-medoids. Mais especificamente, o algoritmo das k-médias é iniciado com um conjunto inicial de k centros escolhidos dentre os n pontos a serem agrupados. Após a definição dos centroides, verifica-se, em cada um desses grupos, qual o objeto que representa um medoid. Em seguida, selecionam-se dois grupos com os centroides mais próximos e escolhe-se um desses grupos aleatoriamente. Para esse grupo, serão selecionados p objetos pertencentes aos *(k-1)* grupos restantes. Cada um desses objetos é avaliado como possível medoid para substituir o medoid do grupo fixado efetuando-se trocas, ou seja, aplicando-se uma busca local em cada um dos p objetos. Havendo redução no valor da função objetivo, retorna-se a chamada ao algoritmo das k-médias, recalculando-se novos centros.

Finalmente, Barioni *et al.* (2008) propuseram um algoritmo de k-medoids baseado em *Metric Access Method* (MAM) e que pode ser eficientemente aplicado para grandes bases de dados. Ou seja, em vez de aplicar o algoritmo em uma amostra da base de dados (caso do algoritmo CLARA), o algoritmo trabalha com todos os dados. Basicamente, o algoritmo divide o espaço de dados (pontos a serem agrupados em torno dos medoids) em regiões que contêm objetos representativos (medoids) aos quais outros objetos das regiões podem ser associados. Os objetos contidos em cada região estão associados a um nó, o qual possui certo raio de cobertura. Somente os objetos dentro desse raio podem ser definidos como o medoid da região. Tal algoritmo foi combinado com o método PAM (citado anteriormente neste trabalho), de forma a produzir agrupamentos de melhor qualidade.

3 Variáveis Consideradas para Cálculo das Distâncias

Antes de se proceder à apresentação do algoritmo proposto para o problema dos k-medoids capacitado, fazer-se-á uma breve descrição dos tipos de variáveis que foram consideradas no cálculo das distâncias da equação (1) (vide seção 2). É fato conhecido (Kaufman e Rousseeuw, 1989) que a cada tipo de variável corresponde uma particular medida de distância. Em análise de agrupamento, a distância é uma medida matemática

que representa o grau de similaridade ou dissimilaridade entre os objetos que serão agrupados.

3.1 Variáveis Quantitativas

Ao se efetuar o cálculo da distância entre objetos com atributos do tipo quantitativo, deve-se, inicialmente, aplicar uma padronização dos dados. Ou seja, dado um conjunto de n objetos representado por $X = \{o_1, o_2, ..., o_n\}$, com cada objeto o_j possuindo observações de p variáveis, $o_j = (o_j^1, o_j^2, ..., o_j^p)$, os valores originais associados às variáveis devem ser convertidos para valores adimensionais. O processo de padronização consiste em calcular a média μ e o desvio padrão σ dos valores associados a cada um dos atributos, considerando os n objetos, e, para obter-se o *j*-ésimo valor associado à *h*-ésima variável, aplicar a fórmula

$$z_j^h = \frac{o_j^h - \mu_h}{\sigma_h} \qquad (2)$$

onde *j = 1, 2, ... , n (objetos), h = 1, 2, ... , p (variáveis)* e $z_j = (z_j^1, z_j^2, ..., z_j^h, ..., z_j^p)$ é o vetor de variáveis normalizadas para o *j*-ésimo objeto.

Uma vez aplicada a padronização, pode-se utilizar, por exemplo, a fórmula da distância euclidiana (equação 3), considerando dois objetos o_i e o_j e suas p variáveis:

$$d_{ij} = \sqrt{\sum_{h=1}^{p} (z_i^h - z_j^h)^2}, \qquad (3)$$

3.2 Variáveis Qualitativas

As variáveis qualitativas representam a informação que identifica alguma categoria ou característica, não susceptível de medida, mas de classificação. Podem ser divididas em binárias, nominais, ordinais e mistas. A seguir, apresentaremos alguns detalhes sobre esses tipos e suas respectivas fórmulas de cálculo de distâncias, tais como usadas neste trabalho.

3.2.1 Variáveis Binárias

As variáveis binárias assumem valor 0 (não têm o atributo) ou 1 (têm o atributo) e podem ser de dois tipos: simétricas ou assimétricas. As variáveis binárias simétricas são aquelas cujos dois estados influenciam igualmente o processo de agrupamento. Já no caso das variáveis binárias assimétricas, os dois estados têm influência diferenciada no processo de agrupamento.

Considerando os objetos o_i e o_j, medidos através de p variáveis binárias, definem-se os seguintes coeficientes:

a - Número de variáveis presentes (valor 1) nos dois objetos;

b - Número de variáveis presentes em i e ausentes em j;

c - Número de variáveis ausentes em i e presentes em j;

d - Número de variáveis ausentes (valor 0) nos dois objetos.

As equações (4.1) e (4.2) são definidas em função dos coeficientes a,b,c,d e correspondem, respectivamente, às distâncias (o grau de dissimilaridade) entre variáveis binárias simétricas e assimétricas:

$$d_{ij} = \frac{b + c}{a+b+c+d} \tag{4.1}$$

$$d_{ij} = \frac{b + c}{a+b+c} \tag{4.2}$$

Essas distâncias estão associadas ao coeficiente de emparelhamento simples. Alternativamente, seria possível utilizar a distância definida por Rogers e Tanimomoto (1960) ou a distância definida por Sokal e Sneath (1963).

3.2.2 Variáveis Nominais

São variáveis cujos estados não se limitam a dois, como nas variáveis binárias, mas podem assumir um determinado número de estados. Nesse caso, a distância, ou dissimilaridade, entre os dois objetos o_i e o_j pode ser medida através de:

$$d_{ij} = (u - m)/u \tag{5}$$

Na equação (5), u é o número total de variáveis nominais e m é o número de variáveis nominais de mesmo estado nos dois objetos.

3.3 Variáveis Mistas

Comumente, os objetos de um certo conjunto de dados a serem agrupados possuem variáveis mistas. Nesse caso, pode-se determinar a distância d_{ij} entre esses objetos mediante a soma das distâncias associadas aos vários tipos de variáveis, de acordo com a expressão

$$d_{ij} = \frac{\sum_{h=1}^{p} \delta_{ij}^{h} d_{ij}^{h}}{\sum_{h=1}^{p} \delta_{ij}^{h}} \tag{6}$$

Na equação acima, d_{ij}^{h} é a distância entre os objetos o_i e o_j, considerando a h-ésima variável, e δ_{ij}^{h} assume valor um se os valores de x_i^h *e* x_j^h estão definidos para a h-ésima variável. Em caso contrário, δ_{ij}^{h} assume valor zero, e a distância d_{ij}^{h} não é calculada.

4 Algoritmo

Descreve-se, na presente seção, o algoritmo proposto para o problema dos k-medoids capacitado, implementado a partir do estudo do método ILS (*Iterated Local Search*). Inicialmente, de forma a facilitar o entendimento do novo algoritmo, serão apresentados alguns conceitos básicos associados a esse método.

4.1 Conceitos Básicos

(i) Vizinhança: Uma vizinhança $V(s,\varepsilon)$ de uma solução s é um conjunto de soluções s' que podem ser obtidas a partir de s através de uma operação simples ε. Tal operação ε pode significar remover um objeto de s ou adicionar um objeto em s. A troca na posição de dois objetos que compõem uma solução é um outro exemplo dessa operação, que é particularmente comum quando os objetos estão em sequência. A operação ε também é chamada de perturbação.

(ii) Ótimo Local (solução viável): Dada uma solução s para um problema *P*, e sendo f a função objetivo associada a esse problema, dizemos que s é um ótimo local de *P* se existe uma vizinhança $V(s,\varepsilon)$ tal que para toda solução $s' \in V(s,\varepsilon) \Rightarrow f(s) \leq f(s')$ (problema de minimização) ou $\forall s' \in V(s,\varepsilon) \Rightarrow f(s) \geq f(s')$ (problema de maximização).

(iii) Ótimo Global (solução exata): Dada uma solução s para um problema *P* e sendo f a função objetivo associada a *P*, dizemos que s é um ótimo global de *P* se, para toda solução $s' \in$ Dom(f) , temos $f(s) \leq f(s')$ (problema de minimização) ou $f(s) \geq f(s')$.

(iv) Busca Local: Um procedimento de busca local é baseado em uma ideia simples e geral, considerando um problema de otimização *P*, sua função objetivo f associada e a solução atual s, que, no momento, é um ótimo local de *P*. O procedimento de busca local explora uma determinada vizinhança $V(s,\varepsilon)$, de forma a atingir um novo ótimo local $s' \in V(s,\varepsilon)$ tal que $f(s) > f(s')$ (problema de minimização) ou $f(s) < f(s')$ (problema de maximização). Havendo melhoria no valor da função objetivo, toma-se s' como nova solução, define-se uma nova vizinhança para s' e repete-se o procedimento de busca.

4.2 Método ILS

O método ILS (*Iterated Local Search*), descrito em Lourenço *et al*. (2002), consiste, essencialmente, na aplicação iterativa de um procedimento de busca local em uma solução inicial s_o, previamente obtida a partir da utilização de um algoritmo de

construção (Campello e Maculan, 1994). No caso particular do ILS, a busca local tem por finalidade melhorar a solução inicial s_o e as soluções obtidas após a aplicação do procedimento de perturbação.

O êxito desse método está diretamente associado à definição do **procedimento de busca local**, do **procedimento de perturbação** aplicado sobre a solução atual e de um **critério de aceitação das soluções**. Observamos que a implementação desses procedimentos, bem como do critério de aceitação, está intrinsecamente associada ao problema que será resolvido.

O pseudocódigo da figura 1 contém os passos essenciais do método ILS. No passo (1), temos a construção de uma solução inicial s_o, e no passo (2) aplica-se uma busca local nessa solução, produzindo uma solução S^*. Em seguida, com a finalidade de obter soluções de qualidade superior à solução obtida no passo (2), temos, nos passos (3), (4) e (5), a aplicação dos procedimentos de perturbação e busca local e de um critério de aceitação. O procedimento de perturbação (passo (3)) produz soluções intermediárias S' que podem ser melhoradas através da aplicação da busca local (passo 4). Esse procedimento impede a estagnação em pontos de ótimo local de baixa qualidade.

A solução S'' produzida a partir dessa busca será comparada com a melhor solução S^* obtida até o momento. Se S'' passa no teste de aceitação (passo 5), definimos $S^* = S''$; caso contrário, mantém-se S^*, aplicando-se novamente os procedimentos de perturbação e busca local. Ao final de *m* iterações, S^* será a melhor solução obtida.

```
1. so = Gerar_Solução_Inicial;
2. s* = Busca_Local(so);
Repita
  3. s' = Perturbação(s*);
  4. s" = Busca_Local(s') ;
  5. s* = Critério_Aceitação(s*,s");
Até (Sejam efetuadas m iterações);
```

Figura 1 - **Pseudocódigo do Método ILS**

4.3 Algoritmo ILS para o Problema dos k-Medoids

A presente seção traz uma descrição detalhada dos quatro procedimentos do algoritmo ILS proposto para a resolução do problema dos k-medoids.

O procedimento de geração da solução inicial consiste na construção de q soluções, cada uma com k objetos que correspondem aos medoids do problema capacitado. Em seguida, dentre essas q soluções, selecionamos a de menor custo, de acordo com o valor da função objetivo (equação 1 da seção 2). Algumas das soluções restantes, as de melhor qualidade segundo o valor da função objetivo, são armazenadas em um conjunto E que é atualizado em cada iteração do algoritmo, substituindo-se a pior solução de E pela solução obtida mediante a aplicação da busca local.

O procedimento de perturbação adotado nesse algoritmo consiste em selecionar, de forma aleatória, um dos medoids que compõem a solução s^* e também um objeto o_j dentre os $(n-k)$ objetos que não são os medoids da solução atual, trocando-o, em seguida, pelo medoid selecionado.

A busca local implementada nesse algoritmo consistiu em um procedimento denominado ***Trocas*** e em um procedimento denominado ***Reconexão por Caminhos***. No primeiro procedimento, aplicado em todas as *m* iterações do algoritmo, selecionam-se dois medoids, med_r e med_s, que compõem a solução s', obtida após a aplicação do procedimento de perturbação.

A partir desses medoids, são aplicados dois tipos de movimentos com a finalidade de reduzir o valor da função objetivo: (1) Efetuam-se l trocas de objetos entre os grupos definidos por med_r e med_s; (2) Tenta-se substituir cada um dos k-medoids pelos $(n-k)$ objetos restantes, fazendo-se a realocação dos objetos ao seu medoid mais próximo.

No caso do algoritmo ILS implementado neste trabalho, em vez de efetuar-se os passos um e dois apenas uma vez e os passos três, quatro e cinco por um certo número m de iterações, obtendo-se, ao final destas, a melhor solução (vide figura 1), realiza-se uma chamada do algoritmo ILS por w iterações "principais".

Assim, em cada uma das w iterações, são executados os passos um e dois (geração e busca local) e, em seguida, os passos três, quatro e cinco (perturbação, busca local e critério de aceitação). Tal procedimento, conhecido como ***multistart*** (Lourenço *et al.* , 2002), pode contribuir para a obtenção de soluções de boa qualidade. A utilização de um procedimento *multistart* está baseada no seguinte fato: métodos de busca local que têm por finalidade encontrar ótimos locais de boa qualidade, ou até mesmo o ótimo global, geralmente necessitam de algum tipo de diversificação para "escapar" de ótimos locais de baixa qualidade. Sem a diversificação, esses métodos podem iniciar em "áreas pequenas", associadas ao espaço de soluções, tornando impossível a obtenção de ótimos de boa qualidade. Ao longo dos últimos anos, muitos métodos têm trabalhado com *multistart*, ou seja, buscando a solução ótima e tendo por ponto de partida não apenas uma, mas várias soluções iniciais.

Ao final de cada uma das w iterações, aplica-se o segundo procedimento de busca local (***Reconexão por Caminhos***) considerando a melhor solução s^* obtida do algoritmo e as soluções $s^e \in E$.

Com base nessas soluções, o procedimento de ***Reconexão por Caminhos*** consistirá em "transformar" cada solução $s^e \in E$ na solução s^*, através de movimentos (incrementos) em cada um dos medoids que compõem s^e. Para cada movimento aplicado em s^e, produz-se uma nova solução intermediária s^i até que o conjunto dos medoids de s^i fique igual ao dos medoids de s^*.

A partir da aplicação desse procedimento, tem-se a expectativa de gerar soluções intermediárias s^i que tenham custo inferior ao da solução s^*, isto é, $f(s^*) > f(s^i)$.

Em Glover (1996), são fornecidos maiores detalhes sobre esse procedimento, que também tem sido utilizado conjuntamente com outros métodos implementados e aplicados em diversos problemas de otimização.

A seguir, apresentamos os passos principais dos procedimentos envolvidos no algoritmo ILS para o problema dos k-medoids capacitado.

Procedimento de Geração (Obter s⁰)

- Gerar q soluções iniciais s^0, cada uma com k-medoids escolhidos aleatoriamente a partir dos n objetos.
- Para cada solução s^0, alocar os $(n-k)$ objetos restantes ao seu medoid $med_i \in M, i \in \{1,...,k\}$, mais próximo, considerando a função de distância da equação (1) e não ultrapassando o número máximo (T) de objetos por grupo.
- Escolher, dentre as q soluções, a solução s^0 de menor custo, considerando a equação (1).
- Guardar as q^* ($q^* < q$) melhores soluções restantes (com menores valores de acordo com a equação (1)) em um conjunto E.

Procedimento de Perturbação (Obter s')

- Considerando a solução s^*, selecionar, aleatoriamente, dentre os $(n-k)$ objetos restantes, um objeto o_j e selecionar, aleatoriamente, um medoid med_i dentre os k-medoids que compõem s^*.
- Substituir o medoid med_i pelo objeto o_j para obter s'.
- Realocar os $(n-k)$ objetos aos medoids que compõem s' e recalcular o valor da função (equação (1)).

Procedimento de Busca Local (Obter s") – *Trocas*

- Selecionar em s^0 ou em s' dois medoids (aleatoriamente), med_r e med_s, $(1 \le r \le k, 1 \le s \le k, r \ne s)$.
- Selecionar, aleatoriamente, um objeto o_a pertencente ao agrupamento definido por med_r e um objeto o_b pertencente ao agrupamento definido por med_s e trocar o_a com o_b. Ou seja, o_a vai para o agrupamento de med_s e o_b vai para o agrupamento de med_r. Efetuar tal procedimento de troca l vezes, avaliando se houve uma redução no valor da função definida pela equação (1).

- Realizar tentativas de substituir cada medoid $med_i\ (1 \le i \le k)$ por outro medoid, considerando como possíveis novos medoids apenas os $(n-k)$ objetos que não são medoids na solução atual, de forma a reduzir o valor da função da equação (1).

Procedimento de Busca Local (Obter s'') – *Reconexão por Caminhos*

- Selecionar a melhor solução s^* obtida após m iterações do algoritmo ILS.
- Considerando s^* e cada solução $s^e \in E$, aplicar o procedimento de ***Reconexão por Caminhos*** (RC) e obter a melhor solução intermediária s^i.
- Para cada med_j de s^* e cada med_i de s^e, efetuar movimentos intermediários, caracterizados por um incremento ou um decremento no valor de med_i, de forma a obter med_j, avaliando se em cada movimento $f(s^i) < f(s^*)$, mediante a realocação dos $(n-k)$ objetos aos novos medoids que compõem s^i.

Na tabela abaixo, apresentamos um exemplo da aplicação do procedimento de reconexão por caminhos (RC), observando que foram produzidas cinco soluções intermediárias, em que cada valor destacado corresponde à mudança no valor de um medoid.

Tabela 1 - Ilustração do procedimento (RC)

Solução	*Medoid 1*	*Medoid 2*	*Medoid 3*
s^*	**80**	**12**	**24**
s^e	**77**	**14**	**23**
s^i	**78**	**14**	**23**
s^i	**79**	**14**	**23**
s^i	**80**	**14**	**23**
s^i	**80**	**13**	**23**
s^i	**80**	**12**	**23**
$s^i = s^*$	**80**	**12**	**24**

O procedimento de reconexão por caminhos diferencia-se de um procedimento de cruzamento, presente na meta-heurística algoritmos genéticos (Linden, 2008), por trabalhar com toda a solução, e não apenas com "trechos" da mesma, onde são geradas apenas duas novas soluções "filhas". Ou seja, a troca entre dois cromossomos não considera muito a intensificação no espaço de busca e nem a diversificação das soluções.

No caso da reconexão por caminhos, são geradas várias soluções intermediárias entre uma solução da busca local e do conjunto elite, sendo explorada uma "trajetória" entre essas duas soluções, de forma a tentar produzir pelo menos um mínimo local de melhor qualidade do que o atual. Em cada movimento, a nova solução é avaliada e comparada com a solução atual.

Critério de Aceitação (Obter s^*)

Se $f(s'') < f(s^*)$, então substitua s^* por s'' e atualize E

Senão, Se $|f(s^*) - f(s'')| / f(s^*) \leq \varepsilon$, substitua s^* por s'' e atualize E

Senão, matenha s^*.

Obs1: f é o valor da função objetivo (equação 1) considerando a solução s'' ou s^*, e ε é um fator de tolerância definido *a priori*. Observa-se que tal desigualdade permite a aceitação de soluções de pior qualidade do que a melhor solução obtida até o momento. Tal questão está intrinsecamente associada à característica do algoritmo ILS e de algoritmos baseados em outras metaheurísticas. Permite-se que a solução piore, na expectativa de produzir soluções melhores a partir desta, mediante a perturbação e a busca local. Tal critério, conjuntamente com a perturbação, evita que o algoritmo fique "estagnado" em pontos de mínimo local de baixa qualidade.

Obs2: A atualização da lista E corresponderá à substituição da pior solução dessa lista pela solução s'', considerando o valor da função f (equação 1).

5 Resultados Computacionais

A presente seção contém um conjunto de resultados computacionais obtidos para o problema dos k-medoids capacitado, mediante a aplicação do algoritmo ILS e da formulação de programação matemática descrita em Kaufman e Rousseeuw (1989).

A formulação foi implementada a partir da utilização do *software* de otimização LINGO (http://www.lindo.com), e o algoritmo foi implementado em linguagem Delphi (versão 6.0).

Todos os experimentos computacionais com o algoritmo e a formulação foram realizados em um computador AMD-Athlon 64 X 2 com 2 Gb de memória RAM e dotado de um processador de 2.31 GHz (*Dual Core*).

Antes da apresentação dos resultados, faremos uma breve descrição dos dados utilizados no experimento.

5.1 Dados Utilizados

A avaliação do algoritmo ILS foi efetuada utilizando-se uma base de dados do IBGE (*Instituto Brasileiro de Geografia e Estatística*), composta pelos domicílios da amostra do Censo Demográfico de 2000 (CENSO DEMOGRÁFICO 2000, *Primeiros Resultados da Amostra, Parte I*, 2001, IBGE/CDDI).

A partir dessa base, foram selecionadas nove áreas de ponderação[3] no estado do Paraná e oito áreas de ponderação no estado de Pernambuco, sendo escolhidos, em cada uma dessas áreas, de 50 a 400 registros (domicílios). Foram definidos, dessa forma, dezessete arquivos (ou problemas teste) utilizados nos experimentos. Os domicílios selecionados correspondem aos objetos que foram agrupados.

Observamos, ainda, que cada registro selecionado continha seis campos correspondentes às variáveis utilizadas no cálculo das distâncias d_{ij} (equação 1), quais sejam: o número de pessoas no domicílio (quantitativa), o número de banheiros no domicílio (quantitativa), o domicílio possui iluminação (binária assimétrica), o domicílio possui computador (binária simétrica), o tipo de domicílio (nominal) e o total de rendimentos do domicílio em salários mínimos (quantitativa).

Maiores informações sobre conceitos de área de ponderação e de setor censitário e sobre as variáveis consideradas podem ser obtidas consultando-se a *Metodologia do Censo Demográfico 2000* / IBGE – Rio de Janeiro: IBGE, 2003, Relatórios Metodológicos.

[3] Uma área de ponderação é formada por um agrupamento mutuamente exclusivo de subáreas chamadas setores censitários, os quais englobam, cada um, um conjunto de domicílios.

5.2 Experimentos Computacionais

A tabela 1 sumariza os resultados obtidos a partir da aplicação do algoritmo ILS e da formulação, considerando os dados descritos na seção anterior.

A primeira coluna dessa tabela contém a identificação dos problemas utilizados no experimento. Ou seja, no "rótulo" definido por **Censo_*x*_*y*_*z***, ***x*** indica o número do problema teste, ***y*** indica o número de domicílios selecionados e ***z*** indica o número máximo de domicílios por grupo, ou seja, o número máximo de objetos associados a cada medoid.

Na coluna dois, temos o número de medoids definidos para a utilização do algoritmo ILS (no presente experimento, tal número variou entre 2 e 5). Finalmente, nas colunas 3, 4, 5 e 6, temos, respectivamente, o valor da solução (de acordo com a equação (1)) e o tempo de processamento (em segundos), obtidos a partir da aplicação do algoritmo ILS e da formulação.

Em relação aos parâmetros do algoritmo, ou seja, o número de iterações "principais" e o número de execuções dos procedimentos de perturbação, busca_local e critério de aceitação, foram definidos, respectivamente, os seguintes valores: $w = 25$ e $m = 20$. Além disso, em cada chamada ao procedimento de geração, foram produzidas $q = 30$ soluções iniciais, sendo selecionadas, dentre estas, as 20 melhores soluções para compor o conjunto E.

Para a resolução de cada problema mediante a aplicação da formulação, estabeleceu-se o tempo limite de seis horas, obtendo-se a solução no decorrer desse período caso o algoritmo de *Branch and Bound* (Nemhauser e Wolsey, 1999) utilizado pela formulação termine, ou seja, convirja. Ao final das seis horas, caso esse algoritmo não tenha sido finalizado, toma-se a melhor solução viável gerada pela formulação até o momento. Tal solução corresponde a um ótimo local.

A partir dos resultados da tabela 1, pode-se observar que, para a maioria dos problemas, o algoritmo ILS produziu soluções melhores ou iguais àquelas advindas da formulação exata. Mais especificamente, foi possível, mediante a utilização do procedimento ILS, obter o ótimo global para 40 dentre os 68 problemas resolvidos (17 problemas x número de agrupamentos), o que corresponde a um percentual de 59%.

Para encontrar o ótimo global, o algoritmo ILS consumiu, em média, um tempo de 7 segundos, enquanto a formulação consumiu, em média, um tempo de 4.005 segundos. A tabela 2 traz a média e o resumo de cinco números associados a esses tempos de processamento (em segundos). Com base nessa tabela, é claramente visível que o algoritmo ILS demandou um tempo de processamento significativamente inferior ao da formulação no que concerne à obtenção dos ótimos globais.

Além disso, para os problemas onde só foi possível obter o ótimo local (assinalados na tabela), ou seja, o algoritmo da formulação não finalizou (gastou o tempo de 21.600 segundos), o algoritmo ILS forneceu ótimos locais melhores do que os da formulação em 18 dos 68 problemas, o que corresponde a um percentual de 26%. Para esses problemas, o tempo médio de execução do algoritmo ILS foi de 48 segundos. E em relação à qualidade das soluções, os maiores ganhos percentuais dos ótimos locais produzidos pelo ILS em relação aos ótimos da formulação foram obtidos para o último problema (**Censo_17_400_280**), aonde a diferença entre essas soluções chega a quase 9%.

Essas observações evidenciam uma boa performance do algoritmo ILS em sua aplicação ao problema dos k-medoids capacitado, especialmente quando houver a necessidade de se trabalhar com problemas de dimensão mais elevada (com mais objetos).

Tabela 1 – Resultados do Algoritmo ILS e da Formulação (Dados do Censo Demográfico)

Problema	*Nº Grupos*	*Algoritmo*	*Tempo*	*Formulação*	*Tempo*
	2 (a)	32.995	4	32.995	261
Censo_1_90_50	3 (a)	30,454	7	30,454	5711
	4 (b)	28,128	11	28,128	21600
	5 (b)	25,848	17	25,848	21600
	2 (a)	34,925	5	34,925	405
Censo_2_100_70	3 (a)	32,095	8	32,095	12988
	4 (d)	29,555	13	29,572	21600
	5 (d)	28,028	22	28,077	21600
	2 (c)	30,437	4	30,175	157
Censo_3_80_40	3 (a)	26,860	6	26,860	2111
	4 (a)	24,639	9	24,639	15691
	5 (b)	23,298	14	23,298	21600
	2 (a)	17,752	3	17,752	22
Censo_4_50_30	3 (a)	15,340	4	15,340	68
	4 (a)	13,711	6	13,711	88
	5 (a)	12,712	9	12,712	64
	2 (a)	40,628	6	40,628	787
Censo_5_120_70	3 (a)	35,392	10	35,392	21387
	4 (d)	33,391	16	33,422	21600
	5 (d)	31,500	26	31,565	21600
	2 (a)	24,841	4	24,841	71
Censo_6_70_40	3 (a)	22,719	6	22,719	958
	4 (a)	20,892	9	20,892	7171
	5 (a)	19,309	11	19,309	2545
	2 (a)	55,639	8	55,639	3987
Censo_7_150_100	3 (b)	48,068	13	48,068	21600
	4 (d)	43,179	23	43,250	21600
	5 (d)	41,272	38	41,628	21600
	2 (a)	32,981	4	32,981	156
Censo_8_90_70	3 (a)	30,125	8	30,125	3163
	4 (a)	28,447	11	28,447	17027
	5 (b)	27,314	14	23,314	21600
	2 (a)	20,048	3	20,048	50
Censo_9_60_45	3 (a)	17,763	5	17,763	340
	4 (a)	16,069	6	16,069	1236
	5 (a)	15,076	8	15,076	1968
	2 (a)	23,581	3	23,581	58
Censo_10_70_50	3 (a)	21,172	6	21,172	858
	4 (a)	19,380	9	19,380	2726
	5 (a)	17,941	13	17,941	896
	2 (a)	28,240	4	28,240	124
Censo_11_80_60	3 (a)	25,018	6	25,018	1252
	4 (a)	23,361	10	23,361	11025
	5 (b)	22,158	14	22,158	21600
	2 (a)	31,974	4	31,974	232
Censo_12_90_50	3 (a)	29,684	8	29,684	5171
	4 (a)	27,435	12	27,435	21600
	5 (d)	25,749	20	25,758	21600
	2 (a)	34,505	5	34,505	313
Censo_13_100_80	3 (a)	31,738	8	31,738	7914
	4 (b)	29,701	12	29,701	21600
	5 (d)	27,770	20	27,888	21600
	2 (a)	29,550	4	29,550	252
Censo_14_90_60	3 (a)	26,132	7	26,132	2821
	4 (b)	24,467	11	24,467	21600
	5 (b)	22,971	18	22,971	21600
	2 (a)	51,323	6	51,323	2609
Censo_15_140_80	3 (d)	44,756	13	44,758	21600
	4 (d)	42,403	24	42,585	21600
	5 (d)	40,379	37	40,678	21600
	2 (a)	57,984	8	57,984	4131
Censo_16_160_90	3 (d)	52,313	15	52,591	21600
	4 (d)	48,944	26	49,436	21600
	5 (d)	46,330	44	46,704	21600
	2 (d)	139,816	33	143,022	21600
Censo_17_400_280	3 (d)	126,008	72	133,347	21600
	4 (d)	118,678	160	123,231	21600
	5 (d)	112,109	261	121,769	21600

a= ótimo global produziu pela formulação e pelo algoritmo ILS; b= ótimo local da formulação igual ao do algoritmo ILS ;
c= formulação produziu ótimo global e o algoritmo ILS produziu ótimo local; d= ótimo local do algoritmo ILS melhor do que o da formulação

Tabela 2 – Média e Resumo dos Cinco Números para os quarenta Problemas com Ótimo Global

Medidas	Tempo Algoritmo ILS	Tempo Formulação
Mínimo	3,0	22,0
1º Quartil	4,8	247
Mediana	6,0	1244,0
Média	6,8	4005,9
3º Quartil	8,3	4391,0
Máximo	13,0	21600,0

Os procedimentos de busca local e de perturbação, bem como o critério de aceitação, implementados no algoritmo ILS, são apenas tentativas iniciais no sentido de produzir ótimos locais de boa qualidade para o problema dos k-medoids em um tempo de processamento razoável. Em trabalhos futuros, podem ser levados em conta outros objetivos:

- Implementar e incorporar a esse algoritmo procedimentos e critérios de aceitação mais sofisticados, com a expectativa de obter soluções melhores do que as apresentadas no presente estudo.
- Implementar um dos algoritmos clássicos da literatura (PAM, CLARA, CLARANS), agregando-se a restrição do número máximo de objetos por grupo, com a finalidade de possibilitar uma comparação com o algoritmo ILS.
- Modificar o algoritmo ILS, de forma a resolver o problema dos k-medoids com outro tipo de restrição de capacidade. Ou seja, considerar um total máximo por grupo, correspondente à soma dos valores das observações de uma variável quantitativa, tal como o número de pessoas ou a renda.

Referências bibliográficas

Barioni M.C.N, Razente, H.L., Traina A.J.M. and Traina Jr. C. (2008). Accelerating k-medoid-based Algorithms Trough Metric Access Methods. Journal of Systems and Software, 81,3, 343-355.

Campello, R. E. e Maculan, N. (1994). Algoritmos e Heurísticas. Desenvolvimento e Avaliação de Performance. Editora da Universidade Federal Fluminense.

Chu, S.C., Roddick J.F and Pan J.S. (2008). Improved Search Strategies and Extensions to k-medoids-based Clustering Algorithms. International Journal of Business Intelligence and Data Mining, 3,2, 212-231.

Everitt, B.S. (2001). Cluster Analysis (fourth ed.), Edward Arnold, London.

Freitas, M.P.S, Lila, M.F., Azevedo, R.V. e Antonaci, Giuseppe de Abreu (2007). Amostra Mestra para o Sistema Integrado de Pesquisas Domiciliares, Textos para Discussão, 23, Diretoria de Pesquisas, IBGE.

Garey, M. R. e Johnson, D. S. (1979). Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: Freeman.

Glover, F. (1996). Tabu Search and Adaptive Memory Programming – Advances, Applications and Challenges. In: R. S. Barr, R. V. Helgason and J. L. Kennington (eds.), Interfaces in Computer Science and Operations Research. Kluwer, pp. 1-75.

Grira N. and Houle M.E. (2007). Best of Both: A Hybridized Centroid-Medoid Clustering Heuristic. International Conference on Machine Learning. http://www.machinelearning.org/proceedings/icml2007/papers/216.pdf.

Han, J. e Ng, R. (2002). "CLARANS:A Method for Clustering Objects for Spatial Data Mining. IEEE Transactions Knowledge of Data Enginnering 14(5), pp. 1003-1016.

Hansen, P. e Jaumard, B. (1997). Cluster Analysis and Mathematical Programming. Math. Programming, 79, pp. 191-215.

Hartigan, J. A. e Wong, M. A. (1979). A k-means clustering algorithm, Applied Statistics, 28, pp. 100-108.

Johnson A.R. e Wichern D.W. (2002). Applied Multivariate Statistical Analysis. Prentice Hall. Fifth Edition.

Kaufman, L. e Rousseeuw, P. J. (1989). Finding Groups in Data – An Introduction to Cluster Analysis. Wiley-Interscience Publication.

Linden, R. (2008). Algoritmos Genéticos – Uma Importante Ferramenta da Inteligência Computacional, Editora Brasport.

Lourenço H.R. , Martin, O. and Stützle T. (2002). Iterated local search. In F. Glover and G. Kochenberger, editors, Handbook of Metaheuristics, volume 57 of International Series in Operations Research & Management Science, pages 321-353. Kluwer Academic Publishers, Norwell, MA.

Negreiros, M.J.G, Almeida, P.G, Bezerra, A.G.F. e Xavier, A.E. (2002). Análise de agrupamentos para a taxa de resíduos sólidos de Fortaleza via sistema visual TAX. Revista da Associação Brasileira de Limpeza Pública, 57, 10-17.

Nemhauser G. and Wolsey L. (1999). Integer and Combinatorial Optimization. Wiley – Interscience Publication.

Park H.S. and Jun C.H. (2009). A Simple and Fast Algorithm for K-medoids Clustering. Expert Systems with Applications, 36,2, 3336-3341.

Rogers, D.J. and Tanimoto,T.T. (1960). A Computer Program for Classifying Plants. Science, 132, 1115-1118.

Sheng W. amd Liu X. (2004). A Hybrid Algorithm for K-medoid Clustering of Large Data Sets. Congress on Evolutionary Computation, 1, 77-82.

Sokal, R.R. and Sneath, P.H.A (1963). Principles of Numerical Taxonomy, Freeman, San Francisco.

Späth, H. (1980). Cluster Analysis Algorithms for Data Reduction and Classification of Objects. John Wiley & Sons.

Wei et al. (2003). Empirical Comparison of Fast Partitioning-Based Clustering Algorithms for Large Data Sets. Expert Systems with Applications, 24, 4, 351-363.

Zhang and Couloigner (2005). A New Efficient k-medoid Algorithm for Spatial Clustering. Lecture Notes in Computer Science, v3482, 181-189.

Abstract

We describe a new algorithm to solve a classical clustering problem known as *k*-medoids problem. This problem is analogous to the k-means problem, which is commonly related to stratification in sample surveys. However, the centroids are now replaced by medoids, aiming to obtain more robust and homogeneous clusters. Given *n* objects with *p* attributes and a fixed number *k* of desired clusters, we must select *k* objects (medoids), in such a way to minimize the sum of distances from each one of the remaining *(n-k)* objects to its respective nearest medoid. Constraining the maximum number of objects per cluster, we obtain the capacitated *k*-medoids problem. The algorithm proposed in this paper is based on the ILS (Iterated Local Search) method. Computational results using IBGE demographic census reveal a reasonable performance of the algorithm in terms of computational time consumption and quality of the solutions.

Key-Words: Clustering, k-Medoids and ILS

ANÁLISE DOS DADOS DE DOAÇÃO DE SANGUE DA FUNDAÇÃO HEMOMINAS -NÚCLEO REGIONAL DE SÃO JOÃO DEL REI

Profa. Rejane Correa Rocha[1]
Profa. Thelma Safadi[2]

Resumo

Este trabalho teve como principal objetivo ajustar modelos, utilizando a metodologia de Box e Jenkins, para as séries de doação de sangue da Fundação Hemominas - Núcleo Regional de São João del Rei, com a finalidade de fazer previsões, comparando-as com metas propostas pelo Ministério da Saúde. Foram utilizadas as séries número de doadores, número de inaptidões clínicas, número de inaptidões sorológicas e número de doações espontâneas, todas com medição mensal. As três primeiras foram coletadas no período de janeiro de 1993 a dezembro de 2006, num total de 168 observações e a quarta foi coletada entre agosto de 1995 e dezembro de 2006, totalizando 137 observações. As quatro últimas observações de cada uma das séries foram reservadas para serem comparadas com as previsões. A adequação dos modelos foi verificada por meio dos critérios erro quadrado médio de previsão (EQMP) e erro percentual médio absoluto (MAPE). Os índices previstos para os quatro últimos meses de 2006 para as séries do número de doadores, de inaptidões clínicas, de inaptidões sorológicas e de doações espontâneas foram de 0,531%, 1,6%, 11,5% e 53%, respectivamente, aproximando-se dos índices reais para o mesmo período.

Palavras-chave: Modelo SARIMA, doadores de sangue, modelos de intervenção, séries temporais.

[1] Universidade Federal de São João del Rei (Departamento de Matemática, Estatística e Ciências da Computação), 36305302 São João del Rei - MG, Brasil.
[2] Universidade Federal de Lavras (Ciências Exatas), Cx.P 3037, CEP 37200-000 Lavras - MG, Brasil

1 Introdução

De acordo com a Organização Mundial de Saúde (OMS), os índices de doação de sangue contam pontos para estabelecer o Índice de Desenvolvimento Humano (IDH). Cerca de 1,7% da população brasileira doa sangue, índice considerado baixo, se comparado ao recomendado pela OMS, que é de 3%.

No Brasil, a política de sangue, componentes e hemoderivados, foi impulsionada em 1998, quando o Programa Nacional de Qualidade e Produtividade (PNQP), reeditado pelo governo federal, definiu como meta mobilizadora nacional para a saúde o tema: "Sangue com garantia total em todo o seu processo até 2003". Dentre as metas do Ministério da Saúde para o setor de sangue e hemoderivados estão: atingir um índice de 2,2%, em 2006 e 3%, em 2007, de doadores em relação à população total, reduzir os índices de inaptidões clínicas para 11,3% e os de inaptidões sorológicas para 8,3% e, ainda, aumentar o índice de doações espontâneas para 100%.

É considerado doador todo o cidadão que chega ao posto de coleta e se cadastra com a finalidade de doar sangue. Se a doação não é vinculada a nenhum paciente específico, é considerada espontânea. Mas, se é vinculada a um paciente específico, com a finalidade de repor o estoque, é considerada de reposição. Após o cadastro, o doador passa por uma triagem clínica. Se não for aprovado, ele é considerado inapto clínico. Se for aprovado, é considerado apto e seu sangue é coletado. Durante a coleta, além das bolsas de sangue, também são coletadas amostras nas quais são realizados exames sorológicos. Se algum desses exames der resultado positivo, o doador é considerado inapto sorológico.

Devido ao fato de o Ministério da Saúde trabalhar baseado em metas e, sendo o processo de doação de sangue complexo e de alto custo, é importante fazer uma análise estatística, por meio de séries temporais, desses dados. Sobretudo no que se refere aos valores previstos pelo modelo ajustado, pois, de acordo com essas análises, o Ministério da Saúde poderá propor estratégias para que as metas propostas sejam alcançadas.

Existem vários estudos que utilizaram os dados de doação de sangue, na maioria deles, as análises são feitas por meio da estatística descritiva.

Urrutia, Duarte e Joanini (1999) fizeram uma análise dos dados dos doadores do Hemocentro da Universidade Católica de Campinas que apresentaram, no primeiro exame sorológico, resultados positivos ou inconclusivos. Faria (2004) também fez uma análise do doador inapto sorológico da Fundação Hemominas de Belo Horizonte. Camara (2003) analisa dados referentes às metas do Programa Nacional de Doação Voluntária (PNDVS) e à triagem laboratorial para doenças transmissíveis pelo sangue. Borges et al. (2005) analisaram os dados referentes à satisfação de doadores de sangue do Hemocentro de Ribeirão Preto, com o objetivo de avaliar a fidedignidade desses doadores. Foi utilizada a técnica de análise de correspondência, tendo os itens de maior insatisfação sido a dificuldade de acesso ao Hemocentro e o tempo gasto na doação. Os itens de maior satisfação referem-se à confiança no serviço, nos funcionários e atendimento. Silva et al. (2005) descrevem o processo de doação no HEMORIO e procuraram por meio de técnicas de simulação, um modelo que diminuísse o tempo total de doação e incentivasse os doadores a repetirem o ato voluntário. Loureiro (2005) também fez um estudo relativo à criação de um modelo de atendimento em triagem e coleta em bancos de sangue, utilizando os dados da Fundação Hemominas - Hemocentro de Belo Horizonte.

O objetivo deste trabalho é verificar se os modelos de Box e Jenkins se ajustam às séries do número de doadores, do número de inaptidões clínicas, do número de inaptidões sorológicas e do número de doações espontâneas da Fundação Hemominas - Núcleo Regional de São João del Rei e, a partir dos modelos ajustados, fazer previsões.

2 Material e métodos

A base de dados do presente trabalho foi obtida na Fundação Hemominas - Núcleo Regional de São João del Rei. Essa instituição fica localizada na cidade de São João del Rei e atende a demanda das microrregiões de São João del Rei e Lavras e, ainda, o município de Minduri.

Para o ajustamento dos modelos de séries temporais, foram propostas as séries: do número de doadores, do número de inaptidões clínicas, do número de inaptidões

sorológicas e do número de doações espontâneas. Os dados relativos às séries deste estudo são registrados diariamente, consolidados mensalmente e lançados no sistema de informação de produção hemoterápica (HEMOPROD). A apuração para a consolidação mensal dos dados é feita no período do dia 20 do mês anterior até o dia 19 do mês corrente. Por exemplo, o número de doadores, referente ao mês de dezembro de 2005, é a consolidação da apuração diária do número de doadores do dia 20 de novembro de 2005 ao dia 19 de dezembro de 2005.

As séries número de doadores, número de inaptidões clínicas e número de inaptidões sorológicas foram coletadas no período de janeiro de 1993 a dezembro de 2006, num total de 168 observações para cada série. A série número de doações espontâneas foi coletada no período agosto de 1995 a dezembro de 2006, totalizando 137 observações.

As observações do período de setembro de 2006 a dezembro de 2006, de cada uma das séries, foram reservadas para serem comparadas com as previsões.

Um modelo clássico para uma série temporal supõe que a série Z1 ,Z2, ..., Zn possa ser escrita como: Zt = Tt + St + at , em que Zt é a série, Tt é a tendência, St é a sazonalidade, at é uma componente aleatória e t = 1,2, ..., n.

Segundo Morettin e Toloi (2004), a tendência pode ser entendida como aumento ou diminuição gradual das observações ao longo do período. A sazonalidade mostra flutuações ocorridas em períodos (menores que um ano). E a componente aleatória mostra as oscilações aleatórias irregulares. A suposição usual é que at seja uma série puramente aleatória ou um ruído branco independente, com média zero e variância constante.

Na prática, a maioria das séries temporais apresenta as componentes tendência e ou sazonalidade. Neste trabalho, para verificar a presença da tendência foi aplicado o teste de Cox-Stuart e para a sazonalidade, foi feita a análise do periodograma e aplicado o teste de Fisher. Esses testes são descritos em Morettin e Toloi (2004).

Para o ajuste das séries propostas foi utilizado o modelo auto-repressivo integrado médias móveis, ARIMA(p,d,q), dado por:

$$\phi(B)\Delta^d Z_t = \theta(B)a_t, \qquad (1)$$

em que $\phi(B) = 1 - \phi_1 B^1 - \ldots - \phi_p B^p$ é o polinômio auto-regressivo de ordem p; $\Delta^d = (1-B)^d$ é o operador diferença e d é o número de diferenças necessárias para retirar a tendência da série; $\theta(B) = 1 - \theta_1 B^1 - \ldots - \theta_q B^q$ é o polinômio de médias móveis de ordem q.

Para lidar com a sazonalidade, Box e Jenkins (1976) generalizaram o modelo ARIMA e definiram o modelo ARIMA sazonal multiplicativo, denominado SARIMA de ordem (p,d,q)x(P,D,Q), dado por

$$\phi(B)\Phi(B^s)\Delta^d \Delta_s^D Z_t = \theta(B)\Theta(B^s)a_t \qquad (2)$$

em que $\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps}$ é o polinômio auto-regressivo sazonal de ordem *P*; $\Delta_s^D = (1-B^s)^D$ é o operador diferença generalizado, quando duas observações estão distantes entre si de s intervalos de tempos que apresentam alguma semelhança, e D é o número de diferenças de *lags* necessárias para retirar a sazonalidade da série; $\Theta(B^s) = 1 - \Theta_s B^s - \ldots - \Theta_Q B^{Qs}$ é o polinômio de médias móveis sazonal de ordem Q.

Segundo Morettin e Toloi (2004), a intervenção constitui em uma mudança de nível ou inclinação ocorrida com os dados num determinado instante do tempo, podendo ter efeito temporário ou permanente. Esses fenômenos são estimados pelo modelo $Y_t = \sum_{i=1}^{k} \upsilon_i(B) x_{i,t} + N_t$, em que Yt é o valor observado da série no tempo t, k é o número de intervenções da série, $\upsilon_i(B)$ é o valor da função de transferência, $x_{i,t}$ é a variável binária e Nt é um ruído branco representado pelo modelo SARIMA.

Para escolha do modelo mais adequado foi utilizado o critério do erro quadrático médio de previsão (EQMP), descrito em Morettin e Toloi (2004). E para avaliar se as previsões desses modelos foram satisfatórias calculou-se o erro percentual médio absoluto (MAPE), que é dado por:

$$MAPE = \frac{1}{h}\sum_{1}^{h}\left|\frac{e_t(h)}{Z_{t+h}}\right| \times 100, \qquad (3)$$

em que $e_t(h) = Z_{t+h} - \hat{Z}_t(h)$ é o erro de previsão, sendo Z_{t+h} o valor real, $\hat{Z}_t(h)$ o valor previsto e h o número de previsões.

Para todas as análises descritas acima foram utilizados o pacote stats do software R. 2.6.1 (R Development Core Team, 2007).

3 Resultados e Discussão

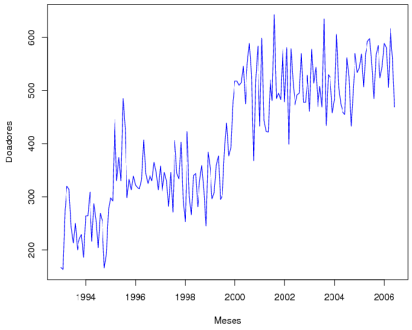
3.1 Análise descritiva

Na Tabela 1 é apresentada a consolidação dos dados de doação de sangue da Fundação Hemominas - Núcleo Regional de São João del Rei para o ano de 2006 e a população total atendida por essa instituição.

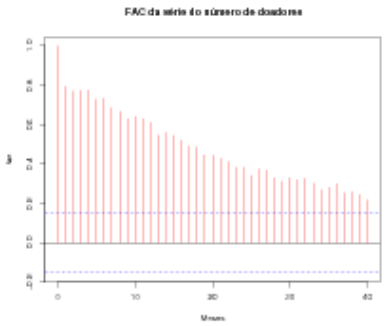
TABELA 1: Dados consolidados para o ano de 2006

Dados	Total
Número de doadores	6.494
Número de inaptidões clínicas	666
Número de inaptidões sorológicas	127
Número de doações espontâneas	3.473
População	≅ 400.000

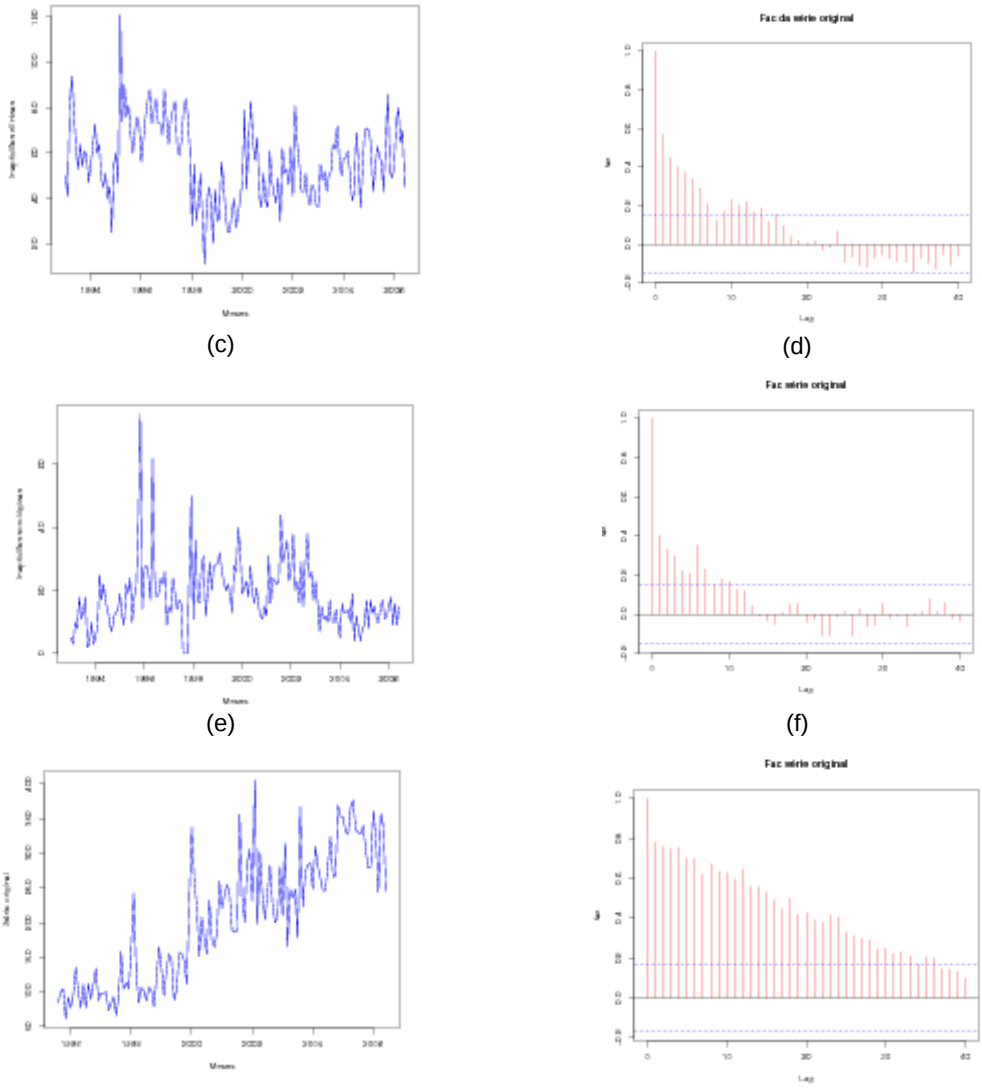
Fonte: HEMOPROD - Fundação Hemominas Núcleo Regional de SJDR e SES-MG



(a)



(b)



(c) (d)

(e) (f)

(g) (h)

Fonte: Fundação Hemominas- NRSJDR.

FIGURA 1: Representação gráfica das séries originais da Fundação Hemominas - Núcleo Regional de São João del Rei: (a) do número de doadores (c) do número de inaptidões clínicas (e) do número de inaptidões sorológicas (g) do número de doações espontâneas e suas respecitvas funções de autocorrelação: (b) Fac da série do número de doadores, (d) Fac da série do número de inaptidões clínicas, (f) Fac da série do número de inaptidões sorológicas e (h) Fac da série do número de doações espontâneas.

Verifica-se na Tabela 2, os índices de inaptidões clínicas, de inaptidões sorológicas, de doações espontâneas e de doadores em relação à população total, para o ano de 2006. Destaca-se o índice de inaptidões sorológicas que foi, aproximadamente 4 vezes menor do que o preconizado pelo Ministério da Saúde.

O índice de doações espontâneas ficou 47% abaixo do preconizado, mas ficou acima da média nacional que é de 51,3%.

TABELA 2: Índices consolidados para o ano de 2006

Dados	Índices Obtidos(%)	Metas(%)
Número de inaptidões clínicas	10,26	11,30
Número de inaptidões sorológicas	1,96	8,30
Número de doações espontâneas	53,48	100,00
Doadores/População total	1,62	2,20

3.2 Ajuste dos modelos de Box e Jenkins para as séries de doações de sangue

Os gráficos das séries originais do número de doadores, do número de inaptidões clínicas, do número de inaptidões sorológicas e do número de doações espontâneas da Fundação Hemominas- Núcleo Regional de São João del Rei e suas respectivas funções de autocorrelação (fac) estão mostrados Figura 1. Pela análise visual desses gráficos, pode-se dizer que todas essas séries não são estacionárias.

Para verificar a presença da componente tendência foi aplicado o teste do sinal de Cox-Stuart e para sazonalidade a análise do periodograma e o teste de Fisher. Os periodogramas das séries do número de doadores, do número de inaptidões clínicas e do número de inaptidões sorológicas, são apresentados nas Figuras 2a, 2b e 2c, respectivamente. Nessas figuras não são observados picos significativos em períodos menores que 12 meses, não caracterizando sazonalidade.

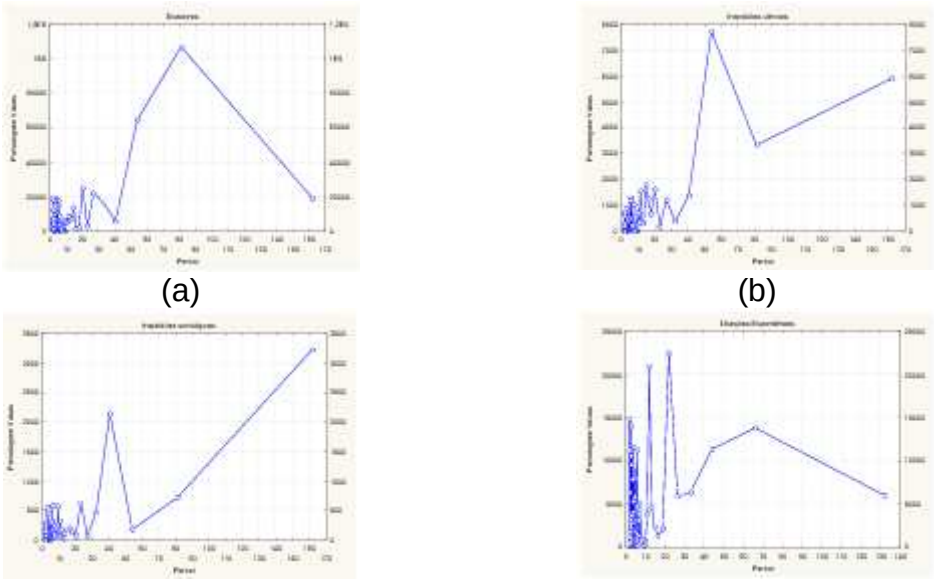


FIGURA 2: Periodogramas das séries originais da Fundação Hemominas - Núcleo Regional de São João del Rei: (a) número de doadores; (b) do número de inaptidões clínicas; (c) do número de inaptidões sorológicas e (d) do número de doações espontâneas.

O periodograma da série do número de doações espontâneas, Figura 2d, apresenta picos significativos nos períodos 12 e 22. Foi aplicado o teste de Fisher no período de 12 meses, cujas estatísticas são dadas por g = 0,071887863 e $z_{0,05}$ = 1,8953. Como g < z0,05, não existem evidências de sazonalidade no período de 12 meses, ao nível de 5% de probabilidade.

TABELA 3: Resultados do teste de Cox Stuart

Série	$P(Z_i < Z_{i+c})$	$P(Z_i > Z_{i+c})$
Número de doadores	1,00	0,00
Número de inaptidões clínicas	0,33	0,67
Número de inaptidões sorológicas	0,46	0,54
Número de doações espontâneas	0,95	0,05

Pelos resultados do teste de Cox-Stuart apresentados na Tabela 3, pode-se afirmar que todas as séries apresentam tendência, sendo necessário tomar a primeira diferença de cada uma das séries para que se retire essa componente.

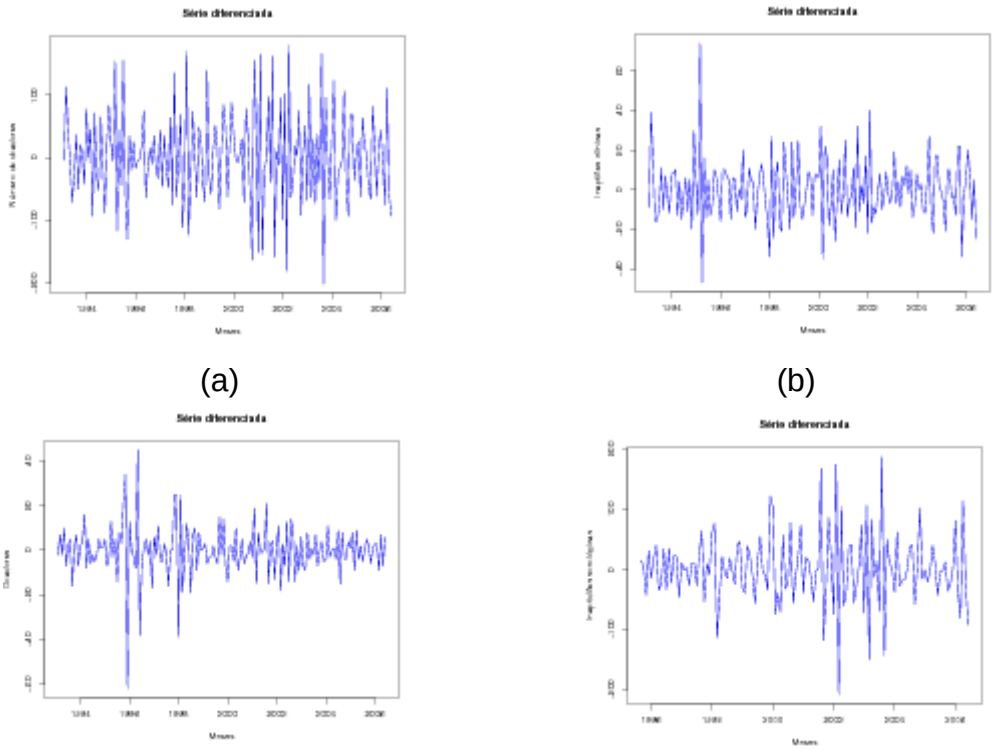


FIGURA 3: Representação gráfica das séries diferenciadas da Fundação Hemominas - Núcleo Regional de São João del Rei: (a) número de doadores;(b) do número de inaptidões clínicas; (c) do número de inaptidões sorológicas e (d) do número de doações espontâneas.

Por meio da Figura 3, observa-se que a tendência foi retirada em todas as séries. Pode-se ver também que existem indícios de intervenção para as séries do número de inaptidões clínicas, inaptidões sorológicas e doações espontâneas, Figuras 3b, 3c e 3d, respectivamente.

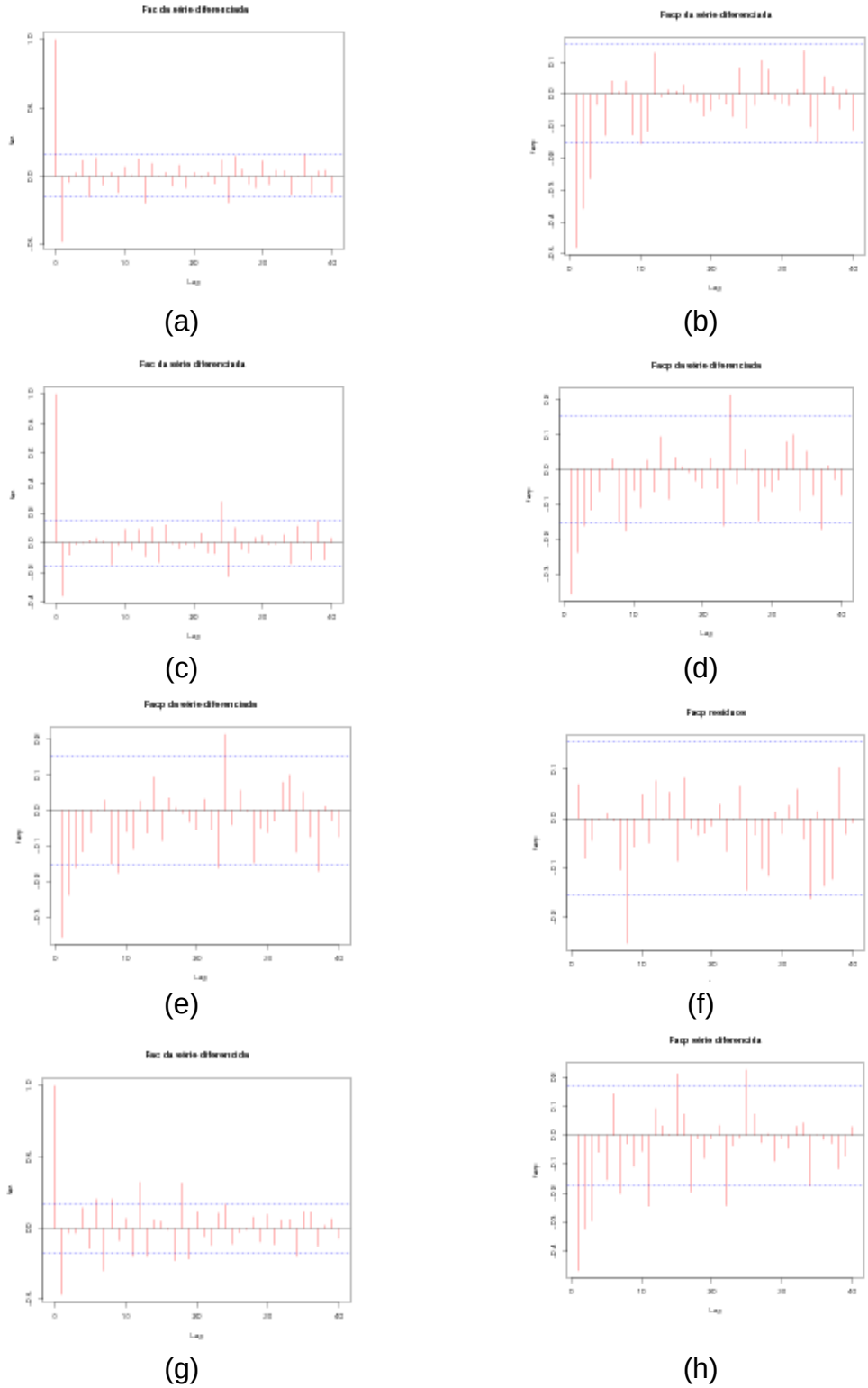


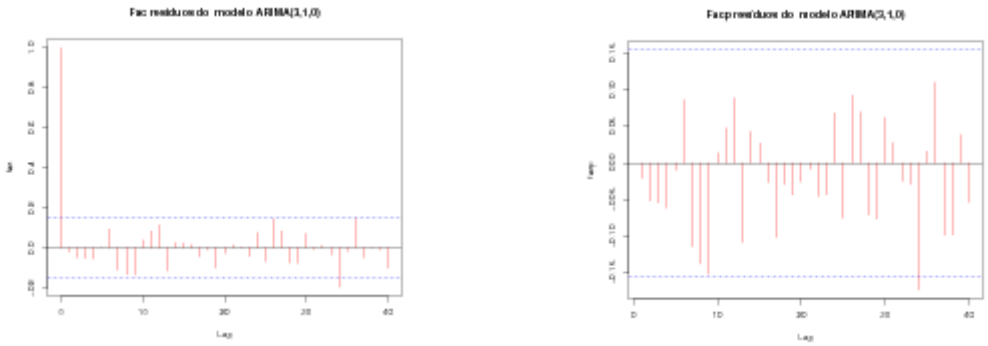
FIGURA 4: Representação gráfica das funções de autocorrelação (fac) e autocorrelação parcial (facp) das séries diferenciadas da Fundação Hemominas - Núcleo Regional de São João del Rei: (a) Fac do número de doadores, (b) Facp da do número de doadores, (c) Fac do número de inaptidões clínicas, (d) Facp da série do número de inaptidões clínicas, (e) Fac do número de inaptidões sorológicas, (f) Facp do número de inaptidões sorológicas, (g) Fac do número de doações espontâneas e (h) Facp do número de doações espontâneas.

A partir da análise dos gráficos das funções de autocorrelação (fac) e autocorrelação parcial (facp) das séries diferenciadas (Figura 4) ajustaram-se alguns modelos para cada uma das séries. Como o objetivo deste trabalho foi fazer previsões, a adequação dos modelos foi verificada por meio dos critérios erro quadrado médio de previsão (EQMP) e erro percentual médio absoluto (MAPE).

3.2.1 Série do número de doadores

Para a série número de doadores o modelo que apresentou menor EQMP foi o ARIMA(3,1,0), dado por:

$$Z_t = \frac{a_t}{(1-B)(1+0{,}768B+0{,}566B^2+0{,}285B^3)}. \qquad (4)$$



(a) (b)

FIGURA 5: Representação gráfica das funções de autocorrelação (a) e função de autocorrelação parcial (b) dos resíduos do modelo ARIMA(3,1,0).

Por meio da análise dos gráficos das funções de autocorrelação e autocorrelação parcial dos resíduos (Figura 5) do modelo ARIMA(3,1,0), pode-se observar que o resíduo at é um ruído branco. Essa análise foi confirmada pelo teste de Box-Pierce cujo resultado é dado por $Q_{40} = 40{,}08 < \chi^2_{37,0.05} = 52{,}19$.

TABELA 4: Valores observados (Z_{t+h}), valores previstos ($\hat{Z}_t(h)$), erro padrão (SE), erro de previsão ($e_{t(h)}$) para a série do número de doadores no período de setembro a dezembro de 2006, utilizando o modelo ARIMA(3,1,0).

Mês	Valor observado (Z_{t+h})	Valor previsto ($\hat{Z}_t(h)$)	Erro padrão (SE)	Erro de previsão ($e_{t(h)}$)
Set/06	559	529,483	62,252	-29,517
out/06	525	520,096	66,294	-4, 904
nov/06	519	537,105	72,331	18,105
dez/06	489	536,272	74,898	47,272
EQMP= 864,420		*MAPE*= 4,84%		

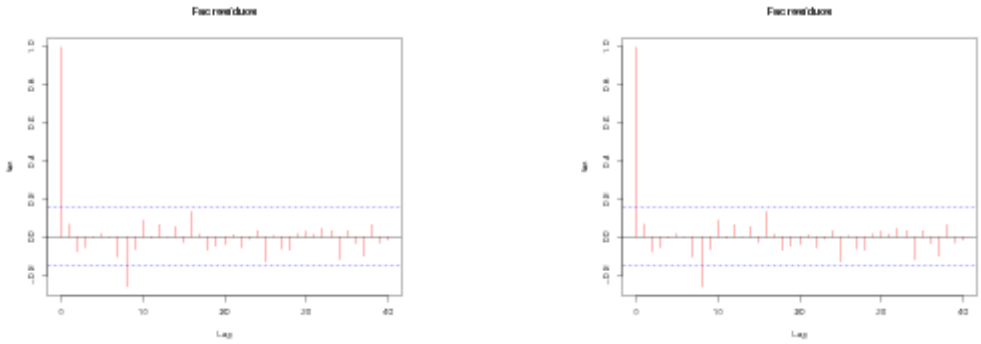
Nota: $e_t(h) = Z_{t+h} - \hat{Z}_t(h)$

Na Tabela 4 são apresentadas as previsões para o modelo ARIMA(3,1,0). É possível observar que as previsões foram bastante satisfatórias, apresentando um MAPE de 4,84%. O índice do número de doadores pela população total previsto para os quatro últimos meses de 2006 é de 0,531%, abaixo da meta estabelecida pelo Ministério da Saúde para esse mesmo período que é, em média, de 0,73%.

3.2.2 Série do número de inaptidões clínicas

É possível observar nas Figuras 4c e 4d que, embora a série não apresente sazonalidade, existe correlação significativa no lag 24, tanto na fac quanto na facp, sugerindo um modelo SARIMA incompleto. Dentre os modelos ajustados o que melhor representou os dados foi o SARIMA (0,1,1)(0,0,2)12 incompleto, ou seja,

$$Z_t = \frac{(1-0{,}567B)(1-0{,}206B^{24})a_t}{(1-B)}. \qquad (5)$$



(a) (b)

FIGURA 6: Representação gráfica das funções de autocorrelação (a) e função de autocorrelação (b) dos resíduos do modelo SARIMA(0,1,1) (0,0,2)12 incompleto.

Por meio da análise dos gráficos representados na Figura 6, observou-se que a fac e a facp dos resíduos não apresentam correlação significativa, indicando que at é um ruído branco. Essa análise pode ser confirmada pelo Teste de Box-Pierce, pois a estatística $Q_{40} = 37,91$ é inferior a $\chi^2_{38,0.05} = 53,38$.

As previsões apresentadas na Tabela 5 não foram boas, com um MAPE de 30,6%. O índice previsto para o período de setembro a dezembro de 2006 é 11,51% próximo ao estabelecido pelo Ministério da Saúde.

TABELA 5: Valores observados (Z_{t+h}), **valores previstos**$(\hat{Z}_t(h))$, **erro padrão** (SE), **erro de previsão** $(e_{t(h)})$ **para a série dos número de inaptidões clínicas no período de setembro a dezembro de 2006, utilizando o modelo SARIMA** $(0,1,1)(0,0,2)_{12}$ **incompleto.**

Mês	Valor observado (Z_{t+h})	Valor previsto $(\hat{Z}_t(h))$	Erro padrão (*SE*)	Erro de previsão $(e_{t(h)})$
Set/06	44	56,855	16,208	12,855
out/06	51	59,935	17,277	8,935
Nov/06	39	63,680	18,284	24,680
dez/06	57	64,040	19,238	7,040
EQMP= 225,941			*MAPE*= 30,60%	

Nota: $e_t(h) = Z_{t+h} - \hat{Z}_t(h)$

3.2.3 Série do número de inaptidões sorológicas

Pela representação gráfica da Figura 3c, observa-se que existem indícios de intervenção nas observações 43 e 59, correspondendo aos períodos de agosto de 1996 e dezembro de 1997, respectivamente.

Essas intervenções têm um efeito abrupto e permanente, sendo a função de transferência dada por: $\upsilon(B) = \omega_i$, em que i é a i-ésima intervenção.

É possível observar, também, que o lag 6 da fac (Figura 4e) é significativo, sugerindo o ajuste de um modelo SARIMA.

Considerando as análises da fac, facp (Figuras 4e e 4f) e das intervenções, o modelo que melhor se ajustou a série do número de inaptidões sorológicas foi SARIMA(0,1,1)(0,0,1)6 com intervenções em agosto/96 e dezembro/97. Esse modelo é dado por:

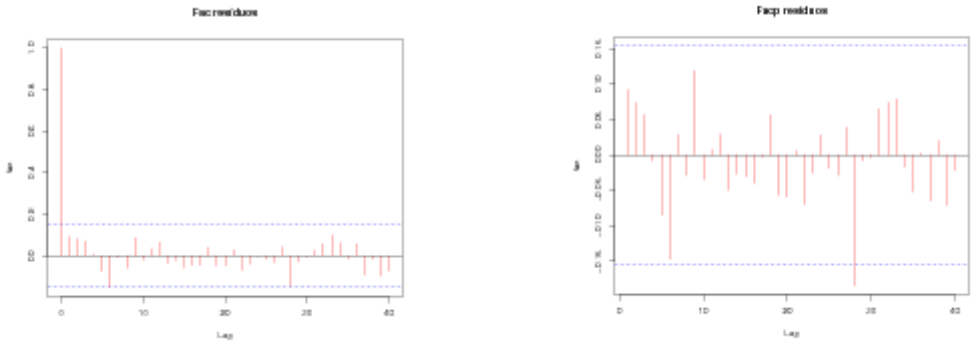
$$Z_t = 12{,}019x_{1,t} + 21{,}28x_{2,t} + \frac{(1-0{,}766B)(1+0{,}183B^6)a_t}{(1-B)}, \qquad (6)$$

para as variáveis dummy $x_{1,t}$ e $x_{2,t}$ tem-se

$$x_{1,t} = \begin{cases} 1 & \text{se } t < 43 \\ 0 & \text{se } t \geq 43 \end{cases} \qquad x_{2,t} = \begin{cases} 1 & \text{se } t < 59 \\ 0 & \text{se } t \geq 59 \end{cases} \qquad (7)$$

Por meio da análise dos gráficos representados pelas Figuras 7a e 7b e do teste de Box-Pierce $Q_{40} = 37{,}01 < \chi^2_{38,0.05} = 53{,}38$, verifica-se que os resíduos do modelo ajustado são um ruído branco.

Na Tabela 6 são apresentadas as previsões para o modelo SARIMA(0,1,1)(0,0,1)6 com intervenções em agosto/96 e dezembro/97. Pode-se dizer que o modelo ajustado mostrou um bom desempenho, pois os valores previstos para a série do número de inaptidões sorológicas apresentam MAPE de 19,96%. O índice previsto para os últimos quatro meses de 2006 foi de 1,69% que é quatro vezes menor do que preconizado pelo Ministério da Saúde.



(a) (b)

FIGURA 7: Representação gráfica das funções de autocorrelação (a) e função de autocorrelação parcial (b) dos resíduos do modelo SARIMA(0,1,1)(0,0,1)6 com intervenções em agosto/96 e dezembro/97.

Utilizando-se o modelo proposto é possível verificar que os efeitos das intervenções causaram uma diminuição em aproximadamente 12 inaptidões sorológicas, em agosto de 1996 e um aumento de 21 inaptidões sorológicas em dezembro de 1997. Não foi possível identificar os fatores que influenciaram essas intervenções.

TABELA 6: Valores observados (Z_{t+h}), valores previstos($\hat{Z}_t(h)$), erro padrão (SE), erro de previsão ($e_{t(h)}$) para a série dos número de inaptidões sorológicas no período de setembro a dezembro de 2006, utilizando o modelo SARIMA$(0,1,1)(0,0,1)_6$ com intervenções em agosto/96 e dezembro/97.

Mês	Valor observado (Z_{t+h})	Valor previsto ($\hat{Z}_t(h)$)	Erro padrão (SE)	Erro de previsão ($e_{t(h)}$)
Set/06	7	9,761	8,797	2,761
out/06	11	8,660	9,059	-2,340
nov/06	9	9,732	9,314	0,732
dez/06	7	7,770	9,562	0,770
EQMP= 2,371			*MAPE*= 19,96%	

Nota: $e_t(h) = Z_{t+h} - \hat{Z}_t(h)$

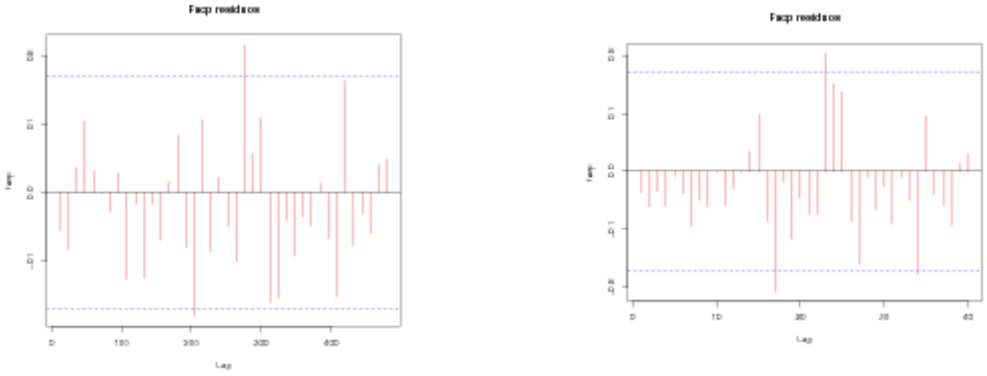
3.2.4 Série do número de doações espontâneas

Nas Figuras 4g e 4h é possível observar que, embora a série não apresente sazonalidade, existe correlação significativa nos lags múltiplos de 12 e também no lag 7, tanto na fac quanto na facp, sugerindo a utilização do modelo SARIMA incompleto.

O modelo SARIMA(7,1,1)(1,0,1) incompleto é dado por

$$Z_t = \frac{(1+,747B)(1+0,558B^{12})a_t}{(1-B)(1+0,214B^{7})(1-0,769B^{12})} \qquad (8)$$

Apesar de superparametrizado, o modelo SARIMA(7,1,1)(1,0,1) incompleto foi o único que se ajustou aos dados. Uma possível justificativa para dificuldade de ajustar um modelo a essa série seja o número de observações da mesma.



(a) (b)

FIGURA 8: Representação gráfica das funções de autocorrelação (a) e função de autocorrelação parcial (b) dos resíduos do modelo SARIMA(7,1,1)(1,0,1) incompleto.

Por meio da Figura 8b, observou-se que a facp dos resíduos não apresentou correlação significativa. Na Figura 8a, a fac apresentou correlação significativa em três lags. Pelo Teste de Box-Pierce, cujo resultado é $Q_{40} = 45,451 < \chi^2_{36,0.05} = 50,998$, tem-se at é um ruído branco.

As previsões do modelo SARIMA(7,1,1)(1,0,1) incompleto (Tabela 7) foram razoáveis, apresentando um MAPE de 26,47%.

O índice previsto para os quatro últimos meses de 2006 foi 57,78%, que é aproximadamente 40% menor que a meta estabelecida pelo Ministério da Saúde, mas está próximo do índice real (53,48%).

TABELA 7: Valores observados (Z_{t+h}), **valores previstos**$(\hat{Z}_t(h))$, **erro padrão** (SE), **erro de previsão** $(e_{t(h)})$ **para a série dos número de doações espontâneas no período de setembro a dezembro de 2006, utilizando o modelo SARIMA(7,1,1)(1,0,1) incompleto.**

Mês	Valor observado (Z_{t+h})	Valor previsto $(\hat{Z}_t(h))$	Erro padrão (SE)	Erro de previsão $(e_{t(h)})$
Set/06	279	308,96	43,99	-29,96
out/06	158	288,58	45,60	-130,58
nov/06	286	299,52	47,14	-13,52
dez/06	335	308,97	48,64	26,03
EQMP= 4702,27			*MAPE*= 26,47%	

Nota: $e_t(h) = Z_{t+h} - \hat{Z}_t(h)$

4 Conclusão

Com relação à análise descritiva dos dados de doação de sangue da Fundação Hemominas – Núcleo Regional de São João del Rei, para o ano de 2006, tem-se que os índices de inaptidões clínicas e sorológicas superaram as expectativas do Ministério da Saúde, sendo que o índice de inaptidões sorológicas é, aproximadamente, 4 vezes menor do que o recomendado pelo Ministério da Saúde.

O índice do número de doadores em relação à população total (1,62%) está abaixo da meta do Ministério da Saúde (2,2%). E finalmente, o índice de doações espontâneas (53,48%) está abaixo do estabelecido pelo Ministério da Saúde (100%).

Com relação aos modelos de séries temporais tem-se que eles são úteis para descrever as séries do número de doadores, do número de inaptidões clínicas, do número de inaptidões sorológicas e do número de doações espontâneas da Fundação Hemominas - Núcleo Regional de São João del Rei.

Os modelos ajustados para as séries do número de doadores e de inaptidões sorológicas geraram boas previsões, apresentando um MAPE de 4,84% e de 19,96%, respectivamente. Para as séries do número de inaptidões clínicas e doações espontâneas obteve-se previsões razoáveis, com o MAPE igual a 30,6% e 26,48%, respectivamente.

Os índices previstos para os quatro últimos meses de 2006 das séries do número de doadores, de inaptidões clínicas, de inaptidões sorológicas e de doações espontâneas foram de 0,531%, 1,6%, 11,5% e 58%, respectivamente, aproximando-se os índices reais para esse mesmo período.

Referências bibliográficas

BORGES, V.; Martinez, E. Z.; Bendini, M. H.; Costa, M. A. G. F.; Ferreira, S.C.L. Avaliação de fidedignidade de um instrumento voltado à satisfação do doador de sangue. Revista Brasileira de Epidemiologia, São Paulo, n. 2, p. 14–17, ago 2005. Disponível em: <http:/www.scielo.br/pdf/rbepid/v8n2/10.pdf>. Acesso em: 07 abr 2005.

BOX, G.; JENKINS, G. Time series analysis, forecasting and control. San Francisco: Holden-Day, 1976. 575 p.

CAMARA, G. O programa de sangue no Brasil: triagem sorológica e controle de qualidade. XIIa Reunión Intergubernamental INCOSUR - Chagas, Santiago, mar 2003. Disponível em: <http:/www.mex.ops-oms.org/Spanish/AD/DPC/CD/dch-XII-INCOSUR-inf-final-bra.pdf>. Acesso em: 05 jan 2006.

FARIA, E. Estudo do doador inapto sorológico da Fundação Hemominas Belo Horizonte 2001-2003. Belo Horizonte: Fundação Hemominas, 2004. 39 p.

LOUREIRO, F. Criação de um modelo de atendimento em triagem e coleta em bancos de sangue. Belo Horizonte: UFMG, 2005. 75 p. (Monografia - Especialização em Saúde Pública).

MORETTIN, P.; TOLOI, C. Análise de séries temporais. São Paulo: Edgard Blücher, 2004. 535 p.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria, 2007. Disponível em: <http://www.R-project.org>. Acesso em: 2007.

Silva, B. B. L.; Roberto, L. F.; Medeiros, P. R. G.; Gomes Netto, W. P. Otimização do processo de doação de sangue no Hemorio: análise dos cenários através de um modelo de simulação. XXV ENEGEP, Porto Alegre, out/nov 2005. Disponível em: <http:/www.gpi.ufrj.br/pdfs/artigos>. Acesso em: 07 abr 2005.

URRUTIA, D.; DUARTE, G.; JOANINI, S. Análise do perfil sócio-econômico dos doadores de sangue que apresentaram no primeiro exame sorológico resultados positivos ou inconclusivos. Perspectivas Médicas, São Paulo, n. 10, p. 14–17, jan/dez 1999. Disponível em: <http:/www.fmj.br/download/revistamedica/revista1999.pdf>. Acesso em: 07 abr 2005.

Abstract

This work had as main objective to fit models, using the methodology of Box and Jenkins, for the series of donations blood Hemominas Foundation - Regional Center of São João del Rei, with the purpose of doing forecasts, comparing them with goals proposed by the Health Ministry. Were used the series number of donors, number of clinical inaptness, number of serologic inaptness and number of spontaneous donation, all with monthly measurement. The first three were collected from January 1993 to december 2006, a total of 168 observations and the fourth was collected between august 1995 and December 2006, totaling 137 observations. The last four observations of each series have been set aside to be compared with the forecasts. The adequacy of the models was verified using the mean squared error of prediction criteria and mean absolute percentage error criteria. The indices provided for the last four months of 2006 to the series of the number of donors, medical disabilities, disabilities of serological and Spontaneous donations were 0,531%, 1,6%, 11,5%and 53%, respectively, coming close itself to the real indices for the same period.

Key-words: SARIMA model, blood donors, intervention models, time series.

Introdução a Cópulas e Aplicações na Avaliação do Desempenho de Empresas

Sumaia A. Latif[1]
Pedro A. Morettin[2]

Resumo.

O objetivo deste artigo é introduzir o conceito de dependência local através de cópula e densidade cópula no contexto de variáveis aleatórias. Para tanto, fazemos uma breve introdução sobre cópula e densidade cópula, descrevendo os seus principais tipos e métodos de estimação. Dados reais referentes a alguns indicadores do desempenho das vendas das maiores empresas no Brasil em 2006 (exceto bancos e seguradoras) foram obtidos, analisando-se as informações obtidas através de coeficientes de associação global usuais e as fornecidas por cópula e densidade cópula, sendo que estas foram estimadas através de método não paramétrico via *kernel* e semi-paramétrico.

Palavras chave: associação, dependência local, copula, densidade cópula.

[1] Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
[2] Instituto de Matemática e Estatística – Universidade de São Paulo

1 Introdução

Para avaliar a associação entre duas variáveis, usualmente são utilizados coeficientes de associação global os quais resultam num único valor numérico. Como exemplo, podemos mencionar a correlação linear de Pearson, o Rho de Spearman e o Tau de Kendall, dentre outros. Porém, os dados podem apresentar diferentes comportamentos de dependência em diferentes subregiões de variação dos dados, e então faz-se necessário o uso de ferramentas que permitam a análise da dependência local.

Algumas propostas para avaliar a dependência local no contexto de variáveis aleatórias podem ser encontradas em Holland e Wang (1987), Bjerve e Doksum (1993) e Bairamov et al. (2003), além de Sibuya (1960) e Nelsen (2006). Sob o enfoque de séries temporais, podemos mencionar o estudo de cópulas tanto por Fermanian e Scaillet (2003) quanto por Morettin et al. (2010) e Genest et al. (2009).

Cópula e densidade cópula (veja Joe, 1997, Nelsen, 2006, Anjos et al., 2004, e Embrechts, 2001, dentre outros) são medidas de dependência local que têm sido utilizadas destacadamente na análise de risco e na análise de séries financeiras. Seja no contexto de variáveis aleatórias ou de séries temporais, não encontramos na literatura artigos com aplicações desta metodologia na análise de empresas.

O objetivo deste artigo é expor os conceitos básicos de análise de dependência local através de cópula e densidade cópula no contexto de variáveis aleatórias. Para tanto, na Seção 2 fazemos uma breve introdução sobre a teoria de cópulas, fornecendo alguns exemplos de cópulas elípticas e Arquimedianas. Os principais métodos de estimação são mencionados na Seção 3. A Seção 4 apresenta a análise da associação e da dependência entre três pares de variáveis indicadoras do desempenho das vendas das maiores empresas no Brasil em 2006. Para isto, foram utilizados os coeficientes Rho de Spearman e Tau de Kendall, gráficos de dispersão dos dados e do log dos dados com e/ou sem os *outliers*, e também os gráficos de curvas de nível das cópulas e densidades cópula estimadas através dos métodos não-paramétrico suavizado por *kernel* e semi-paramétrico. Por fim, descrevemos um resumo e conclusão do trabalho na Seção 5.

2 Noções sobre cópulas

A palavra cópula surgiu pela primeira vez na literatura estatística em Sklar (1959), embora idéias e resultados similares sejam encontrados em Hoeffding (1940).

Vejamos uma definição clássica de cópula.

Definição. Uma cópula d-dimensional é uma função $C:[0,1]^d \to [0,1]$ com as seguintes propriedades:

(i) para cada $\mathbf{u}$ em $[0,1]^d$, $C(\mathbf{u}) = 0$ se pelo menos uma das coordenadas de $\mathbf{u}$ é 0, e $C(\mathbf{u}) = u_k$ se todas as coordenadas de $\mathbf{u}$ são iguais a 1, exceto u_k;

(ii) para todo $\mathbf{a}$ e $\mathbf{b}$ em $[0,1]^d$ tal que $a_i \le b_i, \forall i$, temos que $V_C([\mathbf{a},\mathbf{b}]) \ge 0$, em que $V_C([\mathbf{a},\mathbf{b}])$ representa o volume de *C* no *d*-cubo $[\mathbf{a},\mathbf{b}]$.

A propriedade (i) acima refere-se à uniformidade das distribuições marginais e a propriedade (ii) ao caráter *d*-crescente que no caso bivariado (*d*=2) significa que para (U,V) com função de distribuição *C* então $P[u_1 \le U \le u_2, v_1 \le V \le v_2] \ge 0$ com $(u_1,v_1),(u_2,v_2) \in [0,1]^2$ e tais que $u_1 \le u_2$ e $v_1 \le v_2$.

Portanto, uma cópula *C* é uma função de distribuição conjunta em $[0,1]^d$ com marginais distribuídas uniformemente em [0,1]. Além disso, observamos que para qualquer cópula com $d \ge 3$, cada marginal *k*-dimensional de *C* é uma cópula de dimensão *k*.

Teorema 1 (de Sklar). Seja *F* uma função de distribuição *d*-dimensional com marginais $F_1,\ldots,F_d$. Então existe uma cópula *d*-dimensional *C* tal que para todo ***x*** em $\overline{\Re}^d$ (em que $\overline{\Re} = \Re \bigcup \{-\infty;+\infty\}$),

$$F(x_1,...,x_d) = C(F_1(x_1),...,F_d(x_d)). \quad (1)$$

Se $F_1,...,F_d$ são todas contínuas, então *C* é única; caso contrário, *C* é unicamente determinada no conjunto $\mathrm{Im}(F_1)\times...\times\mathrm{Im}(F_d)$, em que $\mathrm{Im}(F_i)$ é a imagem de F_i. Reciprocamente, se *C* é uma cópula *d*-dimensional e $F_1,...,F_d$ são funções de distribuição, então a função F definida acima é uma função de distribuição conjunta *d*-dimensional com marginais $F_1,...,F_d$.

Pelo teorema de Sklar, que é um dos resultados mais importantes na teoria e aplicações de cópulas, vemos que a cópula é uma função de dependência que "casa" (acopla) as distribuições marginais univariadas formando uma distribuição multivariada,

ou então que uma função de distribuição multivariada pode ser decomposta nas marginais univariadas e na estrutura de dependência dada pela cópula.

A expressão da cópula (fórmula (1)) pode ser reescrita como

$$C(u_1,...,u_d) = F(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d)),$$

(2)

para qualquer $\mathbf{u}$ em $[0,1]^d$, em que F é uma função de distribuição d-dimensional com marginais $F_1,\ldots,F_d$ e cópula C, sendo $F_i^{-1}(p) = \inf\{x \in \Re / F_i(x) \geq p\}$, $\forall p \in [0,1]$, as inversas generalizadas de $F_i, i = 1,\ldots,d$, com $\inf \phi = -\infty$. Se F_i é estritamente crescente, então F_i^{-1} é a inversa usual.

Uma cópula muito referenciada é a cópula produto dada por $\Pi(\mathbf{u}) = u_1 \times \ldots \times u_d$, $\forall \mathbf{u} \in [0,1]^d$, significando que as suas componentes são independentes.

Pode-se verificar que as cópulas são invariantes sob transformações estritamente crescentes das variáveis originais (Teorema 2.4.3 de Nelsen, 2006), e os coeficientes de concordância (ou associação) Rho de Spearman (ρ_S) e Tau de Kendall (τ) também (Teorema 5.1.9 de Nelsen, 2006), sendo que estes medem o grau de dependência monotônica entre duas variáveis. Além disso, ρ_S é proporcional ao volume (acrescido do sinal) entre os gráficos de $C(u,v)$ e $\Pi(u,v)$ (veja seção 5.1 de Nelsen, 2006). Observamos que no caso em que a associação entre duas variáveis aleatórias é positiva, significa que há maior dependência entre os valores maiores das variáveis e/ou entre os valores menores das variáveis; por outro lado, quando a associação é negativa, há maior dependência entre os valores maiores de uma variável e os valores menores da outra variável.

Quanto aos limites de variação de qualquer cópula, temos que são dados pela versão cópula da desigualdade de Fréchet-Hoeffding, isto é

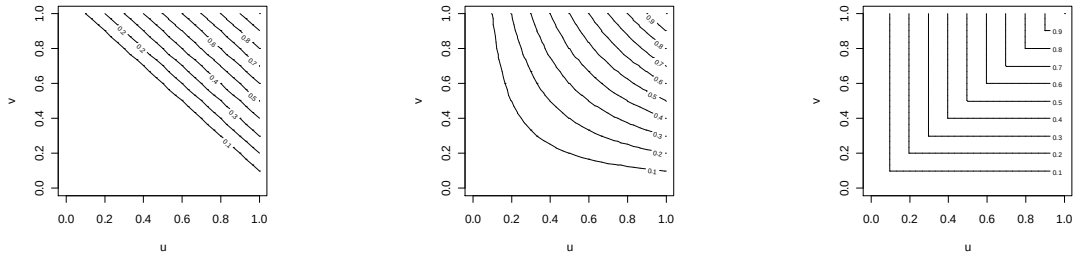
$$W^d(\mathbf{u}) \leq C(\mathbf{u}) \leq M^d(\mathbf{u}),$$

para todo $\mathbf{u}$ em $[0,1]^d$, em que *C* é uma cópula *d*-dimensional qualquer, $M^d(\mathbf{u}) = \min(u_1 + \ldots + u_d)$ é uma cópula *d*-dimensional para todo $d \geq 2$, a qual é denominada limite superior de Fréchet-Hoeffding, e $W^d(\mathbf{u}) = \max(u_1 + \ldots + u_d - d + 1, 0)$ não é uma cópula *d*-dimensional para qualquer $d \geq 3$, sendo esta o limite inferior de Fréchet-Hoeffding.
Por exemplo, no caso bivariado em que X e Y tem distribuição conjunta F e marginais contínuas F_1 e F_2, se $F(x,y) \geq F_1(x)F_2(y)$, $\forall (x,y) \in \Re^2$, então X e Y apresentam dependência de quadrante positiva, e se $F(x,y) \leq F_1(x)F_2(y)$, $\forall (x,y) \in \Re^2$,

então a dependência é de quadrante negativa. Especificamente, se (X,Y) tem cópula M então X e Y são comonotônicas (dependência positiva perfeita), e se (X,Y) tem cópula W então X e Y são ditas ser contramonotônicas (dependência negativa perfeita ou máxima).

Na Figura 1 apresentamos os gráficos de curvas de nível das cópulas W, Π e M para o caso bivariado, em que os expoentes $d=2$ foram removidos para simplificar a notação.

Figura 1. Gráficos de curvas de nível das cópulas W, Π e M em (a), (b) e (c).



(a) (b) (c)

Além da cópula, uma outra função que descreve a estrutura de dependência local, e com mais detalhes, é a densidade cópula *c*, a qual é a derivada parcial mista da cópula, ou seja

$$c(F_1(x_1),...,F_d(x_d)) = \frac{f(x_1,\ldots,x_d)}{f_1(x_1)\times\ldots\times f_d(x_d)}, \quad \forall (x_1,\ldots,x_d)\in \overline{\Re}^d,$$

que pode ser reescrita como

$$c(u_1,\ldots,u_d) = \frac{f(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1))\times\ldots\times f_d(F_d^{-1}(u_d))}. \tag{3}$$

Portanto, a densidade cópula pode ser interpretada como sendo a razão entre a densidade conjunta e o produto das densidades marginais. Então, se a razão entre estas duas quantidades resultar no valor 1 num determinado ponto bivariado da região sob análise, isto significa que as variáveis em questão apresentam independência localmente neste ponto, e se a densidade cópula resultar no valor 1 em todo o suporte das variáveis, então dizemos que as variáveis são globalmente independentes.

Os principais tipos de cópulas são as cópulas elípticas e as cópulas da família Arquimediana. As cópulas elípticas são simplesmente cópulas de distribuições elípticas

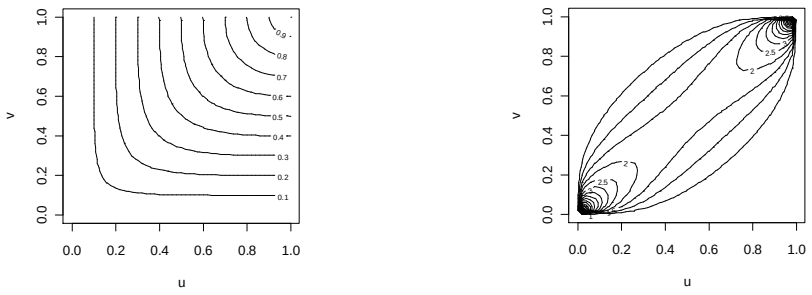
as quais referem-se a distribuições multivariadas que compartilham várias propriedades da distribuição normal multivariada e que são obtidas a partir do teorema de Sklar . Porém, estas cópulas não possuem expressões com formas fechadas e são restritas a ter simetria radial. Por exemplo, quando há maior dependência entre grandes valores de duas variáveis do que entre pequenos valores dessas variáveis (ou vice-versa), as cópulas elípticas não devem ser utilizadas. Já as cópulas Arquimedianas permitem uma grande variedade de estruturas de dependência, inclusive assimétricas, sendo que as usualmente encontradas possuem expressões com forma fechada. Tais cópulas são obtidas através do Teorema 2.

Vejamos alguns exemplos para o caso bivariado.

Seja Φ a função de distribuição univariada normal padrão e R_{12} o coeficiente de correlação linear. Então, a cópula Gaussiana é dada por

$$C(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-R_{12}^2)^{1/2}} \exp\left(-\frac{s^2 - 2R_{12}st + t^2}{2(1-R_{12}^2)}\right) dsdt .$$

Figura 2. Gráficos de curvas de nível da cópula Gaussiana em (a) e da respectiva densidade cópula em (b) com $\rho = 0{,}8$.



(a) (b)

Na Figura 2, com correlação 0,8, observamos os gráficos de curvas de nível da cópula Gaussiana em (a) que exibe dependência positiva, e em (b) da respectiva densidade cópula que além de mostrar que a dependência é positiva (*c* apresenta valores maiores que 1 em pelo menos uma das regiões além do centro da diagonal principal), mostra também que a dependência é simétrica (*c* apresenta intensidades iguais nas

regiões externas ao centro da diagonal), sendo a dependência tanto maior quanto mais nos aproximamos dos extremos da diagonal principal do plano.

A expressão cópula *t* de Student é dada por

$$C_{\upsilon,R}^{t}(u,v)=\int_{-\infty}^{t_\upsilon^{-1}(u)}\int_{-\infty}^{t_\upsilon^{-1}(v)}\frac{1}{2\pi(1-R_{12}^2)^{1/2}}\left(1+\frac{s^2-2R_{12}st+t^2}{\upsilon(1-R_{12}^2)}\right)^{-(\upsilon+2)/2}dsdt\,,$$

em que t_ν é a função de distribuição *t* de Student com $\nu>2$ graus de liberdade e R_{12} é a correlação linear. Na Figura 3, observamos em (a) e (b) os gráficos de curvas de nível da cópula e da densidade cópula *t* de Student com parâmetros 0,8 para a forma da distribuição (pois $\nu=2$) e 2 para os graus de liberdade.

Teorema 2. Seja φ uma função estritamente decrescente e contínua de $[0,1]$ em $[0,\infty]$ tal que $\varphi(1)=0$ e seja $\varphi^{[-1]}:[0,\infty]\to[0,1]$ a pseudo inversa de φ dada por

$$\varphi^{[-1]}(t)=\begin{cases}\varphi^{-1}(t) & ,0\le t\le\varphi(0)\\ 0 & ,\varphi(0)\le t\le\infty\end{cases}.$$

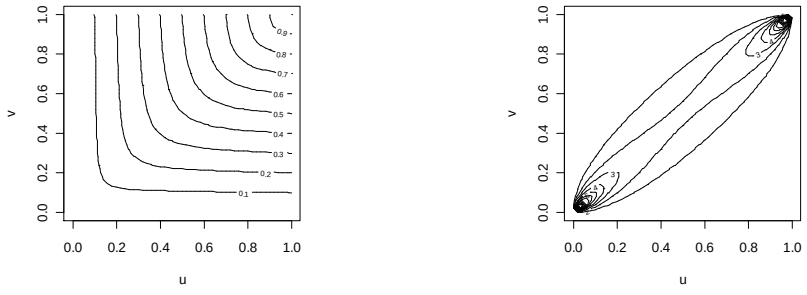
Seja $C:[0,1]^2\to[0,1]$ dada por

$$C(u,v)=\varphi^{[-1]}(\varphi(u)+\varphi(v))\,. \qquad (4)$$

Então *C* é uma cópula se e somente se φ é convexa.

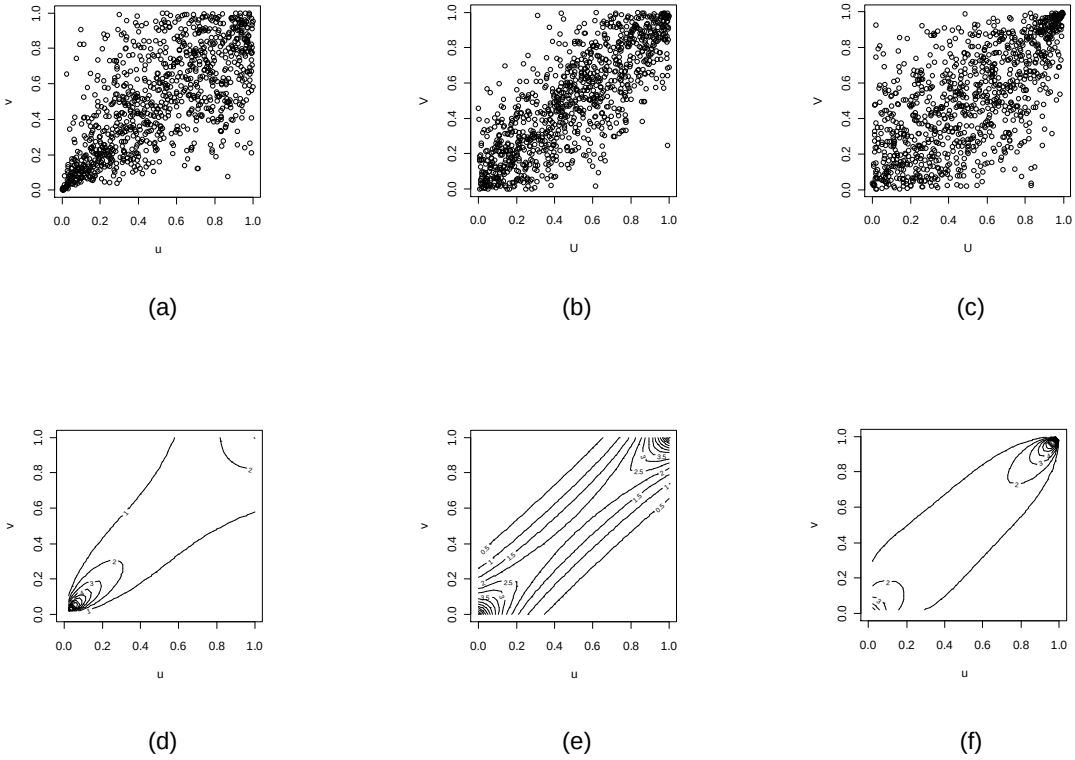
Nas cópulas Arquimedianas (fórmula (4)), a função φ denomina-se gerador da cópula, e se $\varphi(0)=\infty$ então $\varphi^{[-1]}=\varphi^{-1}$, e φ é denominada de gerador estrito e *C* de cópula estrita.

Figura 3. Gráficos de curvas de nível da cópula t de Student com parâmetros $\rho=0{,}8$ e g.l. = 2 em (a) e da respectiva densidade cópula em (b).



(a) (b)

Figura 4. Gráficos de dispersão e de curvas de nível da densidade cópula para a cópula Clayton com parâmetro 2 em (a) e (d), Frank com parâmetro 8 em(b) e (e) e Gumbel com parâmetro 2 em (c) e (f), respectivamente.



(a) (b) (c)

(d) (e) (f)

Por exemplo, considerando a fórmula (4) com $\varphi(t) = (t^{-\theta} - 1)/\theta$ em que $\theta \in [-1, \infty) \setminus \{0\}$, obtemos a família de cópulas Clayton cuja expressão resultante é

$$C_\theta(u, v) = \max([u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}, 0),$$

e se $\theta > 0$ (cópula estrita), temos que $C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$. A simulação da cópula Clayton com parâmetro 2 apresenta-se nos gráficos (a) e (d) da Figura 4, os quais referem-se à dispersão e às curvas de nível da densidade cópula, respectivamente. Podemos observar a ocorrência de dependência positiva assimétrica com dependência destacada entre os baixos valores das variáveis.

A família de cópulas Gumbel é obtida com $\varphi(t) = (-\ln t)^\theta$ em que $\theta \geq 1$, ou seja

$$C_\theta(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)) = \exp(-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}).$$

Uma simulação dessa cópula com parâmetro 2, apresenta-se através dos gráficos (c) e (f) da Figura 4, em que novamente observamos dependência positiva assimétrica, mas agora com dependência entre os maiores valores das variáveis.

Já para $\varphi(t) = -\ln((e^{-\theta t} - 1)/(e^{-\theta} - 1))$ em que $\theta \in \Re \setminus \{0\}$, obtemos a família de cópulas Frank cuja expressão é

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right),$$

que é uma cópula estrita. As cópulas da família Frank são as únicas da classe Arquimediana que apresentam simetria radial. Esta cópula simulada com parâmetro 8 apresenta-se nos gráficos (b) e (e) da Figura 4, em que vemos dependência positiva simétrica.

3 Estimação de cópulas

Considerando o caso bivariado, dada uma amostra $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ de uma variável aleatória (X,Y) com função de distribuição F, funções de distribuição marginais F_1 e F_2 e cópula *C*, queremos estimar *C*. Podemos usar estimadores paramétricos, estimadores semi-paramétricos ou estimadores não-paramétricos. No primeiro caso usamos estimadores de Máxima Verossimilhança, no segundo estimadores de Pseudo-Máxima Verossimilhança, e no terceiro caso podemos usar cópulas empíricas (baseadas em postos) ou estimadores suavizados (via kernels, ondaletas, etc).

As funções de distribuição empíricas correspondentes a F, F_1 e F_2 são dadas por

$$F_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x, Y_i \le y\}, \qquad F_{1n}(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}, \qquad F_{2n}(y) = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i \le y\},$$

em que $-\infty < x, y < +\infty$, sendo que pode-se utilizar o denominador $n+1$ para garantir que as funções estejam no intervalo $[0,1]$. Pelo lema de Glivenko-Cantelli, F_{in} aproxima-se de F_i quando $n \to \infty$, $i = 1,2$. Similarmente, pela transformada de probabilidade, espera-se que $F_{in}(X_j)$, $j = 1,2,...,n$, sejam aproximadamente uniformes, $i = 1,2$. Como vimos, a cópula *C* é também a função de distribuição de $(U,V) = (F_1(X), F_2(Y))$ se F_1 e F_2 são contínuas, de modo que $(U_i,V_i) = (F_1(X_i), F_2(Y_i))$, $i = 1,2,...,n$, formam uma amostra aleatória de *C*. Logo, se considerarmos $(\hat{U}_i,\hat{V}_i) = (F_{1n}(X_i), F_{2n}(Y_i))$, $i = 1,2,...,n$, teremos uma boa idéia de como é a verdadeira cópula C, ou seja, podemos dizer que estes pares formam uma pseudo-amostra de *C*. Note também que $(F_{1n}(X_i), F_{2n}(Y_i)) = (R_i/n, S_i/n)$, $i = 1,2,...,n$, em que R_i é o posto de X_i considerando $X_1, X_2,..., X_n$ e S_i é o posto de Y_i considerando $Y_1, Y_2,..., Y_n$. Portanto, o gráfico dos postos normalizados $(R_i/n, S_i/n)$, $i = 1,2,...,n$, é útil para evidenciar a estrutura de dependência entre X e Y.

3.1 Estimação de máxima verossimilhança

Dada a amostra (x_i, y_i), $i = 1,2,...,n$, de F bivariada com marginais F_1 e F_2, cópula C e densidade cópula *c*, então a densidade conjunta é dada por

$$f(x_i, y_i; \theta) = c(F_1(x_i, \theta_1), F_2(y_i, \theta_2); \theta_{12}) \times f_1(x_i; \theta_1) f_2(y_i; \theta_2)$$

em que $\theta = (\theta_1, \theta_2, \theta_{12})$, com θ_1 representado os parâmetros de F_1, θ_2 os de F_2 e θ_{12} os de *c*.

A log-verossimilhança é

$$l(\theta, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \log c(F_1(x_i, \theta_1), F_2(y_i, \theta_2); \theta_{12}) + \sum_{i=1}^{n} \log f_1(x_i, \theta_1) + \sum_{i=1}^{n} \log f_2(y_i, \theta_2),$$

que maximizada, resulta nos estimadores de Máxima Verossimilhança (MV). Normalmente esta é uma tarefa difícil se há muitos parâmetros. Um procedimento em dois estágios (denominado “inference for margins”, IFM) pode ser utilizado, em que no passo 1 obtemos os estimadores dos parâmetros das marginais,

$$\hat{\theta}_1 = \arg\max \sum_{i=1}^{n} \log f_1(x_i, \theta_1), \ \hat{\theta}_2 = \arg\max \sum_{i=1}^{n} \log f_2(y_i, \theta_2),$$

e no passo 2 obtemos os estimadores dos parâmetros da cópula,

$$\hat{\theta}_{12} = \arg\max \sum_{i=1}^{n} \log c(F_1(x_i, \hat{\theta}_1), F_2(y_i, \hat{\theta}_2); \theta_{12}).$$

Tais estimadores são consistentes e assintoticamente normais (veja Joe e Xu, 1996).

3.2 Estimação de pseudo-máxima verossimilhança

Neste enfoque, F_1 e F_2 são estimadas usando modelos paramétricos, funções de distribuição empíricas, ou uma combinação de funções de distribuição empíricas e ajuste de uma distribuição de valores extremos para as caudas das distribuições, por exemplo, a distribuição generalizada de Pareto. Para detalhes, veja Zivot e Wang (2006). Então, no segundo passo obtemos as pseudo-amostras para a cópula

$$(\hat{u}_i, \hat{v}_i) = (\hat{F}_1(x_i), \hat{F}_2(y_i)), \ i = 1, \ldots, n,$$

e no terceiro passo formamos a log-verossimilhança

$$l(\theta_{12}, \hat{\mathbf{u}}, \hat{\mathbf{v}}) = \sum_{i=1}^{n} \log c(\hat{u}_i, \hat{v}_i; \theta_{12})$$

e a maximizamos em relação a θ_{12} através de métodos numéricos. Este método é também chamado de MV canônica. Especificamente, se F_1 e F_2 são estimadas através de funções de distribuição empíricas, então o método é denominado de estimação semi-paramétrica (veja Genest et al., 1995).

3.3 Estimação não-paramétrica

A estimação não-paramétrica elaborada através do método empírico é dada por

$$C_n(u,v) = F_n(F_{1n}^{-1}(u), F_{2n}^{-2}(v)),\ 0 \le u,v \le 1,$$

em que F_{in}^{-1}, $i=1,2$, são os quantis empíricos. Utilizando o método suavizado via *kernel*, então a estimação pode ser representada por

$$\hat{C}(u,v) = \hat{F}(\hat{F}_1^{-1}(u), \hat{F}_2^{-1}(v)),\ 0 \le u,v \le 1,$$

em que

$$\hat{F}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_{1n}}, \frac{y-Y_i}{h_{2n}}\right),\ K(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} k(u,v)dudv,$$

para alguma função *kernel* bivariada $k: R^2 \to R$ com $\int\int k(u,v)dudv = 1$ e sequências de largura de faixa $h_{in} \to 0$, $i=1,2$, conforme $n \to \infty$.

Mais detalhes podem ser encontrados em Deheuvels (1979, 1981a, 1981b) para o primeiro método e em Fermanian et al. (2004) para o segundo.

4 Aplicações a dados reais

Com o objetivo de analisar a dependência entre os principais índices anuais de desempenho das vendas de grandes empresas, consideramos as maiores empresas privadas e estatais (exceto bancos e seguradoras) no ano de 2006 segundo o critério da revista Exame, totalizando 1018 empresas (veja http://app.exame.abril.com.br/servicos/melhoresemaiores/).

Então, os índices de vendas (US$ milhões), margem sobre as vendas (%) e lucro líquido (US$ milhões) das 864 empresas que continham estas informações foram analisados de forma bivariada utilizando-se os coeficientes de associação Rho de Spearman e Tau de Kendall, a cópula não-paramétrica via *kernel* de Chen e Huang (1999) com largura de faixa de Azzaline (1981) ou Hansen (2004) e a densidade cópula não-paramétrica via *kernel* Beta de Charpentier et al. (2007) com largura de faixa obtida de modo a trazer resultados similares à estimação da densidade cópula não-paramétrica obtida pelo princípio de reflexão (Deheuvels e Hominal, 1979, e Schuster, 1985) e pelo método *kernel* transformado (Chen, 1999 e Charpentier et al., 2007). Também, a cópula e a densidade cópula foram estimadas através do método semi-paramétrico, cujas marginais são obtidas de forma empírica. As análises foram elaboradas através do *software* R, e os resultados apresentam-se a seguir.

(1) Para as variáveis lucro líquido (US$ milhões) e margem sobre as vendas (%), obtivemos $\rho_S = 0{,}846$ e $\tau = 0{,}686$ indicando forte associação global positiva.

Os gráficos de dispersão das variáveis e do log das variáveis (transformadas por locação se o valor mínimo mostrou-se negativo) com *outliers* e sem 6 *outliers* apresentam-se em (a) e (b), respectivamente, da Figura 5, sendo que nestes dois últimos gráficos vemos que há associação positiva entre as variáveis sendo que a dispersão aumenta com o aumento do valor das variáveis.

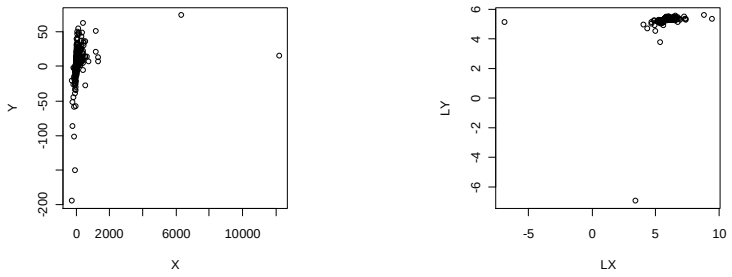
Aplicando-se o teste de independência de Genest e Rémillard (2004), o nível descritivo resultante foi próximo de zero, e então a hipótese nula (de independência) foi rejeitada.

Na Figura 6 os gráficos de dispersão dos postos normalizados (ou pseudo observações) das variáveis com *outliers* em (a) e sem os *outliers* em (b), muito similares, mostram forte dependência positiva assimétrica com maior dependência entre os menores valores das variáveis.

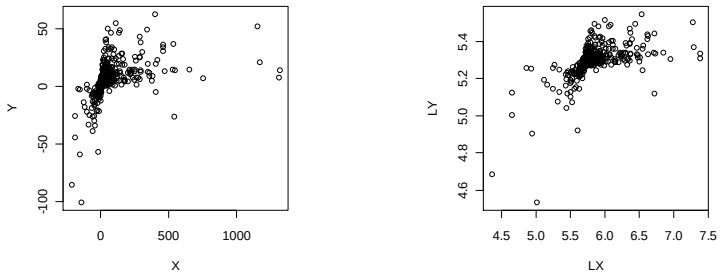
Elaborando o teste de qualidade de ajuste para cópulas com base em processo empírico, que utiliza a estatística de teste de Cramer-von Mises (veja equação (2) em Genest, Remillard e Beaudoin, 2009), considerando as variáveis com e sem *outliers*, tanto a cópula Clayton quanto a cópula Tawn (que é uma extensão assimétrica da cópula Gumbel) foram rejeitadas com nível descritivo inferior a 0,001. Então, investigamos a estrutura de dependência entre variáveis considerando a estimação via *kernel* (Beta, que utiliza pseudo observações), e também via método semi-paramétrico (por pseudo máxima verossimilhança, que utiliza pseudo observações) apenas para um comparativo visual.

Os gráficos das curvas de nível da cópula e da densidade cópula estimadas via kernel (linha cheia) e via método semi-paramétrico (linha pontilhada) com cópula Clayton para as variáveis em (a) e o log das variáveis em (b) considerando os *outliers* apresentam-se na Figura 7 e sem considerar os *outliers* apresentam-se na Figura 8. Já na Figura 9, sem os *outliers*, no método semi-paramétrico foi utilizada a cópula Tawn. Pode-se observar através das curvas de nível dos gráficos de cópulas que as estimativas via *kernel* estão coerentes, pois estão próximas das curvas de uma cópula comonotônica, e também estão mais próximas daquelas via método semi-paramétrico com cópula Clayton do que com cópula Tawn. Veja os resultados para a cópula Clayton na Tabela 1 (o cálculo com a cópula Tawn não convergiu), em que a estimação com os *outliers* apresenta menor AIC.

Figura 5 – Considerando o lucro líquido (X, em US$ milhões) e a margem sobre as vendas (Y, em %), temos os gráficos de dispersão dos dados e do log dos dados com *outliers* em (a) e sem 6 *outliers* em (b).

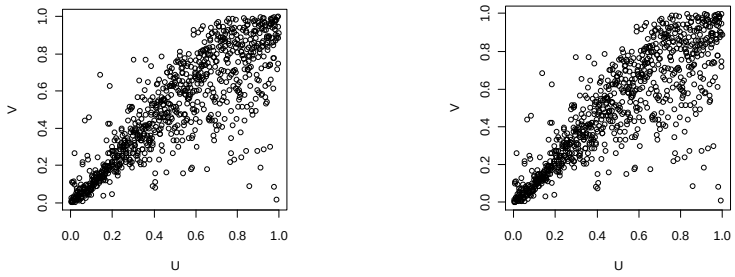


(a)



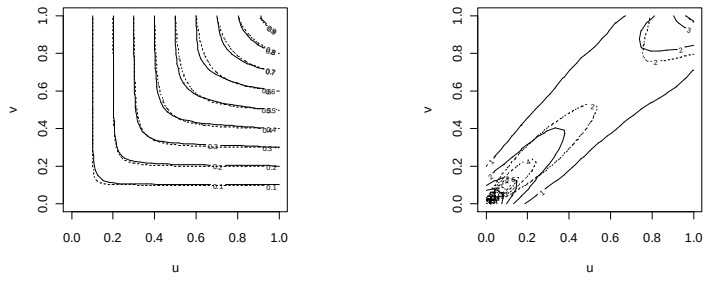
(b)

Figura 6 – Considerando o lucro líquido (X, em US$ milhões) e a margem sobre as vendas (Y, em %), temos os gráficos dos postos normalizados dos dados com *outliers* em (a) e sem 6 *outliers* em (b).

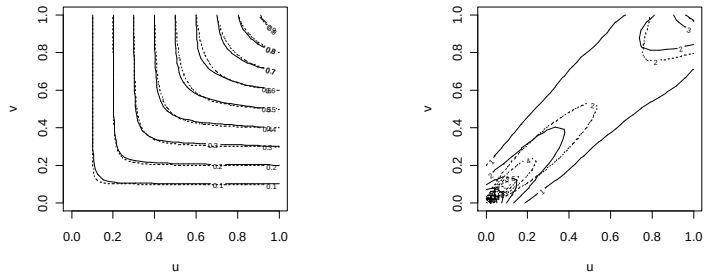


(a) (b)

Figura 7 – Considerando o lucro líquido (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em %), temos os gráficos da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Clayton, considerando os dados em (a) e o log dos dados em (b), com *outliers*.

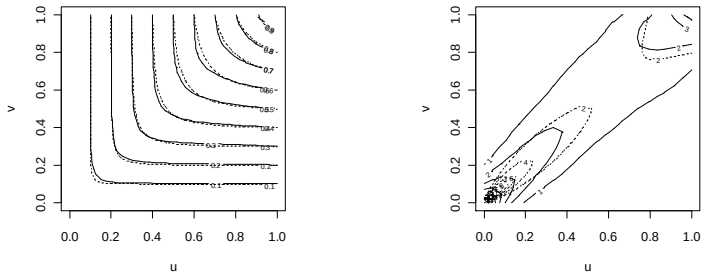


(a)

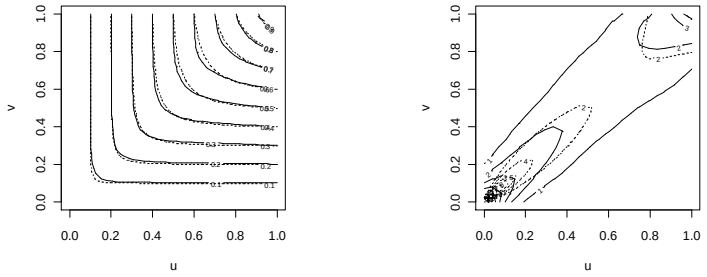


(b)

Figura 8 – Considerando o lucro líquido (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em %), temos os gráficos da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Clayton, considerando os dados em (a) e o log dos dados em (b), sem 6 *outliers*.

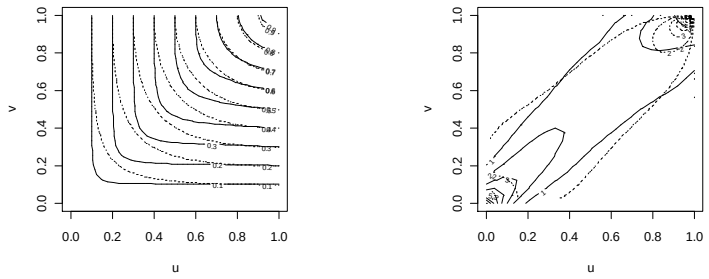


(a)

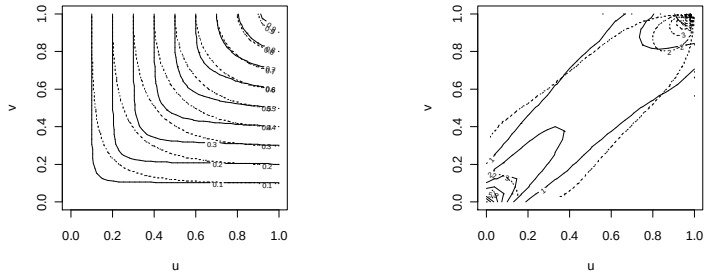


(b)

Figura 9 – Considerando o lucro líquido (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em %), temos os gráficos da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Tawn, considerando os dados em (a) e o log dos dados em (b), sem 6 *outliers*.



(a)



(b)

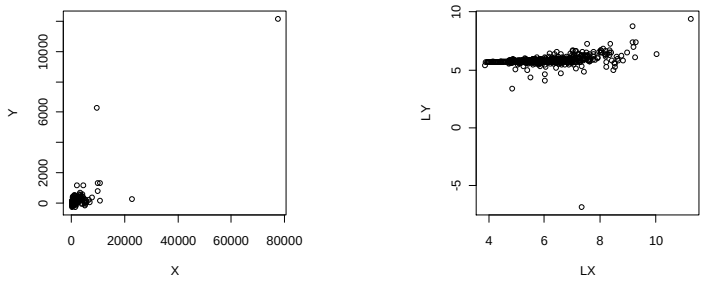
Tabela 1 – Considerando o lucro líquido (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em %), temos os resultados do ajuste semi-paramétrico por pseudo máxima verossimilhança com cópula Clayton para os dados com *outliers* e sem 7 *outliers*.

	Estimativa	Erro-padrão	Log verossimilhança	AIC
Clayton				
com *outliers*	3,566	0,145	600,53	-1199,06
sem 6 *outliers*	3,464	0,155	581,25	-1160,50

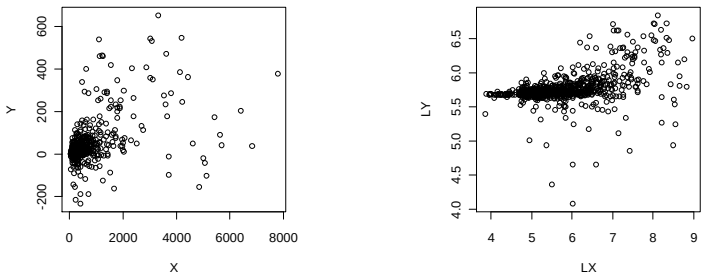
(2) Os coeficientes de associação ρ_S e τ resultaram em 0,464 e 0,336 para as vendas (US$ milhões) e o lucro líquido (US$ milhões) indicando associação positiva (para os gráficos de dispersão, veja a Figura 10), no caso entre os maiores valores das variáveis conforme pode ser observado nos gráficos dos postos normalizados da Figura 11 que utiliza as variáveis com *outliers* em (a) e sem 11 *outliers* em (b). O teste de independência foi elaborado obtendo-se níveis descritivos próximos de zero. Também foi

utilizado o teste de qualidade de ajuste considerando as cópulas Gumbel e Tawn, que resultaram em níveis descritivos menores que 0,001. As figuras 12, 13 e 14 apresentam estimativas via *kernel* comparadas com estimativas por método semi-paramétrico com cópulas Gumbel e Tawn, estas apenas para conhecimento, e não indicando visualmente uma considerável diferença de comportamento. A Tabela 2 apresenta um resumo dos resultados, sendo que o menor AIC refere-se a estimações com *outliers*, mais especificamente com cópula Tawn.

Figura 10 – Considerando as vendas (*X*, em US$ milhões) e o lucro líquido (*Y*, em US$ milhões), temos os gráficos de dispersão dos dados e do log dos dados com *outliers* em (a) e sem 11 *outliers* em (b).

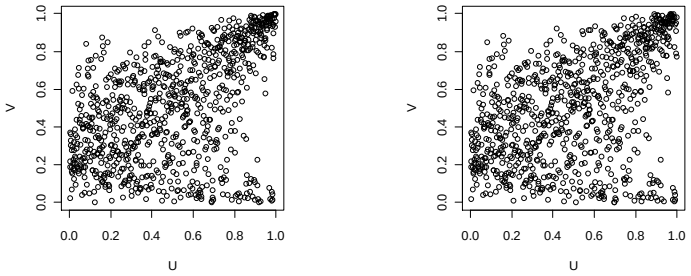


(a)



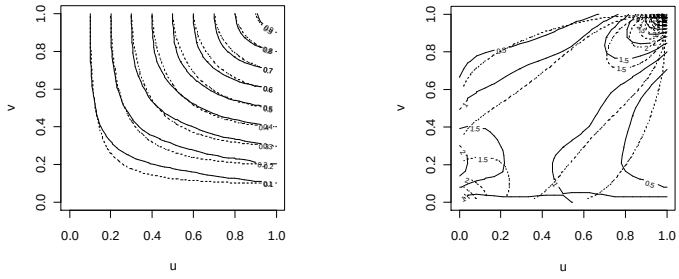
(b)

Figura 11 – Considerando as vendas (*X*, em US$ milhões) e o lucro líquido (*Y*, em US$ milhões), temos os gráficos dos postos normalizados dos dados com *outliers* em (a) e sem 11 *outliers* em (b).

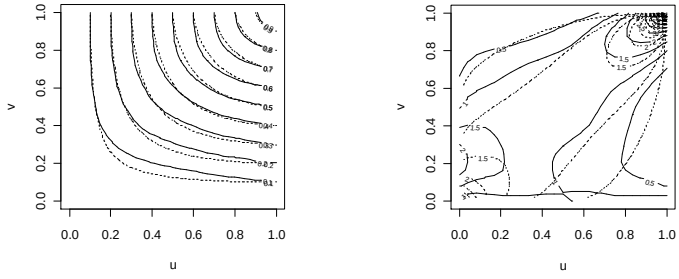


(a) (b)

Figura 12 – Considerando as vendas (*X*, em US$ milhões) e o lucro líquido (*Y*, em US$ milhões), temos os gráficos de curvas de nível da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Gumbel, considerando os dados em (a) e o log dos dados em (b), com *outliers*.

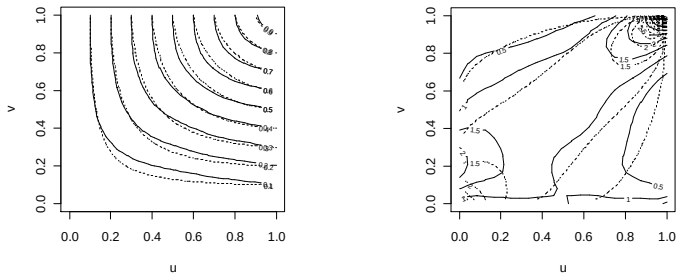


(a)

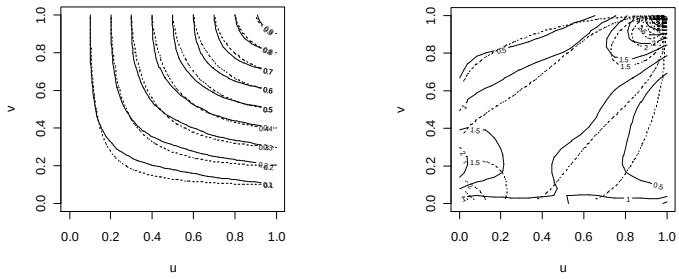


(b)

Figura 13 – Considerando as vendas (*X*, em US$ milhões) e o lucro líquido (*Y*, em US$ milhões), temos os gráficos de curvas de nível da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Gumbel, considerando os dados em (a) e o log dos dados em (b), sem 11 *outliers*.

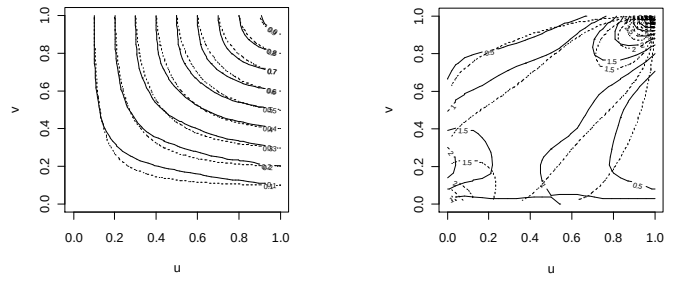


(a)

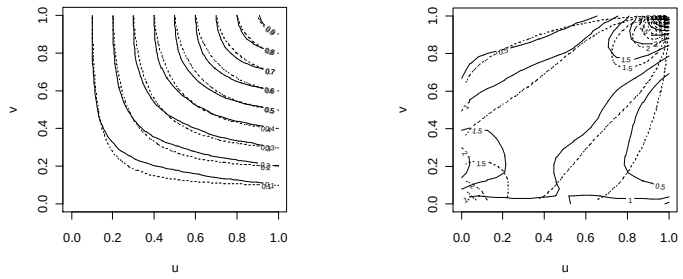


(b)

Figura 14 – Considerando as vendas (*X*, em US$ milhões) e o lucro líquido (*Y*, em US$ milhões), temos os gráficos de curvas de nível da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Tawn considerando os dados em (a) e o log dos dados em (b), sem 11 *outliers*.



(a)



(b)

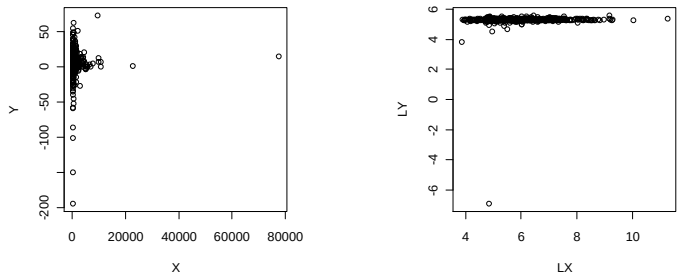
Tabela 2 – Considerando as vendas (*X*, em US$ milhões) e o lucro líquido (*Y*, em US$ milhões), temos os resultados do ajuste semi-paramétrico por pseudo máxima verossimilhança com cópula Gumbel e Tawn para os dados com *outliers* e sem 7 *outliers*.

	Estimativa	Erro-padrão	Log verossimilhança	AIC
Gumbel				
com *outliers*	1,568	0,044	167,11	-332,22
sem 6 *outliers*	1,510	0,038	140,27	-278,53
Tawn				
com *outliers*	0,856	0,036	172,57	-343,14
sem 11 *outliers*	0,828	0,038	146,07	-290,14

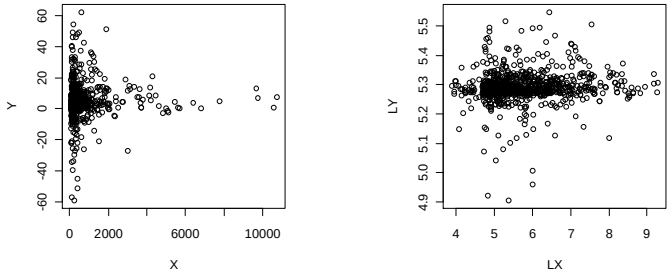
(3) Considerando agora as vendas (US$ milhões) e a margem sobre as vendas (%), temos que $\rho_S = 0{,}131$ e $\tau = 0{,}088$ indicando fraca associação positiva ou quase nula, respectivamente. Os gráficos de dispersão das variáveis e do log das variáveis com *outliers* e sem 7 *outliers* apresentam-se em (a) e (b) da Figura 15, sendo que os dois últimos não indicam associação mas sim redução de variância ao longo dos eixos das abcissas. Na Figura 16 temos os gráficos de dispersão dos postos normalizados que

indicam alguma dependência entre altos valores. Já o teste de independência resultou num nível descritivo próximo de zero. Aplicando-se o teste de qualidade de ajuste com cópulas Gumbel e Tawn aos dados com *outliers* e sem os *outliers*, obtivemos os resultados apresentados na Tabela 3, e ajustando-se estas cópulas com o método semi-paramétrico obtivemos os resultados da Tabela 4, em que o AIC é menor para os ajustes com os *outliers*, mais especificamente, com a cópula Gumbel. Portanto, ambas cópulas podem representar a estrutura de dependência bivariada em estudo. Também, podemos observar que a estimação via *kernel* apresenta as curvas de nível muito próximas daquelas da estimação semi-paramétrica (veja as figuras 17 a 19).

Figura 15 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos os gráficos de dispersão dos dados e do log dos dados com *outliers* em (a) e sem 7 *outliers* em (b).

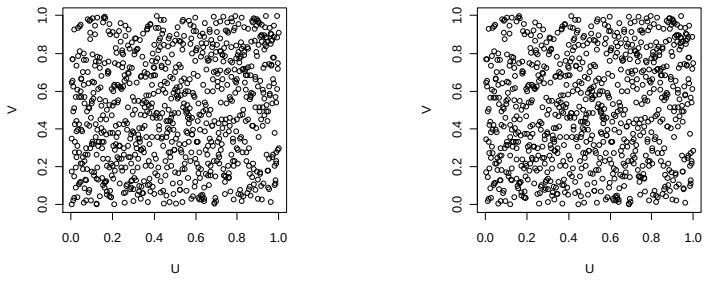


(a)



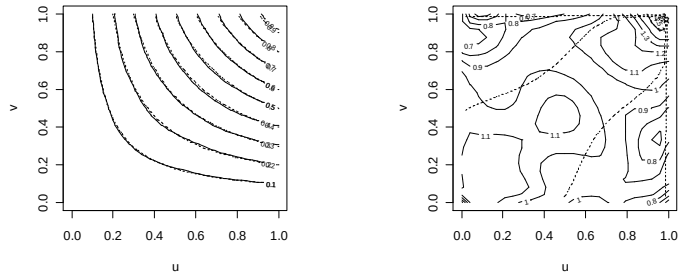
(b)

Figura 16 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos os gráficos dos postos normalizados dos dados com *outliers* em (a) e sem 7 *outliers* em (b).

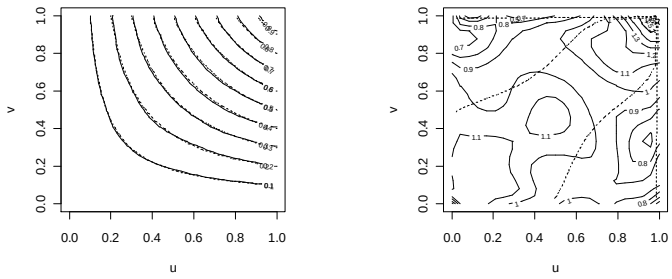


(a) (b)

Figura 17 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos os gráficos de curvas de nível da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Gumbel, considerando os dados em (a) e o log dos dados em (b), com *outliers*.

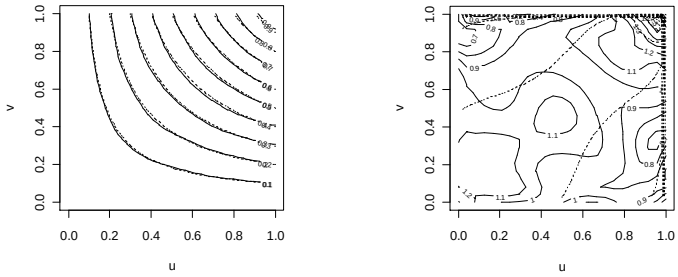


(a)

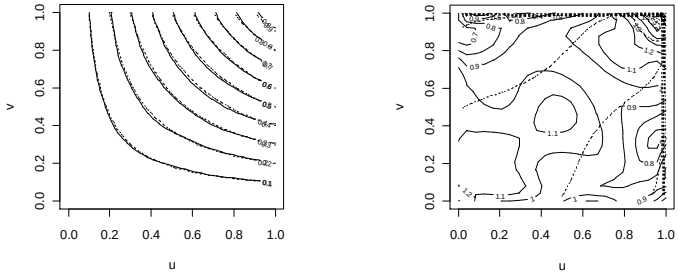


(b)

Figura 18 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos os gráficos de curvas de nível da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Gumbel, considerando os dados em (a) e o log dos dados em (b), sem 7 *outliers*.

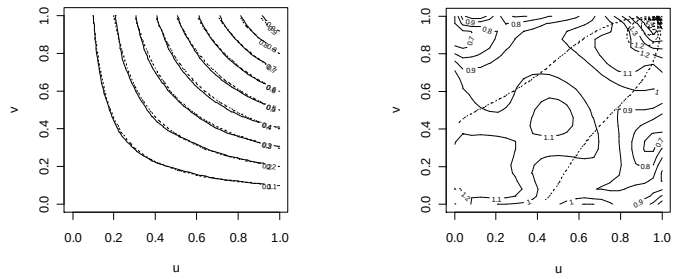


(a)

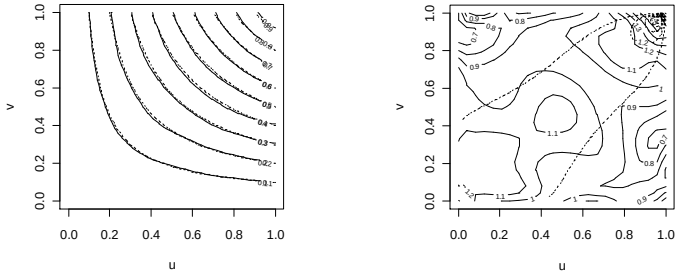


(b)

Figura 19 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos os gráficos de curvas de nível da cópula e da densidade cópula estimadas via *kernel* (em linha cheia) e por método semi-paramétrico (linha pontilhada) com cópula Tawn, considerando os dados em (a) e o log dos dados em (b), sem 7 *outliers*.



(a)



(b)

Tabela 3 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos o nível descritivo (entre parêntesis) do teste de qualidade de ajuste para a cópula Gumbel e a cópula Tawn para os dados com outliers, e o log dos dados sem 7 *outliers*.

	Gumbel (nível descritivo)	Tawn (nível descritivo)
Com *outliers*	0,252	0,499
Sem 7 *outliers*	não convergiu	0,474

Tabela 4 – Considerando as vendas (*X*, em US$ milhões) e a margem sobre as vendas (*Y*, em US$ milhões), temos os resultados do ajuste semi-paramétrico por pseudo máxima verossimilhança com cópula Gumbel e Tawn para os dados com *outliers* e sem 7 *outliers*.

	Estimativa	Erro-padrão	Log verossimilhança	AIC
Gumbel				
com *outliers*	1,084	0,025	7,48	-12,96
sem 6 *outliers*	1,067	0,024	4,07	-6,14
Tawn				
com *outliers*	0,208	0,056	6,72	-11,44
sem 11 *outliers*	0,175	0,060	4,15	-6,30

5 Resumo e conclusão

Para analisar a dependência entre duas variáveis, podemos utilizar índices de associação global em conjunto com funções de dependência local como a densidade cópula.

A correlação linear de Pearson pode ser utilizada para verificar a intensidade da associação entre duas variáveis quando a distribuição dos dados é elíptica ou esférica. Já os coeficientes Tau de Kendall e Rho de Spearman são medidas de dependência global que podem ser utilizada para qualquer distribuição dos dados, sendo esta última uma espécie de distância entre a cópula (que mede a dependência monótona) e a cópula produto (independência).

A densidade cópula é uma medida de dependência local que fornece uma compreensão mais detalhada sobre a forma da dependência local entre duas variáveis do que a cópula. Sua forma pode ser visualizada através do gráfico de dispersão dos postos normalizados (ou das pseudo observações) das variáveis de interesse quando a quantidade de pontos não é excessivamente grande. Além disso, ela pode ser estimada por método não paramétrico via *kernel*, por exemplo. Também pode-se utilizar a estimação semi-paramétrica para obter a estimativa de um (ou mais) parâmetro(s) de uma adequada família de cópulas paramétrica, ou então utilizar a estimação paramétrica em etapa única ou em duas etapas para estimar os parâmetros das distribuições marginais pertinentes e da estrutura de dependência. Sua visualização gráfica pode dar-se através de gráficos de curva de nível.

Analisando a dependência entre os três principais índices de desempenho das vendas das 864 maiores empresas (exceto bancos e seguradoras) no Brasil em 2006, observamos que há forte dependência ($\rho_S = 0{,}846$) entre lucro (US$ milhões) e margem sobre as vendas (%), destacadamente entre os baixos valores das variáveis; o teste de qualidade de ajuste rejeitou a cópula Clayton e a Tawn, sendo que a cópula Clayton estimada apresenta-se mais similar àquela estimada via kernel. Considerando as vendas (US$ milhões) e o lucro líquido (US$ milhões), obteve-se $\rho_S = 0{,}455$ sendo que a dependência concentra-se entre os maiores valores das variáveis; as cópulas Gumbel e Tawn foram rejeitadas pelo teste de qualidade de ajuste, mas seus gráficos de curvas de

nível de cópula são similares entre si e similares ao gráfico com estimação por *kernel*; o menor AIC refere-se ao da cópula Tawn. Finalmente, as vendas (US$ milhões) e a margem sobre as vendas (%) apresentam $\rho_S = 0{,}131$ e alguma dependência entre alguns dos maiores valores das variáveis; o teste de qualidade de ajuste não rejeitou a cópula Gumbel nem a cópula Tawn, e o menor AIC refere-se à cópula Gumbel; as curvas de nível das cópulas apresentaram-se muito similares entre si.

Referências bibliográficas

Anjos, U. U., Ferreira, F. H., Kolev, N. V., Mendes, B. V. M.. Modelando dependências via cópulas. 16º. SINAPE, 2004.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika, v.68, n.1, 326-328.

Bairamov, I., Kotz, S., Kozubowski, T. J.. A new measure of linear local dependence. Statistics, v.37, n.3, p.243-258, 2003.

Bjerve, S., Doksum, K.. Correlation curves: Measures of association as functions of covariate values. Annals of Statistics, v.21, n.2, p.890-902, 1993.

Charpentier, A., Fermanian, J.-D., Scaillet, O.. The estimation of copulas: theory and practice. In: Rank, J. (Ed.), Copulas: From Theory to Application in Finance, p.35-60. London: Risk Books, 2007.

Chen, S. X.. Beta kernel estimator for density functions. Computational Statistics and Data Analysis, v.31, p.131-145, 1999.

Chen, S. X., Huang, T.-M.. Nonparametric estimation of copula functions for dependence modelling. Canadian Journals of Statistics, v.31, p.131-145, 1999.

Deheuvels, P.. La fonction de de´pendance empirique et ses proprie´te´s. Acad. Roy. Belg. Bull. Cl. Sci., v.65, n.5, p.274–292, 1979.

Deheuvels, P.. A Kolmogorov–Smirnov type test for independence and multivariate samples. Rev. Roumaine Math. Pures Appl., v.26, n.2, p.213–226, 1981.

Deheuvels, P.. A nonparametric test for independence. Publ. Inst. Statist. Univ. Paris, v.26, n.2, p.29–50, 1981.

Embrechts, P., Lindskog, F., McNeil, A.. Modelling dependence with copulas and applications to risk management. Publications of ETHZ, 2001.

Fermanian, J. D., Radulovic, D., Wegkamp, M. (2004). Weak convergence of empirical copula processes. Bernoulli 10(5), 847-860.

Fermanian, J.-D., Scaillet, O.. Nonparametric estimation of copulas for time series. Journal of Risk, v.5, p.25-54, 2003.

Genest, C., Ghoudi, K., Rivest, L.-P.. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika. v.82, p.543-552, 1995.

Genest, C., Masiello, E., Tribouley, K.. Estimating copula densities through wavelets. Insurance: Mathematics and Economics. v.44, n.2, p.170-181, 2009.

Genest, C., Rémillard, B.. Tests of Independence and Randomness Based on the Empirical Copula Process. Test, v.13, n.2, p.335-369, 2004.

Genest, C., Rémillard, B., Beaudoin, D..Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics, v.44, p.199-214, 2009.

Hansen, B. E. (2004). Bandwidth selection for nonparametric distribution estimation. Working paper.

Hoeffding, W.. "Masstabinvariante korrelationstheorie," Schriften des Matematischen Instituts und des Instituts für Angewandte Mathematik der Univertät Berlin, v.5, Heft 3, p.179-233, 1940.

Holland, P. W., Wang, Y, J. Dependence function for continuous bivariate densities. Communications in Statistics Theory and Methods, v.16, n.3, p.863-876, 1987.

Joe, H.. Multivariate Models and Dependence Concepts. New York: Chapman & Hall, 1997.

Joe, H., Xu, J.J.. The estimation method of inference functions for margins for multivariate models. Department of Statistics, University of British Columbia: Technical Report, n.166, 1996.

Nelsen, R. B.. An introduction to copulas. Second Edition. New York: Springer-Verlag, 2006.

Morettin, P. A., Toloi, C. M. C., Chiann, C., de Miranda, J. C. S.. Wavelet smoothed empirical copula estimators. Brazilian Review of Finance, v.8, p.263-281, 2010.

Sibuya, M. Bivariate Extreme Statistics. Annals of the Institute of Statistical Mathematics, v.11, p.195-210, 1960.

Sklar, A.. Fonctions de répartition à n dimensions e leurs marges. Publications de l'Institut de Statistique de l'Université de Paris, v.8, p.229-231, 1959.

Zivot, E., Wang, J.. Modeling financial time series with S-plus. Second Edition. New York: Springer, 2006.

Agradecimentos

Este trabalho contou com o apoio da FAPESP, projeto 08/51097-6, e da CAPES/Procad, projeto 177/2007. Os comentários e sugestões dos revisores foram importantes para aprimorar a versão final.

Abstract

The aim of this work is to introduce the concept of local dependence through copula and copula density for random variables. For this, we make a brief introduction about copula and copula density, describing their main types and methods of estimation. Real data of some indicators of sales performance of the largest companies (except banks and insurers) in Brazil in 2006 were obtained, analysing the information provided by global coefficients of association and by copula and copula density, which were estimated by both the nonparametric method given by kernels and the semiparametric method.

Key words: association, local dependence, copula, copula density.

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 111

Centro, Rio de Janeiro – RJ

CEP: 20031-170

Tels.: 55 21 2142-4684 (Sandra Cavalcanti Barros – 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.