Presidente da República
Fernando Henrique Cardoso

Ministro do Planejamento, Orçamento e Gestão
Martus Antônio Rodrigues Tavares

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Sérgio Besserman Vianna

Diretor-Executivo
Nuno Duarte da Costa Bittencourt

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Maria Martha Malard Mayer

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Paulo Roberto Ribeiro da Cunha

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Kaizô Iwakami Beltrão

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 61 número 215 janeiro/junho 2000

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 61, n. 215 p. 1-104, jan./jun. 2000

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

Revista Brasileira de Estatística, ISSN 0034-7175

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

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

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

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

Pedro Luis do Nascimento Silva
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

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

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

Editoração
Helem Ortega da Silva - Departamento de Metodologia - DPE

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2001

Capa
Renato J. Aguiar – Gerência de Criação – CDDI

Ilustração da Capa
Marcos Balster – Gerência de Criação – CDDI

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

v.

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

Continuação de: Revista de economia e estatística.

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

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

ISSN 0034-7175 = Revista brasileira de estatística.

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

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

Impresso no Brasil/*Printed in Brazil*

Sumário

NOTA DO EDITOR

Este novo número da RBEs, com quatro artigos tratando de aplicações em áreas distintas, incluindo aplicações de métodos de amostragem, estimação para pequenas áreas e análise de séries temporais. Abrindo o número, Farias apresenta os estudos que levaram à definição da amostra da Pesquisa Industrial Mensal de Emprego e Salário do IBGE, pesquisa da qual são derivados os indicadores conjunturais da evolução do emprego e salário na indústria brasileira. Em seguida, Castañeda e Moura apresentam uma aplicação de métodos Bayesianos para estimar índices de proficiência escolar para pequenas áreas do Município do Rio de Janeiro, contribuindo para o debate atual sobre avaliação de desempenho escolar e para a disseminação de métodos aplicáveis à estimação para pequenas áreas ou domínios. Toloi e Echeverry fazem uma revisão de testes para comparação de séries temporais, demonstrando aplicabilidade com uma análise de séries de temperatura e salinidade de água, medidas em profundidades diferentes. Finalmente, Borgatto e Sáfadi desenvolvem uma análise de séries de número de passageiros por ônibus, número de assaltos em ônibus e número de acidentes com ônibus, para avaliar os efeitos de intervenções que afetaram o sistema de transporte coletivo urbano do Município de São Paulo, usando métodos de análise de intervenção em séries temporais.

Como já dissemos aqui, a variedade de aplicações ilustra bem como a Estatística vem contribuindo para o enfrentamento e solução de problemas relevantes e atuais (mensuração e acompanhamento conjuntural do emprego na indústria; avaliação educacional; monitoramento ambiental; análise dos efeitos de políticas públicas sobre o setor de transportes coletivos).

Agradecemos novamente pela dedicada colaboração dos revisores que têm avaliado artigos submetidos à RBEs. Aos autores em potencial, convidamos a submeter seus trabalhos, que serão avaliados com base na política editorial em vigor.

Saudações,

Pedro Luis do Nascimento Silva

Editor Responsável

Estudos para definição da amostra da Pesquisa Industrial Mensal de Emprego e Salário

Ana Maria Lima de Farias*

RESUMO

Neste artigo são apresentados os estudos desenvolvidos para definir o desenho amostral da Pesquisa Industrial Mensal de Emprego e Salário. Dada a forte assimetria da distribuição da variável Pessoal Ocupado utilizada na definição do desenho, a estratificação da população foi feita definindo-se um estrato certo com as grandes unidades, que são incluídas na amostra com certeza. Para o restante da população, utilizou-se amostragem aleatória estratificada. Uma ênfase especial é dada ao método de seleção e rotação da amostra, que foram feitas baseadas nos Números Aleatórios Permanentes.

Palavras-chave: distribuição assimétrica; estrato certo; rotação de amostra.

1. Introdução

Este texto tem como objetivo documentar os estudos feitos para a definição da amostra da nova Pesquisa Industrial Mensal de Emprego e Salário - PIMES -, que virá substituir a atual Pesquisa Industrial Mensal - Dados Gerais - PIM-DG -, ambas sob responsabilidade do Departamento de Indústria - DEIND - do Instituto Brasileiro de Geografia e Estatística - IBGE. Nessa reformulação, o objetivo central da pesquisa continua o mesmo: fornecer estimativas de indicadores conjunturais de emprego e salário na indústria. No entanto, para caracterizar mais especificamente os objetivos da pesquisa e as mudanças sendo feitas com relação à pesquisa atual, é necessário especificar a população-alvo, a população de referência, a unidade de investigação e os domínios de

* Endereço para correspondência: Deptº de Estatística - UFF - Rua Álvaro Ramos, 451 ap. 202 – Botafogo – 22280-110 – Rio de Janeiro – RJ - e-mail: amlima.ntg@terra.com.br.

análise para os quais se pretende produzir as estimativas. Cada um desses tópicos será abordado nas seções seguintes, que estão organizadas da seguinte forma: nas próximas três seções descrevem-se as variáveis a serem pesquisadas, a população alvo, o cadastro de seleção e os domínios de análise. Na **seção 5,** são apresentados os estudos para definir os estratos a serem utilizados no desenho amostral por amostragem estratificada. Na **seção 6,** é apresentado o desenho amostral final da pesquisa. Na **seção 7,** apresenta-se o mecanismo de seleção e rotação da amostra, ilustrando-se os procedimentos com um exemplo. Finalmente, na última seção é apresentado um resumo dos resultados e decisões que levaram ao desenho final da pesquisa.

2. Os Indicadores

- As variáveis atualmente investigadas são:
- Pessoal ocupado na produção (POP);
- Admissões;
- Desligamentos;
- Número de horas pagas na produção;
- Valor dos salários contratuais pagos ao POP;
- Valor das horas extras pagas ao POP;
- Valor da folha de pagamento; e
- Valor da produção.

Para essas variáveis, são estimados apenas índices mensais e anuais, sem a divulgação dos totais envolvidos.

- Na nova proposta, as variáveis passam a ser:
- Pessoal ocupado total;
- Admissões;
- Desligamentos;
- Número de horas pagas na produção; e
- Valor da folha de pagamentos.

E serão produzidas estimativas de totais e os índices mensais e anuais.

3. Unidade de investigação, população-alvo e cadastro de seleção

Dada a necessidade da pesquisa produzir estimativas por região e atividade industrial, a unidade de investigação da PIMES foi definida como a unidade local[1] (UL) e não a empresa[2]. A população-alvo é definida

[1] Unidade Local é o endereço de atuação de uma empresa, ocupando geralmente uma área contínua na qual são desenvolvidas uma ou mais atividades econômicas.

[2] Empresa é a unidade jurídica que responde por uma firma ou Razão Social, englobando o conjunto de atividades econômicas exercidas em uma ou mais unidades locais.

como o conjunto de todas as ULs industriais (seções C e D da Classificação Nacional de Atividades Econômicas - CNAE) ativas no ano da pesquisa.

Com relação ao cadastro de seleção da amostra, foram consideradas duas possibilidades: o Cadastro Central de Empresas - CEMPRE - do IBGE e o cadastro de informantes da Pesquisa Industrial Anual - PIA -, também realizada pelo DEIND. A segunda alternativa significaria tratar a amostra da PIMES como uma sub-amostra da PIA.

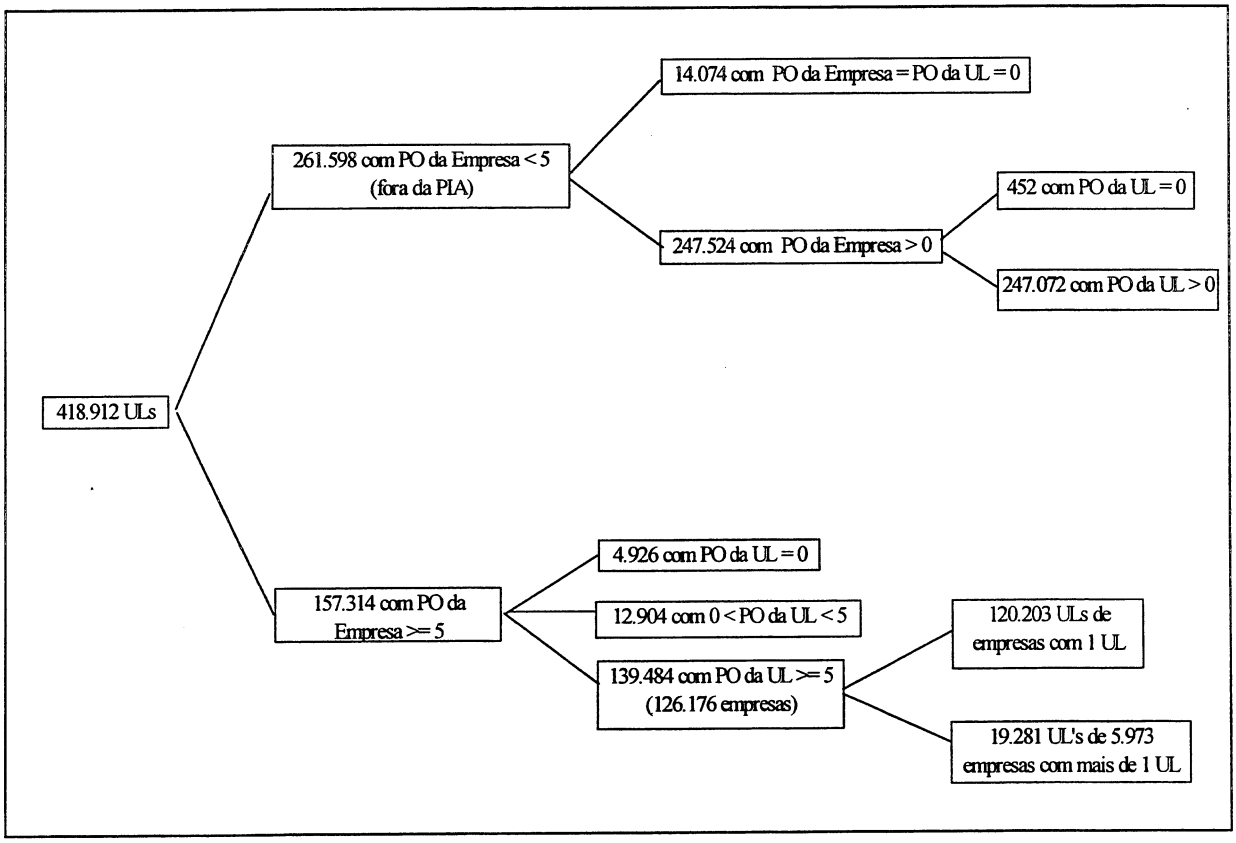
A grande vantagem de considerar a amostra da PIMES como uma subamostra da PIA seria garantir uma alta taxa de informantes em atividade; no entanto, embora essa pudesse ser uma vantagem em termos operacionais, é provável que um viés fosse introduzido nas estimativas, uma vez que estaríamos considerando apenas informantes ativos. Além disso, como as unidades de investigação da PIMES e da PIA são distintas (na PIA, a unidade de investigação é a empresa), a possibilidade de unificação de procedimentos fica bastante reduzida. Sendo assim, decidiu-se trabalhar com o CEMPRE, o que garante uma maior atualização do cadastro de seleção. [Para maiores detalhes sobre as vantagens e desvantagens de cada alternativa, veja Farias (1999).]

Uma versão preliminar do cadastro básico de seleção da amostra da PIMES foi gerada pela Diretoria de Informática, constando de 420 656 observações, correspondentes às ULs industriais de 389 022 empresas industriais. Dessas 420 656 ULs, 1 744 não tinham informação sobre o Pessoal Ocupado (PO). Como essa informação é imprescindível para os estudos do desenho amostral, optou-se por excluir estas ULs; sendo assim, os estudos seguintes baseiam-se no cadastro formado pelas 418 912 ULs restantes, que correspondem a 387 974 empresas.

Das ULs remanescentes, 261 598 (62,45%) pertencem às empresas com PO < 5, ou seja, às empresas que estão fora do âmbito da Pesquisa Industrial Anual. Dessas, 14 074 têm PO nulo, tanto para a empresa quanto para a UL. As 247 524 ULs pertencentes às empresas com PO não-nulo, mas menor que 5, pertencem a 245 712 empresas, mas 452 têm PO nulo.

Para as 157 314 ULs pertencentes às empresas do âmbito da PIA, 4 926 têm PO = 0, enquanto 12 904 têm PO < 5. Na **Figura 1**, temos o esquema desse cadastro intermediário.

Figura 1- Composição do cadastro de seleção da PIMES



No total, há 399 460 (247 072+12 904+139 484) ULs com PO não-nulo empregando 5 573 042 pessoas. No entanto, as 259 976 ULs com PO menor que 5 empregam apenas 509 246 pessoas, correspondendo a 9,14% do pessoal total. Sendo assim, a exemplo do que já é feito na PIA, a população de referência da PIMES será definida como o conjunto das ULs com 5 ou mais pessoas empregadas. Logo, o cadastro final para seleção da amostra da PIM-DG será formado pelas 139 484 ULs industriais com PO ≥ 5.

4. Domínios de análise

Os domínios de análise estão baseados na localização geográfica e na atividade industrial desenvolvida na UL.

4.1 – Localização geográfica

A proposta inicial para os domínios de análise considerava o seguinte detalhamento para a localização geográfica:

- Regiões Norte e Centro-Oeste;

- Região Nordeste;
- Minas Gerais;
- Rio de Janeiro;
- São Paulo;
- Região Sudeste; e
- Região Sul.

A diferença fundamental entre essa proposta e a versão atual da PIM-DG diz respeito ao Estado do Espírito Santo: na versão atual, esse estado está incluído no estrato denominado Complemento Brasil, junto com as Regiões Norte e Centro-Oeste, para o qual não se divulgam as estimativas. Na nova versão, esse estado apenas complementa a Região Sudeste.

Uma proposta alternativa consistiu em desmembrar os estados da Região Sul e alguns da Região Nordeste, de modo que as regiões consideradas passam a ser:

- Regiões Norte e Centro-Oeste;
- Ceará;
- Pernambuco;
- Bahia;
- Região Nordeste;
- Minas Gerais;
- Rio de Janeiro;
- São Paulo;
- Região Sudeste;
- Paraná;
- Santa Catarina; e
- Rio Grande do Sul.

4.2 – Atividade econômica

A proposta inicial considera a atividade econômica definida a partir da Classificação Nacional de Atividades Econômicas - CNAE -, trabalhando-se com o nível de 2 dígitos (divisão), o que totaliza 27 divisões. Como esse detalhamento poderia resultar exagerado, propôs-se alternativamente um agrupamento de divisões, conforme detalhado na **Tabela 1**.

Tabela 1 - Proposta de classificação de atividades para a PIMES

Divisões da CNAE (2 dígitos)	Agrupamento de divisões	
10 Extração de carvão mineral 11 Extração de petróleo e serviços correlatos 13 Extração de minerais metálicos 14 Extração de minerais não-metálicos	10+11+13+14	Indústrias Extrativas
15 Fabr. de produtos alimentícios e bebidas 16 Fabr. de produtos do fumo	15+16	Fabricação de alimentos, bebidas e produtos do fumo
17 Fabr. de produtos têxteis	17	Fabricação de produtos têxteis
18 Confecção de artigos do vestuário e acessórios	18	Confecção de artigos do vestuário e acessórios
19 Preparação de couros e fabr. de artefatos de couro, artigos de viagem e calçados	19	Indústria do calçado, inclusive preparação de artigos de couro
21 Fabr. de produtos de celulose, papel e produtos de papel 22 Edição, impressão e reprodução de gravações	21+22	Indústria do papel e gráfica
23 Fabr. de coque, refino de petróleo, elaboração de combustíveis nucleares, prod. de álcool	23	Coque, refino de petróleo, combustíveis nucleares e álcool
24 Fabr. de produtos químicos	24	Fabricação de produtos químicos
25 Fabr. de produtos de borracha e plástico	25	Fabricação de produtos de borracha e plástico
26 Fabr. de produtos de minerais não-metálicos	26	Fabricação de produtos de minerais não-metálicos
27 Metalurgia básica	27	Metalurgia básica
28 Fabr. de produtos de metal, exclusive máquinas e equipamentos	28	Fabr. de produtos de metal, exclusive máquinas e equipamentos
29 Fabr. de máquinas e equipamentos 30 Fabr. de máquinas para escritório e equipamentos de informativa	29+30	Fabr. de máquinas equipamentos, exclusive elétricos, eletrônicos, de precisão e de comunicação
31 Fabr. de máquinas, aparelhos e materiais elétricos 32 Fabr. de material eletrônico e de aparelhos e equipamentos de comunicação 33 Fabr. de equip. instr. médico-hospitalares,instr. precisão e óticos, equip. para automação industrial, cronômetros e relógios	31+32+33	Fabr. de máquinas e aparelhos elétricos, eletrônicos, de precisão e de comunicação
34 Fabr. e montagem de veículos automotores, reboques e carrocerias 35 Fabricação de outros equipamentos de transporte	34+35	Fabricação de meios de transporte
20 Fabr. de produtos de madeira 36 Fabr. de móveis e indústrias diversas 37 Reciclagem	20+36+37	Fabricação de outros produtos da indústria de transformação

5. Desenho da amostra

Das variáveis existentes no cadastro de seleção, optou-se por utilizar o número de pessoas ocupadas (PO) para definir o desenho da amostra. Sendo assim, a amostra será desenhada para garantir um determinado coeficiente de variação (CV) para o total do PO em cada estrato natural.

Para permitir a estimação dos indicadores para os domínios de análise desejados, a população foi dividida em estratos naturais construídos a partir dos cruzamentos da localização geográfica e da atividade econômica. Considerando, então, as duas propostas de domínios de análise anteriores, os dois conjuntos de estratos naturais são definidos da seguinte forma:

Proposta Inicial (189 estratos naturais):

- Regiões Norte e Centro-Oeste; Região Nordeste; Minas Gerais; Espírito Santo; Rio de Janeiro; São Paulo; Região Sul; e
- 27 divisões.

Proposta alternativa (192 estratos naturais):

- Regiões Norte e Centro-Oeste; Ceará; Pernambuco; Bahia; Região Nordeste exclusive Ceará, Pernambuco e Bahia; Minas Gerais; Espírito Santo; Rio de Janeiro; São Paulo; Paraná; Santa Catarina; Rio Grande do Sul; e
- 16 agrupamentos de divisões.

Dada a forte assimetria da distribuição da variável PO, decidiu-se dividir a população em cada estrato natural em um estrato certo (*take-all*) e um ou mais estratos amostrados (*take- some*) [veja Hidiroglou (1986) e Lavallée-Hidiroglou (1988)]. A questão que se coloca, agora, é a definição desses estratos finais, construídos a partir da variável PO.

5.1 - Estratos finais definidos segundo Hidiroglou (1986)

Hidiroglou (1986) propôs um método para estratificação de uma população assimétrica em um estrato certo (*take-all*) e outro estrato amostrado (*take-some*), do qual uma amostra aleatória simples é retirada sem reposição. O algoritmo proposto apresenta o valor ótimo de corte que minimiza o tamanho da amostra para um dado coeficiente de variação (CV) do estimador do total da variável de análise.

Na **Tabela 2**, temos o tamanho total da amostra, considerando-se os valores de 10% e 15% para o CV, para as duas propostas de estratos naturais. Os tamanhos de amostra são bastante altos, mesmo para CV=15%, o que nos leva a excluir a possibilidade de adoção desse método.

Tabela 2 - Tamanho da amostra pelo método de Hidiroglou (1986)

Estratos	Estratos naturais			
	Proposta inicial		Proposta alternativa	
	CV=10%	CV=15%	CV=10%	CV=15%
Amostrados	4 270	3 245	4 241	3 402
Certos	5 089	3 249	5 438	3 554
TOTAL	9 359	6 494	9 679	6 956

5.2 - Estratos finais definidos pelo método de Lavallée-Hidiroglou (1988)

Lavallée e Hidiroglou propuseram, em 1988, um algoritmo iterativo para decompor uma população assimétrica em um estrato certo e um dado número de estratos amostrados, de modo a minimizar o tamanho total da amostra, dados o coeficiente de variação desejado para o estimador do total da variável de análise e o método de alocação da amostra nos estratos amostrados.

Na **Tabela 3**, temos o tamanho da amostra obtido para o total de 4 estratos (3 amostrados + 1 certo), CV=10% e alocação de Neyman nos estratos amostrados. Por questões numéricas do algoritmo, só foram feitos os cálculos para os estratos naturais com população maior que 19.

Tabela 3 - Tamanho da amostra pelo método de Lavallée-Hidiroglou (1988)

CV=10%

Estratos	Estrato natural	
	Proposta inicial	Proposta alternativa
1	352	368
2	460	487
3	560	582
Certo	887	1 127
TOTAL	2 265	2 574

Embora haja uma forte redução no tamanho da amostra, uma desvantagem da aplicação desse método em uma pesquisa mensal é a variabilidade dos limites dos estratos, o que pode dificultar o processo mensal de crítica e controle da amostra. O próximo passo, então, consistiu em analisar os limites resultantes do algoritmo para tentar unificar a estratificação. Nesse sentido foi feita uma análise dos resultados gerais para os limites dos diversos estratos para as duas propostas, obtendo-se os resultados apresentados na **Tabela 4**.

Tabela 4 - Limites dos estratos – Lavallée-Hidiroglou (1988)

Domínios de análise	Média	Mediana	Mínimo	Máximo
Proposta inicial				
Estrato 1	20,67	17	6	73
Estrato 2	93,03	65	10	580
Estrato 3	449,48	284	24	3 537
Proposta alternativa				
Estrato 1	19,41	17	7	69
Estrato 2	81,93	61	13	473
Estrato 3	395,11	270	43	2 714

Com base nesses resultados, chegou-se às seguintes propostas para os estratos finais:

Proposta 1	Proposta 2
[5,30)	[5,30)
[30,100)	[30,100)
[100,500)	[100,400)
≥ 500 (estrato certo)	≥ 400 (estrato certo)

A escolha de PO=30 para limite do primeiro estrato (e não PO=20) foi tomada para se manter uma analogia com a PIA e também porque as diferenças nos tamanhos de amostra resultaram muito pequenas.

5.3 - Cálculo do tamanho da amostra para as diversas propostas

O desenho amostral considerado será o de amostragem aleatória estratificada nos estratos naturais, com alocação de Neyman nos estratos finais. Os estratos naturais são definidos pelos cruzamentos de localização geográfica e atividade econômica e os estratos finais pelo porte da UL, medido através do seu pessoal ocupado. De acordo com as discussões anteriores, serão consideradas as duas propostas para as três variáveis definidoras do desenho amostral.

O tamanho da amostra para o estrato amostrado em cada estrato natural foi calculado de forma que o CV para o estimador do total de pessoal ocupado fosse de 10% ou 15%. Assim, o tamanho da amostra em cada estrato natural é dado por:

$$n = N_C + \frac{\left(\sum_{h=1}^{3} N_h S_h\right)^2}{c^2 Y_A^2 + \sum_{h=1}^{3} N_h S_h^2} \qquad (1)$$

onde

n é o tamanho da amostra no estrato natural em questão;

N_C é o tamanho da população no estrato certo do estrato natural em questão;

N_h é o tamanho da população no estrato amostrado final h, h=1,2,3;

S_h^2 é a variância populacional do PO no estrato amostrado final h, h=1,2,3;

Y_A é o total populacional do PO no estrato amostrado do estrato natural em questão; e

c é o CV pré-fixado para o estimador do PO total na parte amostrada do estrato natural em questão.

Definido o tamanho total da amostra para cada estrato natural, o tamanho da amostra n_h em cada estrato amostrado final é dado por:

$$n_h = n \times \frac{N_h S_h}{\sum_h N_h S_h} \qquad (2)$$

segundo a alocação ótima de Neyman.

Os tamanhos de amostra calculados por (1) ou por (2), quando fracionários, foram sempre arredondados para o inteiro imediatamente maior. Além disso, para evitar problemas operacionais com amostras muito pequenas, foi arbitrado um valor mínimo de 5 ULs para o tamanho da amostra em cada estrato final. Então, quando o valor encontrado para n_h pela expressão (2) foi menor que 5, tomou-se o tamanho da amostra igual a 5, caso $N_h > 5$. Sempre que $N_h \leq 5$, tomou-se o tamanho da amostra no estrato igual ao tamanho da população, o que eqüivale a incluir na amostra, com certeza, todas as ULs do estrato final em questão. Nas tabelas que se seguem, o valor obtido por (2), arredondado, será denotado Am. Exata ou nl, enquanto o tamanho final, estipulado obedecendo à restrição de tamanho mínimo de 5 ULs, será denotado Am.Final ou n.

Vamos considerar inicialmente o estrato certo definido pelo corte de 500 empregados. Na **Tabela 5** temos os resultados referentes às diferentes propostas consideradas.

Tabela 5 - Tamanho da amostra - Amostragem aleatória estratificada com alocação de Neyman

Estrato	Atividade econômica: 27 divisões				Atividade econômica: 16 agrupamentos de divisões			
	Região geográfica				Região geográfica			
	Proposta inicial		Proposta alternativa		Proposta inicial		Proposta alternativa	
	CV=10%	CV=15%	CV=10%	CV=15%	CV=10%	CV=15%	CV=10%	CV=15%
Amostrado								
Am. exata	2 972	1 626	4 487	2 546	2 184	1137	3 389	1 832
Am. final	3 320	2 522	5 104	4 014	2 324	1 671	3 697	2 803
Certo	1199	1199	1199	1199	1199	1199	1199	1199
Total								
Am. exata	4 171	2 825	5 686	3 745	3 383	2 336	4 588	3 031
Am. final	4 519	3 721	6 303	5 213	3 523	2 870	4 896	4 002

Analisando os resultados, decidiu-se fixar o coeficiente de variação teórico em 10%, principalmente pelo fato de que também serão estimados totais para outras variáveis, além da variável de análise PO.

O tamanho da amostra obtido para o detalhamento máximo da atividade econômica e das regiões geográficas, 6 303, foi considerado grande, dadas as restrições orçamentárias e de pessoal. Como o agrupamento de divisões foi acordado pelos usuários internos da pesquisa, decidiu-se mantê-lo, junto com o detalhamento maior para a localização geográfica.

Trabalhando com o estrato certo definido pelo corte de 400 empregados, houve um aumento no tamanho total da amostra em virtude do aumento do tamanho do estrato certo, que passa de 1199 ULs para 1690 ULs. Sendo assim, essa alternativa também foi descartada.

Uma nova análise dos agrupamentos da CNAE foi feita e, devido ao fato de o agrupamento "Fabricação de Outros Produtos da Indústria de Transformação" ter a segunda maior participação no PO total (veja **Tabela 6**), decidiu-se isolar a divisão "Fabricação de Produtos de Madeira" (CNAE 20). Além disso, decidiu-se separar também a divisão "Fabricação de Produtos do Fumo" (CNAE 16) em virtude da forte sazonalidade da produção e também do espalhamento geográfico das ULs. No entanto, a amostra para essa divisão resultou ser bastante grande (130) quando comparada com o tamanho da população (176). Dados esse resultado e a já citada peculiaridade dessa atividade, decidiu-se incluir na amostra todas as ULs de tal divisão, ou seja, a divisão Fumo passa a ser considerada como um estrato certo.

Tabela 6 - Participação dos agrupamentos de CNAE no PO total

Agrupamento de CNAE	Pessoal Ocupado	
	Total	%
15	933 052	18,41
20+36+37	476 756	9,41
18	369 932	7,30
34+35	334 468	6,60
21+22	330 504	6,52
29+30	317 978	6,28
28	298 735	5,90
26	284 888	5,62
24	279 678	5,52
17	255 221	5,04
31+32+33	253 634	5,01
19	252 094	4,98
25	244 274	4,82
27	190 317	3,76
10+11+13+14	118 676	2,34
23	105 987	2,09
16	20 902	0,41
Total	5 067 096	100,00

6. A amostra final da PIMES

Diante dos resultados obtidos, o desenho da amostra da PIMES contemplará os seguintes aspectos:

- Unidade de investigação: Unidade Local;
- Variável de Análise: Pessoal Ocupado;
- Tipo de amostragem: Amostragem aleatória estratificada com alocação de Neyman;
- Coeficiente de variação: 10%; e
- Estratos naturais: 216 cruzamentos de localização geográfica e atividade econômica.
  - Localização geográfica: Regiões Norte e Centro-Oeste; Região Nordeste exclusive CE, PE, BA; Ceará; Pernambuco; Bahia; Minas Gerais; Espírito Santo; Rio de Janeiro; São Paulo; Paraná; Santa Catarina; Rio Grande do Sul.
  - Atividade econômica: Indústrias Extrativas; Fabricação de alimentos e bebidas; Fabricação de produtos do fumo; Fabricação de produtos têxteis; Confecção de artigos do vestuário e acessórios; Indústria do

calçado, inclusive preparação de artigos de couro; Fabricação de produtos de madeira; Indústria do papel e gráfica; Coque, refino de petróleo, combustíveis nucleares e álcool; Fabricação de produtos químicos; Fabricação de produtos de borracha e plástico; Fabricação de produtos de minerais não-metálicos; Metalurgia básica; Fabricação de produtos de metal, exclusive máquinas e equipamentos; Fabricação de máquinas e equipamentos, exclusive elétricos, eletrônicos, de precisão e de comunicação; Fabricação de máquinas e aparelhos elétricos, eletrônicos, de precisão e de comunicação; Fabricação de meios de transporte; Fabricação de outros produtos da indústria de transformação.

- Estratos finais: 4 classes de PO.
  - [5, 30); [30,100); [100,500); [500,∞).

A análise dos resultados da PIA mostra uma taxa de morte de empresas na ordem de 20%, o que revela uma desatualização do cadastro. De modo a evitar que a representatividade da amostra da PIMES fique prejudicada em função de tal desatualização, decidiu-se também aumentar em 20% o tamanho da amostra em cada estrato natural. Na **Tabela 7,** temos a distribuição da amostra final da PIMES por região geográfica. Tal amostra, de 5 862 ULs, está distribuída em 787 estratos finais, dos quais 527 são amostrados, 151 são estratos certos e 109 são estratos definidos gerencialmente como certos (fumo ou tamanho mínimo da amostra igual a 5).

Tabela 7 -Tamanho da população e da amostra final da PIMES por região geográfica

Região	Popu-lação	Amostra – CV=10%				
		Est. amostrado		Est. certo	Total	
		Am. exata	Am. final	(>= 500)	Am. exata	Am. final
Norte e Centro-Oeste	10 462	390	412	61	451	473
Nordeste, excl. CE,PE,BA	4 607	359	370	70	429	440
Ceará	2 974	286	309	41	327	350
Pernambuco	3 290	301	326	48	349	374
Bahia	3 368	324	344	18	342	362
Minas Gerais	16 847	438	445	99	537	544
Espírito Santo	2 914	261	281	12	273	293
Rio de Janeiro	10 803	414	421	79	493	500
São Paulo	50 135	463	472	496	959	968
Paraná	11 157	419	429	65	484	494
Santa Catarina	9 388	383	389	88	471	477
Rio Grande do Sul	13 539	461	465	122	583	587
TOTAL	139 484	4 499	4 663	1 199	5 698	5 862

7. Seleção e rotação da amostra

A exemplo do que já é feito nas Pesquisas Anuais da Indústria e do Comércio, o cadastro de seleção da Pesquisa Industrial Mensal de Emprego e Salário será extraído, a cada ano, do cadastro central com as informações mais atualizadas. Uma seleção independente da amostra a cada ano, no entanto, poderia provocar variações muito bruscas nos índices. Sendo assim, será adotado um mecanismo de seleção e rotação da amostra que permitirá que as ULs dos estratos amostrados permaneçam por um número máximo esperado de rodadas da pesquisa. Tal mecanismo é totalmente análogo ao empregado na PIA, mas utiliza os Números Aleatórios Permanentes (NAP) [veja Ohlsson (1995)] das ULs, uma vez que a unidade de investigação da PIMES é a UL.

Definido o tempo esperado de permanência na amostra das ULs dos estratos amostrados (na PIA, esse tempo é de 4 anos para as pequenas empresas), o mesmo algoritmo usado na PIA [veja Silva et al. (1998)] definirá a nova amostra da PIMES, assegurando a coordenação negativa das amostras em anos adjacentes. A diferença fundamental será o fato de que, em vez de ser feita em um único instante de tempo, a rotação será distribuída ao longo do ano, de modo a evitar variações bruscas nos índices devidas simplesmente à substituição de parte da amostra.

No que segue, estaremos descrevendo o procedimento a ser aplicado em cada um dos estratos amostrados finais da amostra da PIMES que são definidos pelos cruzamentos de região geográfica, agrupamentos da CNAE a 2 dígitos e as 3 classes de PO dadas por (5,30), (30,100) e (100,500).

7.1 - Algoritmo de seleção e rotação da amostra

A seguir descrevem-se as etapas do procedimento de seleção e rotação da amostra nos estratos amostrados:

- As ULs do cadastro são estratificadas de acordo com os critérios de estratificação definidos no plano amostral, sendo a estratificação refeita a cada ano, considerando as informações mais atualizadas disponíveis;
- Cada UL, ao entrar no cadastro do IBGE, recebe um número aleatório permanente (NAP) que não será modificado enquanto a UL permanecer ativa no cadastro. Esses NAPs já foram gerados quando da montagem do cadastro para seleção das amostras da PIA;
- Em cada estrato amostrado final, as ULs são ordenadas crescentemente segundo os NAPs a elas associados. Obtém-se, assim, a população ordenada de números aleatórios $A_{(1)}, A_{(2)}, \ldots, A_{(N)}$, onde N aqui representa o tamanho da população do estrato final em questão. Partindo da população ordenada, determinam-se as posições das ULs, ou seja, os postos $P_1, P_2, \ldots, P_N$ das Uls, segundo os números aleatórios;
- Em cada rodada da pesquisa, calcula-se novamente o tamanho da amostra para todos os estratos, usando-se as informações mais atualizadas;
- As posições inicial e final das ULs a serem incluídas na amostra ao longo do ano são dadas por:

$$\text{Início} = \left\lfloor (r-1) \times \frac{n}{T} \right\rfloor \bmod N + 1 \tag{3}$$

$$\textit{Final} = \textit{Início} + n - 1 \tag{4}$$

onde

$\lfloor \cdot \rfloor$ representa o menor inteiro maior ou igual a $\cdot$;

r é a rodada da pesquisa;

n é o tamanho da amostra no estrato final em questão;

T é o limite máximo de rodadas que se espera que as ULs permaneçam na amostra;

N é o tamanho da população no estrato final em questão; e

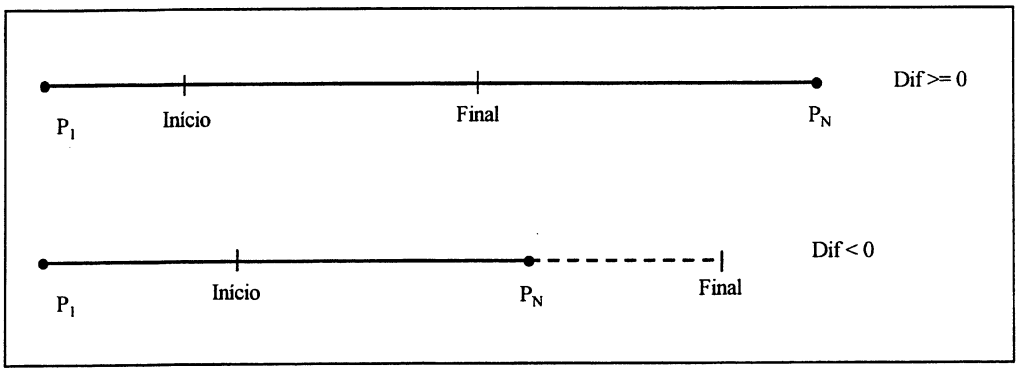
mod é a função módulo, que retorna o resto da divisão.

- Caso o valor de Final seja maior que o tamanho N da população, o procedimento consiste em completar a amostra com o número necessário de elementos do início do estrato (menores NAPs). Então, calcula-se a diferença

$$Dif = Final - N = Início + n - 1 - N = n + (Início - N - 1) \tag{5}$$

que controla a disponibilidade de ULs para a rotação (veja **Figura 2**).

Figura 2 - Ilustração do processo de rotação – Parte 1



- A regra final de inclusão das ULs na amostra é a seguinte (veja **Figura 3**):

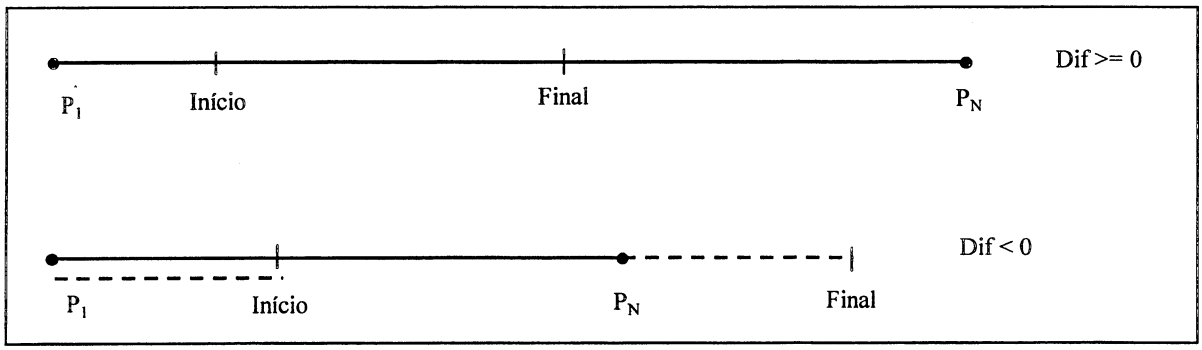
(i) se $Dif \leq 0$ incluem-se as ULs para as quais

$$Início \leq P_i \leq Final \tag{6}$$

(ii) se $Dif > 0$ (isto é, Final > N) incluem-se as ULs para as quais

$$\begin{cases} Início \leq P_i \leq Final \\ \quad ou \\ P_i \leq Dif \end{cases} \tag{7}$$

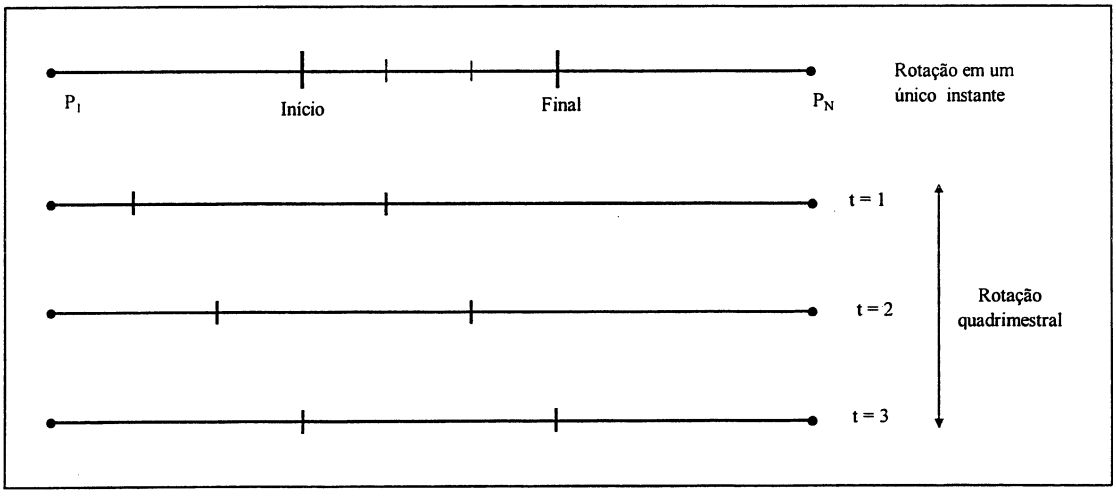
Figura 3 - Ilustração do processo de rotação – Parte 2



Até aqui, o processo é idêntico ao da PIA. Então, no instante t_1^r onde é iniciada a r-ésima rodada da pesquisa, tem-se, com o procedimento acima, todas as ULs que serão incluídas na amostra ao longo do ano. Na primeira rodada da pesquisa (r = 1), são incluídas as ULs com os n primeiros postos. A partir da primeira rodada, espera-se, em média, uma rotação de $\left(\frac{1}{T}\times 100\right)\%$ da amostra, isto é, $\left(\frac{1}{T}\times 100\right)\%$ da amostra é substituída ao longo do ano.

A diferença na PIMES é que, a partir da segunda rodada, a substituição de parte da amostra não será feita em um único instante, mas, sim, distribuída por trimestre ou por quadrimestre (veja **Figura 4**).

Figura 4 - Ilustração do processo de rotação – parte 3



Vamos formalizar, agora, a regra de inclusão das ULs, supondo que a substituição será feita em cada trimestre. Consideremos, então, a r-ésima rodada da pesquisa, $r \geq 2$. A substituição de parte da amostra será feita em 4 instantes:

t=1 instante da rotação;

t=2 primeiro mês do segundo trimestre;

t=3 primeiro mês do terceiro trimestre; e

t=4 primeiro mês do quarto trimestre.

Para definir as ULs que entrarão no trimestre t da r-ésima rodada, definem-se:

$$Início_t^r = \left\lfloor \left(r - 2 + \frac{t}{4} \right) \times \frac{n}{T} \right\rfloor \bmod N + 1 \qquad r = 2, 3, \ldots; \quad t = 1,2,3,4 \tag{8}$$

$$Final_t^r = Início_t^r + n - 1 \tag{9}$$

$$Dif_t^r = Final_t^r - N = n - (N - Início_t^r + 1) \tag{10}$$

e selecionam-se as ULs com postos P_i tais que

(i) se $Dif_t^r \leq 0$, $\qquad Início_t^r \leq P_i \leq Final_t^r$ (11)

(ii) se $Dif_t^r > 0$ $\qquad \begin{cases} Início_t^r \leq P_i \leq N \\ \quad ou \\ P_i \leq Dif_t^r \end{cases}$ (12)

Caso se opte por fazer a rotação por quadrimestre, o procedimento é análogo: a substituição será feita em 3 instantes de tempo (t=1,2,3) e nas fórmulas, em vez de $\frac{t}{4}$, teremos $\frac{t}{3}$.

7.2 - Exemplo

Consideremos um estrato final onde o tamanho da população é N = 22 e de onde se pretende tirar uma amostra de tamanho n = 11. Suponhamos que se pretenda manter cada UL por no máximo 3 anos, isto é, T = 3. Suponhamos, também, que a população permaneça fixa, de modo que a cada rodada temos o mesmo tamanho de população e amostra. É claro que esse exemplo é uma simplificação da realidade, uma vez que, na prática, teremos, a cada ano, tamanhos de população e amostra diferentes.

1ª rodada

No início da pesquisa (r = 1), são selecionadas as ULs com os 11 primeiros postos.

2ª rodada (r = 2)

t=1 $$Início_1^2 = \left\lfloor \left(2-2+\frac{1}{4}\right)\times\frac{11}{3}\right\rfloor \bmod 22+1 = \lfloor 0{,}92 \rfloor \bmod 22+1 = 1 \bmod 22+1 = 1+1 = 2$$

$$Final_1^2 = 2+11-1 = 12 \quad \Rightarrow \quad 2 \le P_i \le 12$$

t=2 $$Início_2^2 = \left\lfloor \left(2-2+\frac{2}{4}\right)\times\frac{11}{3}\right\rfloor \bmod 22+1 = \lfloor 1{,}83 \rfloor \bmod 22+1 = 2 \bmod 22+1 = 3$$

$$Final_2^2 = 3+11-1 = 13 \quad \Rightarrow \quad 3 \le P_i \le 13$$

t=3 $$Início_3^2 = \left\lfloor \left(2-2+\frac{3}{4}\right)\times\frac{11}{3}\right\rfloor \bmod 22+1 = \lfloor 2{,}75 \rfloor \bmod 22+1 = 3 \bmod 22+1 = 3+1 = 4$$

$$Final_3^2 = 4+11-1 = 14 \quad \Rightarrow \quad 4 \le P_i \le 14$$

t=4 $$Início_4^2 = \left\lfloor \left(2-2+\frac{4}{4}\right)\times\frac{11}{3}\right\rfloor \bmod 22+1 = \lfloor 3{,}67 \rfloor \bmod 22+1 = 4 \bmod 22+1 = 4+1 = 5$$

$$Final_4^2 = 5+11-1 = 15 \quad \Rightarrow \quad 5 \le P_i \le 15$$

3ª rodada (r = 3)

t=1 $6 \le P_i \le 16$

t=2 $7 \le P_i \le 17$

t=3 $8 \le P_i \le 18$

t=4 $9 \le P_i \le 19$

e assim sucessivamente. Na **Figura 5** abaixo ilustra-se o processo de rotação para o exemplo.

Figura 5 -Ilustração do processo de rotação para o exemplo

r=1		1	2	3	4	5	6	7	8	9	10	11										
r=2	t=1		2	3	4	5	6	7	8	9	10	11	12									
	t=2			3	4	5	6	7	8	9	10	11	12	13								
	t=3				4	5	6	7	8	9	10	11	12	13	14							
	t=4					5	6	7	8	9	10	11	12	13	14	15						
r=3	t=1						6	7	8	9	10	11	12	13	14	15	16					
	t=2							7	8	9	10	11	12	13	14	15	16	17				
	t=3								8	9	10	11	12	13	14	15	16	17	18			
	t=4									9	10	11	12	13	14	15	16	17	18	19		
r=4	t=1										10	11	12	13	14	15	16	17	18	19	20	
	t=2											11	12	13	14	15	16	17	18	19	20	21
	t=3												12	13	14	15	16	17	18	19	20	21 22
	t=4												12	13	14	15	16	17	18	19	20	21 22
r=5	t=1	1												13	14	15	16	17	18	19	20	21 22
	t=2	1	2												14	15	16	17	18	19	20	21 22
	t=3	1	2	3												15	16	17	18	19	20	21 22
	t=4	1	2	3	4												16	17	18	19	20	21 22
r=6	t=1	1	2	3	4	5												17	18	19	20	21 22
	t=2	1	2	3	4	5	6												18	19	20	21 22
	t=3	1	2	3	4	5	6	7												19	20	21 22
	t=4	1	2	3	4	5	6	7	8												20	21 22

7.3 - Algumas observações sobre a operacionalização do processo

1) Para alguns estratos finais, a diferença entre o tamanho da população e o tamanho da amostra é pequena. Com esse processo de rotação, é possível que, nesses estratos, algumas ULs tenham que sair e entrar na amostra em um momento posterior no mesmo ano. Operacionalmente, isso pode ser bastante complicado. Decidiu-se, então, que os estratos onde a diferença entre os tamanhos da população e amostra fosse menor que 5 seriam definidos como estratos certos.
2) No instante da rotação, isto é, da atualização cadastral, usando as fórmulas (6) a (10), são identificadas as ULs que entrarão na amostra e o trimestre do ingresso. Comparando com a amostra da rodada anterior, é possível identificar as ULs que deixarão a amostra e as ULs que entrarão na amostra. As primeiras devem ser notificadas que deixarão de participar da pesquisa por um certo tempo, mas poderão retornar ao painel de informantes em um momento futuro. As ULs novas têm que ser notificadas e treinadas para responder o questionário.
3) A questão da retirada e inclusão de ULs na amostra tem que ser tratada com bastante cuidado. Fazer a exclusão e a inclusão simultaneamente pressupõe que os novos informantes responderão o questionário a contento na primeira entrevista e essa pode ser uma suposição errônea. Então, os novos informantes serão solicitados a preencher o questionário um mês antes da real inclusão dos seus dados nos cálculos dos

estimadores. Esse procedimento, embora aumente um pouco a carga de trabalho de coleta, certamente reduzirá o problema de não-resposta.

7.4 - Simulação

Para avaliar o impacto do processo de rotação, foi feito um estudo utilizando cadastros de dois anos consecutivos: 1997 e 1998. Com base no cadastro de 1997, selecionou-se o que seria a primeira amostra da pesquisa e no cadastro de 1998 aplicou-se o processo de rotação para gerar a amostra da segunda rodada. Nesse estudo, considerou-se também como estrato certo gerencial aquele onde a diferença entre os tamanhos da população e da amostra era menor que 5 (veja observação 1). Essa decisão acarretou um aumento de 116 ULs na amostra de 1997 e de 107 ULs na amostra de 1998.

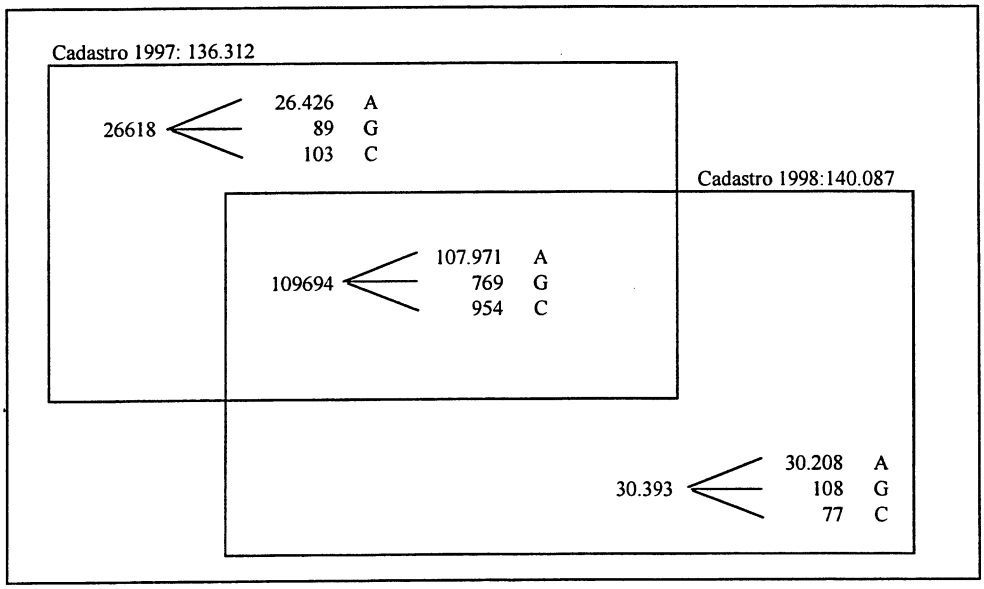
A composição desses dois cadastros está ilustrada na **Figura 6** a seguir, onde utilizou-se a seguinte notação:

A estrato amostrado

G estrato certo gerencial

C estrato certo

Figura 6 - Composição dos cadastros de 1997 e 1998



Pode-se observar que na parte comum dos dois cadastros houve um remanejamento das ULs dentro dos estratos finais, detalhada na **Tabela 8**. Das 107 971 ULs dos estratos amostrados do cadastro de 1998, 107 546 (99,6%) pertenciam aos estratos amostrados no ano anterior; das 769 ULs dos estratos certos gerenciais, 604 (78,5%) também pertenciam a esse tipo de estrato em 1997, enquanto das 954 ULs do estrato certo, 851 (89,2%) também eram do estrato certo em 1997. Por outro lado, 26 618 ULs que estavam no cadastro de 1997 não aparecem no cadastro de 1998; é interessante notar que essas mortes correspondem aproximadamente a 20% do total de ULs, o que é um indicativo que a folga de 20% dada no tamanho da amostra é razoável.

Tabela 8 - Remanejamento das ULs comuns aos 2 cadastros

		Cadastro 1997			Total
		Amostrado	Gerencial	Certo	
Cadastro 1998	Amostrado	107 546	210	215	107 971
	Gerencial	138	604	27	769
	Certo	93	10	851	954

O tamanho da amostra inicial da pesquisa, baseada no cadastro de 1997, é de 5 938 ULs, das quais 1 196 pertencem ao estrato certo, 913 aos estratos certos gerenciais e 3 829 aos estratos amostrados. O tamanho da amostra para a segunda rodada é de 5 870 ULs, das quais 1.031 pertencem ao estrato certo, 877 aos estratos certos gerenciais e 3 962 aos estratos amostrados (veja **Tabela 9**).

Tabela 9 - Distribuição das amostras e do número de estratos, segundo o tipo de estrato

	Tipo de estrato			Total
	Amostrado	Gerencial	Certo	
1ª amostra (1997)	3 829 (472)	913 (165)	1 196 (151)	5 938
2ª amostra (1998)	3 962 (479)	877 (158)	1 031 (153)	5 870

O mecanismo de rotação aplica-se à parte dos 479 estratos amostrados e vamos considerar a rotação distribuída por trimestre (4 instantes de tempo no ano). Assim, a cada trimestre teremos uma "nova" amostra de 3 962 ULs, mais as 1 908 (877 + 1 031) ULs dos estratos certos. O tempo esperado de permanência das ULs na amostra considerado nesse exemplo foi de 4 anos (T=4).

Com relação às ULs que entrarão na amostra ao longo de todo o ano (1º, 2º, 3º ou 4º trimestre), temos a seguinte situação, ilustrada na **Figura 7**: 2 379 (42 + 53 + 67 + 1 277 + 136 + 226 + 234 +192 + 152) ULs são novas e 4 225 (141 + 219 + 212 + 1 951 + 47 + 46 + 45 + 685 + 879) já pertenciam à amostra anterior. Da amostra anterior, por sua vez, 1 713 (28,9%) deixam de participar do painel.

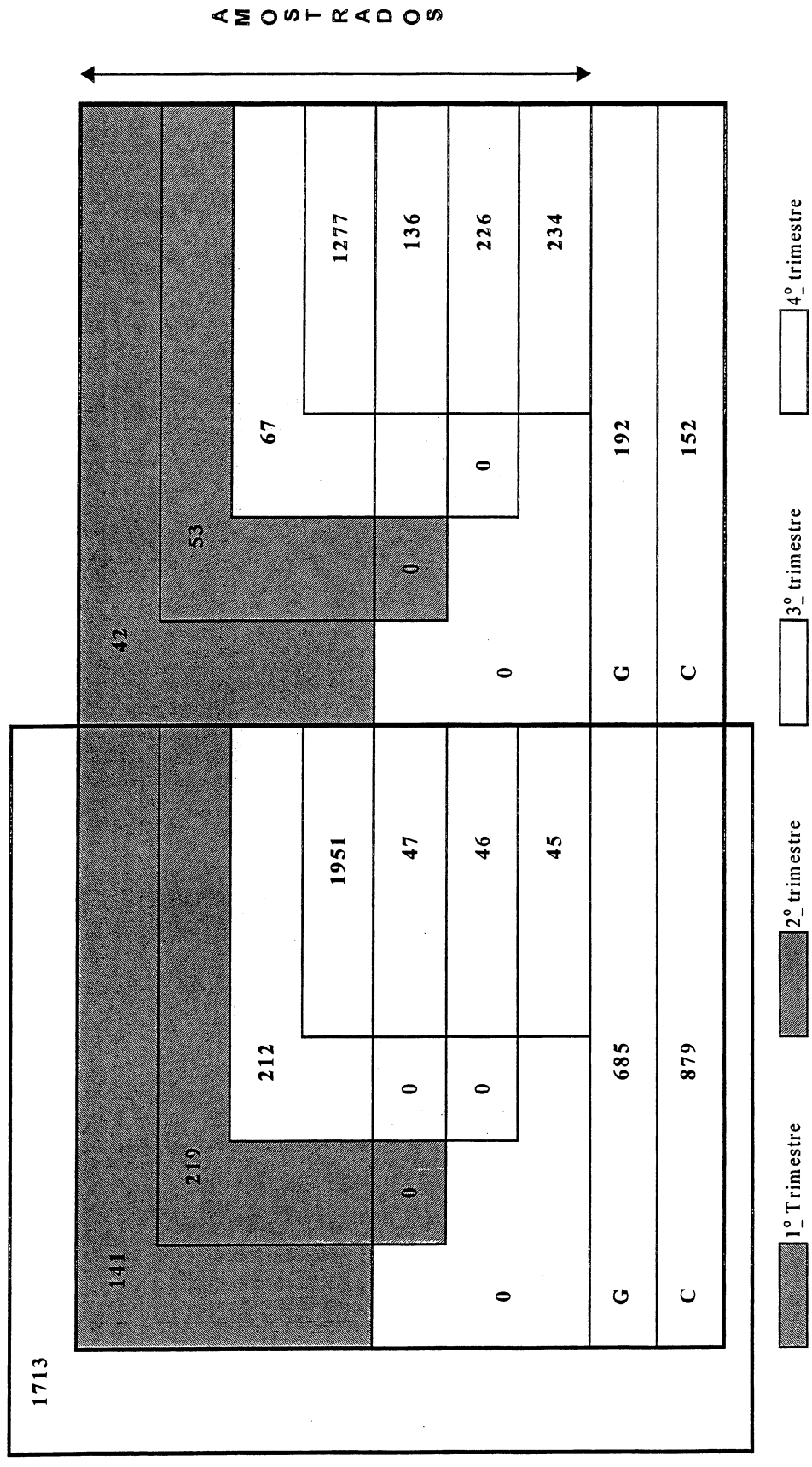
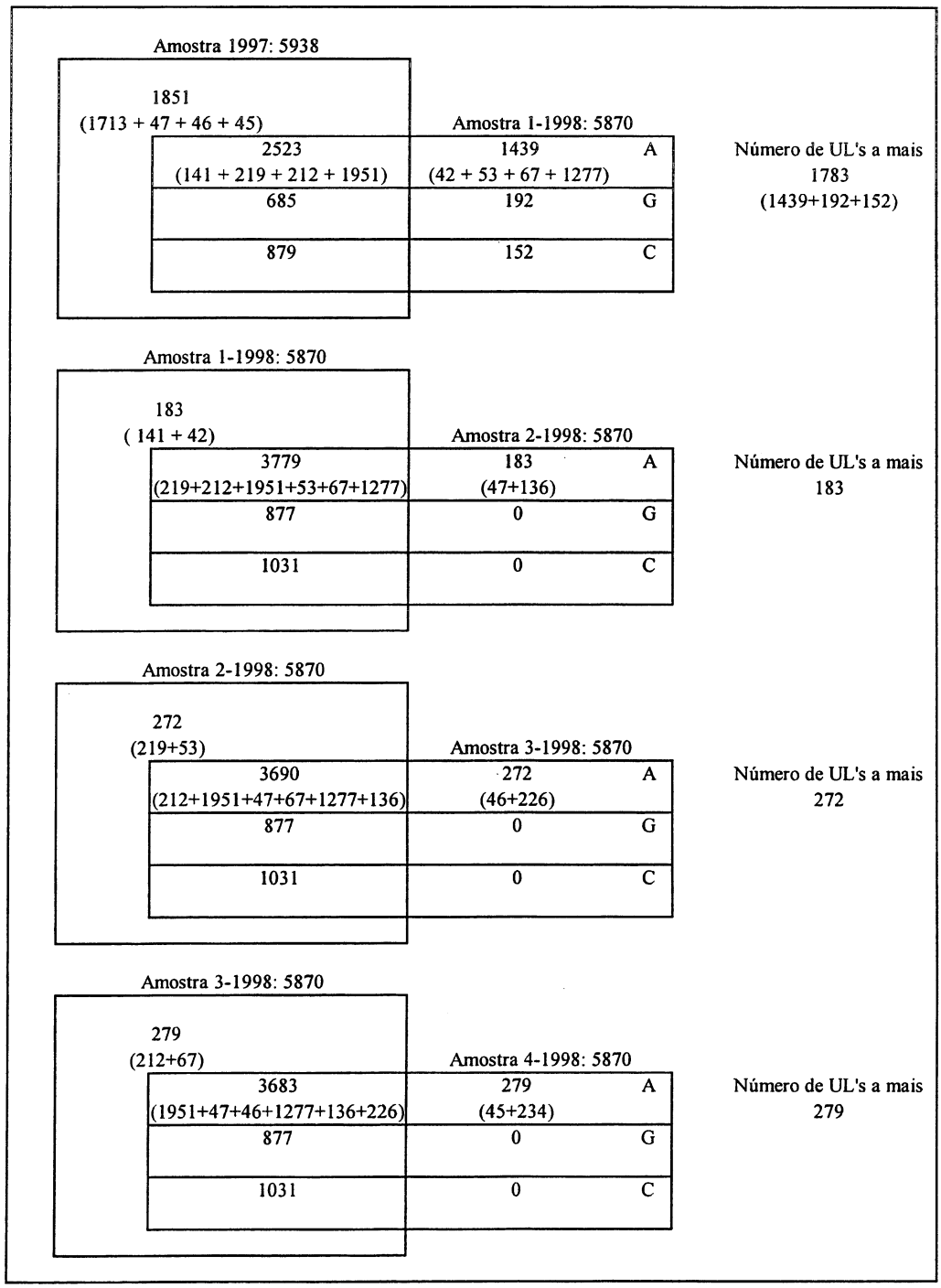


Figura 7 - ULs das amostras – rotação trimestral – T=4

Vamos analisar, agora, as amostras de períodos consecutivos, para mensurar o impacto de superposição de entrevistas que ocorre devido ao fato de os novos informantes responderem o questionário um mês antes da sua inclusão (veja observação 3). Na **Figura 8** temos os resultados dessas análises. O impacto forte, com maior taxa de sobreposição de amostras, ocorre no momento da troca de cadastros, ou seja, no primeiro trimestre, quando deverão ser visitadas 1 783 ULs a mais. Se a rotação fosse feita em um único instante, esse número passaria para 2 217.

Figura 8 - ULs das amostras para rotação efetuada trimestralmente – T=4



Análises alternativas foram feitas, considerando a rotação distribuída por quadrimestre e o tempo esperado de permanência na amostra de 3 anos. Analisados os resultados, essas novas alternativas foram descartadas e manteve-se o esquema de rotação caracterizado pelos seguintes parâmetros: tempo esperado de permanência na amostra T=4 anos e rotação distribuída ao longo dos 4 trimestres. A escolha do primeiro parâmetro baseou-se principalmente no fato de que, a cada ano, já há uma rotação natural em função da mudança do cadastro. Já a escolha do segundo parâmetro baseou-se na conveniência de se ter uma maior uniformidade na distribuição da carga de trabalho da equipe de coleta.

Para auxiliar a compreensão do processo de rotação, vamos considerar dois estratos finais específicos, correspondentes à Região Nordeste exclusive Ceará, Pernambuco e Bahia, classe de PO de 5 a 30 e atividades Metalurgia Básica e Produtos de Metal, exclusive máquinas e equipamentos, respectivamente. Nas **Figuras 9** e **10,** temos o esquema da população nos dois cadastros e os tamanhos das amostras, assim como o posto das ULs que devem entrar na amostra em cada trimestre de 1998.

Figura 9 - Nordeste exclusive CE, PE, BA – Metalurgia Básica - (5,30)

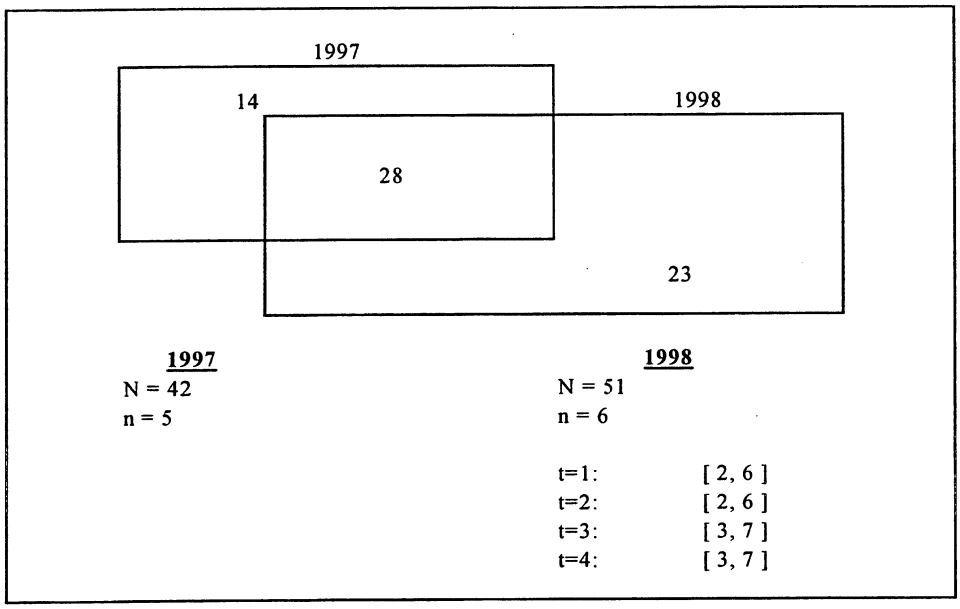
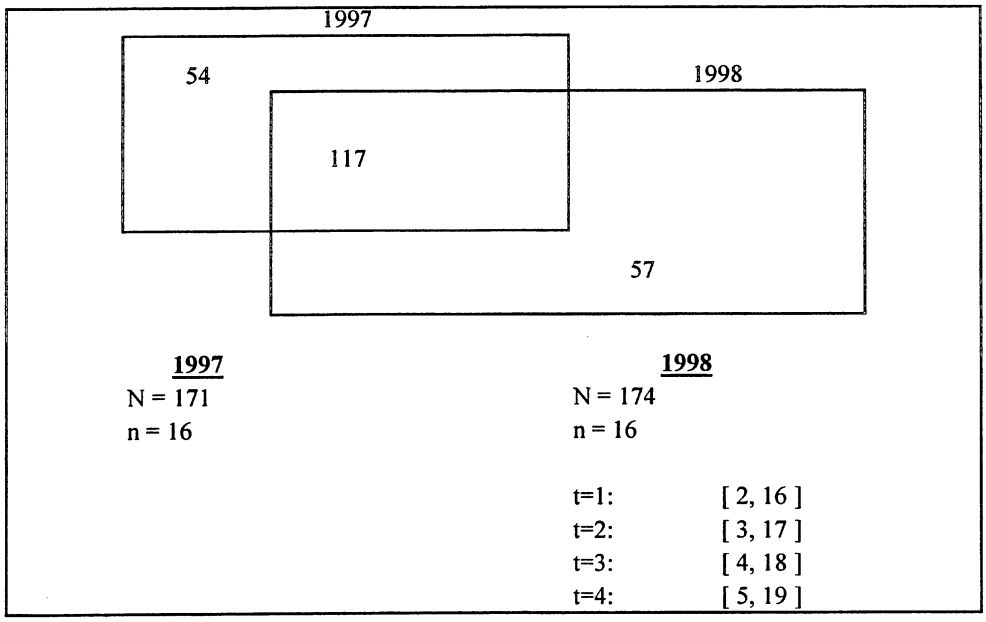


Figura 10 - Nordeste exclusive Ceará, Pernambuco e Bahia – Produto Metal exclusive máquinas e equipamentos – (5,30)



Nas **Tabelas 10** e **11** temos a relação das ULs dos dois cadastros e das amostras. Da primeira pode-se ver que uma empresa nova, correspondente à observação 49, não entra na amostra, devido ao fato de o seu NAP ser menor que o da primeira observação a ser incluída na amostra. Da segunda tabela, constata-se que uma empresa nova, correspondente à observação 107, entra na amostra por apenas um trimestre. Embora seja uma situação operacionalmente inconveniente, não há maneira de evitar esse tipo de problema, dada a aleatoriedade dos NAPs. No entanto, a tendência é que haja poucas ocorrências desse tipo.

Tabela 10 - Nordeste exclusive Ceará, Pernambuco e Bahia – metalurgia básica – (5,30)

Obs	Empresa	UL	Aleat
60	69577666	1	0,01783
58	63410427	1	0,03357
61	70041082	1	0,06845
50	35141381	1	0,07481
38	10937886	1	0,07769
7	1005101	1	0,11205
22	5804836	1	0,15103
26	8223687	2	0,17272
15	1950539	1	0,18073
47	23600430	1	0,21119
30	8490716	1	0,24135
44	12924452	1	0,24661
52	35581529	1	0,26291
13	1759484	1	0,27749
9	1324669	1	0,31278
8	1117740	1	0,34608
54	40982076	1	0,35245
39	11889292	1	0,35841
31	9093386	1	0,40456
24	7159809	1	0,42605
42	12574927	1	0,42890
45	13414156	1	0,44915
65	97464689	1	0,45336
34	10288892	1	0,45392
64	97432579	1	0,45412
43	12636429	1	0,45448
2	386643	1	0,49374
40	11895190	1	0,54029
37	10745255	1	0,55264
3	526052	1	0,60821
28	8424210	1	0,62726
6	993944	1	0,64417
62	70111521	1	0,65201
35	10306322	1	0,75828
32	9222928	1	0,77806
46	23513013	1	0,78159
29	8471476	1	0,84029
53	40924227	1	0,85682
33	9364977	1	0,86906
55	41002882	1	0,89812
51	35361351	1	0,92118
27	8341752	1	0,96743

Legenda

	Cadastro 97
	Cadastro 98
	Cadastros 97 e 98

Obs	Empresa	UL	Aleat	
49	24366361	1	0,02733	
5	887309	1	0,04920	
56	41145798	1	0,06201	
61	70041082	1	0,06845	
18	2166039	1	0,07480	t=1,2
50	35141381	1	0,07481	t=3,4
4	799182	1	0,09574	
22	5804836	1	0,15103	
26	8223687	2	0,17272	
47	23600430	1	0,21119	
13	1759484	1	0,27749	
23	5804836	2	0,28826	
9	1324669	1	0,31278	
39	11889292	1	0,35841	
17	2165005	1	0,39565	
63	70323944	1	0,40327	
31	9093386	1	0,40456	
42	12574927	1	0,42890	
45	13414156	1	0,44915	
65	97464689	1	0,45336	
34	10288892	1	0,45392	
64	97432579	1	0,45412	
14	1794492	1	0,46096	
16	2021305	1	0,49123	
36	10744209	1	0,50791	
21	2891966	1	0,52760	
59	69385227	1	0,52771	
48	24102360	1	0,53481	
19	2428701	1	0,53873	
40	11895190	1	0,54029	
37	10745255	1	0,55264	
25	7628886	1	0,58509	
3	526052	1	0,60821	
28	8424210	1	0,62726	
6	993944	1	0,64417	
62	70111521	1	0,65201	
32	9222928	1	0,77806	
46	23513013	1	0,78159	
10	1422866	1	0,78210	
1	207452	1	0,78967	
20	2638660	1	0,83626	
29	8471476	1	0,84029	
12	1748000	1	0,84904	
53	40924227	1	0,85682	
33	9364977	1	0,86906	
55	41002882	1	0,89812	
51	35361351	1	0,92118	
41	12171716	1	0,92668	
27	8341752	1	0,96743	
57	41199886	1	0,96875	
11	1610480	1	0,98626	

Tabela 11

Nordeste exclusive Ceará, Pernambuco, Bahia – produtos de metal, exclusive máquina e equipamentos. – (5,30)

Obs	Empresa	UL	Aleat
116	10952638	1	0,00587
199	41188871	1	0,01095
83	8518433	1	0,01503
226	74030354	1	0,01538
112	10751253	1	0,01559
201	41499286	1	0,01613
152	23502404	1	0,01717
178	32892648	1	0,01829
147	15595242	1	0,02025
69	5806070	1	0,02262
127	12513248	1	0,02297
221	70102025	1	0,02347
183	35370584	1	0,03240
159	24218505	1	0,03789
39	1679962	1	0,05788
202	41518408	1	0,07205
9	485473	1	0,08192
173	32839318	1	0,08434
90	8822108	1	0,09403
228	86779741	1	0,10156
145	15121817	1	0,10317
129	12740064	1	0,10492
81	8467110	1	0,12362
191	40761652	1	0,13576
114	10835536	1	0,13624
141	13382619	1	0,14099
130	12827952	1	0,14256
97	9169046	1	0,14517
36	1517563	1	0,14589
48	1843210	1	0,15701
66	5489869	1	0,16533
103	9280116	1	0,16722
162	24306862	1	0,17010
153	23505373	1	0,17391
179	35149632	1	0,17788
64	5233705	1	0,18622
96	9135419	1	0,19473
217	70009808	1	0,21368
67	5633623	1	0,21419
192	40775280	1	0,21752
109	10332880	1	0,219
220	70092044	1	0,22128
77	8209942	1	0,22601
210	69424018	1	0,22782
...			

Obs	Empresa	UL	Aleat
116	10952638	1	0,00587
107	10330264	1	0,00713
199	41188871	1	0,01095
224	70132170	1	0,01183
31	1281002	1	0,01193
83	8518433	1	0,01503
112	10751253	1	0,01559
201	41499286	1	0,01613
152	23502404	1	0,01717
211	69577666	1	0,01783
54	1962146	1	0,01943
59	2231854	1	0,01999
147	15595242	1	0,02025
69	5806070	1	0,02262
127	12513248	1	0,02297
221	70102025	1	0,02347
183	35370584	1	0,03240
159	24218505	1	0,03789
39	1679962	1	0,05788
143	15056021	1	0,05978
62	2703480	1	0,06341
9	485473	1	0,08192
90	8822108	1	0,09403
228	86779741	1	0,10156
145	15121817	1	0,10317
55	2105317	1	0,10438
129	12740064	1	0,10492
50	1925398	1	0,10647
63	2800037	1	0,11807
195	41005208	1	0,12609
108	10330777	1	0,13247
114	10835536	1	0,13624
141	13382619	1	0,14099
36	1517563	1	0,14589
142	13946272	1	0,14729
35	1503052	1	0,15437
48	1843210	1	0,15701
66	5489869	1	0,16533
103	9280116	1	0,16722
45	1755670	1	0,1681
17	864988	1	0,16821
162	24306862	1	0,1701
153	23505373	1	0,17391
179	35149632	1	0,17788
...			

t=1 t=2 t=3 t=4

Legenda

	Cadastro 97
	Cadastro 98
	Cadastros 97 e 98

8. Conclusão

Nesse texto apresentou-se o estudo realizado para definir o desenho amostral da Pesquisa Industrial Mensal de Emprego e Salário a ser realizada pelo Departamento de Indústria do IBGE. O desenho amostral acordado é o de amostragem aleatória estratificada com alocação de Neyman nos estratos finais. Os estratos naturais são definidos pelos cruzamentos de Região Geográfica e Atividade Econômica e os estratos finais são definidos pelo número de pessoas ocupadas. Todas as ULs com 500 ou mais pessoas ocupadas entram na amostra com probabilidade um (estrato certo) e na parte amostrada de cada estrato natural toma-se uma amostra de modo a garantir um coeficiente de variação de 10% para o estimador do total da variável de análise, que é o Pessoal Ocupado (PO).

As regiões geográficas consideradas são: Norte e Centro-Oeste, Nordeste exclusive Ceará, Pernambuco e Bahia, Ceará, Pernambuco, Bahia, Minas Gerais, Espírito Santo, Rio de Janeiro, São Paulo, Paraná, Santa Catarina, Rio Grande do Sul.

As atividades econômicas são definidas pelos seguintes agrupamentos da CNAE a 2 dígitos: Indústria Extrativa (10+11+13+14), Alimentos e Bebidas (15), Fumo (16), Têxteis (17), Vestuário (18), Couros (19), Madeira (20), Papel e Gráfica (21+22), Coque e Petróleo (23), Produtos Químicos (24), Borracha e plástico (25), Minerais não-metálicos (26), Metalurgia básica (27), Produtos de metal, exclusive máquinas e equipamentos (28), Máquinas e equipamentos, exclusive eletrônicos, elétricos, de precisão e de comunicação (29+30), Equipamentos elétricos, eletrônicos, de precisão e de comunicação (31+32+33), Meios de transporte (34+35), e Outros produtos da indústria de transformação (36+37).

O tamanho mínimo da amostra em cada estrato final foi definido como 5 e, assim, os estratos finais com população menor que 5 foram definidos como estratos certos gerenciais. Todas as ULs da divisão Fumo também foram incluídas com certeza na amostra, definindo novos estratos certos gerenciais. Para contornar problemas devidos à desatualização cadastral, o tamanho da amostra em cada estrato natural amostrado foi aumentado em 20%. Finalmente, definiu-se como estrato certo gerencial todo aquele em que a diferença entre os tamanhos da população e da amostra fosse menor que 5.

A seleção e rotação da amostra será feita com base nos Números Aleatórios Permanentes. A cada ano a amostra será rotacionada, de modo que cada UL permaneça, em média, 4 anos no painel. A substituição das ULs, devida ao mecanismo de rotação, será feita trimestralmente, mantendo-se os velhos e novos informantes simultaneamente por 1 mês.

Referências bibliográficas

FARIAS, A.M.L. (1999) *Relatório das Atividades de Consultoria*, Relatório Interno DEIND/IBGE, dezembro.

HIDIROGLOU, M.A. (1986) The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

LAVALLÉE, P. E HIDIROGLOU, M.A. (1988) On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

OHLSSON, E. (1995) Coordination of Samples using Permanent Random Numbers. In Cox, Binder, Chinnappa, Christianson, Colledge e Kott (eds.) *Business Survey Methods*, New York, Wiley, p. 153-169.

SILVA, P.LN. et al. (1998) *Planejamento Amostral para as Pesquisas Anuais da Indústria e do Comércio.* Textos para Discussão nº 92, outubro de 1998, Rio de Janeiro: IBGE.

Agradecimentos

A autora agradece aos técnicos dos Departamentos de Indústria e de Metodologia do IBGE pelo suporte oferecido durante a realização deste trabalho, em especial a Wasmália Bivar, Fernanda Santis, Sílvio Salles, Pedro Luis do Nascimento Silva, Sonia Albieri e Antonio José Ribeiro Dias.

Este trabalho foi desenvolvido com suporte financeiro do Banco Mundial.

ABSTRACT

The paper describes studies carried out to design the sample of the Monthly Survey of Manufacturing. The size measure available for each unit was number of employees, which has a very skewed distribution. Then the stratification of the population was done defining "take-all" and "take-some" strata. Special emphasis is given to the selection and rotation of the sample, which were based on the Permanent Random Numbers tecnhique.

Key words: skewed distribution; take-all stratum; sample rotation.

Estimação de índices de proficiência escolar para pequenas áreas do Município do Rio de Janeiro via modelos logísticos hierárquicos[1]

Fernando Antônio da Silva Moura (UFRJ)*

Daniel Francisco Neyra Castañeda*

RESUMO

Pesquisas por amostragem são usualmente desenhadas para proverem estimativas a um nível relativamente desagregado (*ex.: estados*). Contudo há interesse em serem obtidas estimativas para domínios ou regiões mais desagregadas, onde os respectivos tamanhos da amostra são geralmente pequenos. Desta maneira, o uso de estimadores diretos de expansão tais como aqueles encontrados em Cochran (1977) podem resultar em estimativas imprecisas. Neste caso, abordagens baseadas em modelos são recomendadas. Neste trabalho, apresentam-se a abordagem Clássica e a Bayesiana, para serem estimadas proporções em pequenas áreas. Diferenças entre as pequenas áreas, são consideradas através de modelos hierárquicos de regressão logística. Metodologias Bayesianas e Clássicas são aplicados para dados educacionais do Município do Rio de Janeiro. Finalmente uma análise crítica sobre os dois tipos de abordagem é feita.

Palavras-Chave: Estimação em pequenas áreas, regressão logística hierárquica, MCMC.

1. Introdução

O termo "Pequena área" ou "Área local" é freqüentemente utilizado para denotar uma pequena área geográfica, tal como um município, um bairro, ou um setor censitário.

1 Este trabalho é parte da Tese de mestrado de Daniel Francisco Neyra Castañeda e foi financiado pelo CNPq.

* Endereço para correspondência: Instituto de Matemática – UFRJ – Caixa Postal : 68530 - e-mail: fmoura@dme.ufrj.br.

A produção de estatísticas para as pequenas áreas foi introduzida há vários séculos. Brackstone (1987), menciona a existência de tais estatísticas no Século XI na Inglaterra e no Século XVII no Canadá. Nas últimas quatro décadas, com o desenvolvimento da Teoria da Amostragem, os métodos de enumeração completa da população foram substituídos por pesquisas por amostragem. As pesquisas por amostragem têm sido aplicadas com sucesso para a produção de estimadores com aceitável precisão para totais e médias de áreas em domínios grandes. Contudo a utilização de estimadores simples de expansão para as pequenas áreas, provavelmente produziria resultados com precisão inaceitável, devido ao pequeno tamanho da amostra.

Uma técnica freqüentemente empregada na estimação de pequenas áreas é a combinação de dados amostrais com uma fonte auxiliar de informação, tais como registros ou censos administrativos. Nesta linha, têm-se desenvolvido estimadores baseados em modelos que tomam emprestada a informação de áreas vizinhas com a finalidade de obter estimadores mais precisos. Royall (1970) usa uma abordagem baseada em modelos de superpopulação, e considera o problema de estimar totais em populações finitas, quando uma informação auxiliar está disponível para todos os elementos da população. Battesse, Harter e Fuller (1988) propuseram e aplicaram um modelo misto de intercepto aleatório para estimação da produção de soja em pequenas áreas (12 condados) no Estado de Iowa, usando dados amostrais e como variável auxiliar dados proporcionados pelo satélite Landsat. Seus estudos mostraram a superioridade dos estimadores baseados no modelo de intercepto aleatório em relação aos estimadores baseados nos modelos de regressão simples. Moura e Holt (1999) estenderam o trabalho de Prasad e Rao (1990) para um modelo hierárquico mais geral em que todos os coeficientes do modelo podem ser aleatórios e as variáveis em nível de pequenas áreas podem também ser utilizadas para explicarem possíveis diferenças entre as mesmas.

Mais recentemente tem surgido na literatura algumas aplicações de modelos hierárquicos ao problema de estimação em pequenas áreas sob a abordagem Bayesiana. You e Rao (1997) baseados em Moura (1994), propuseram modelos hierárquicos com erros heterocedásticos no último nível, e assumindo modelos com efeitos aleatórios para parâmetros de regressão e para a variância. Eles utilizaram o método de amostrador de Gibbs para serem geradas amostras da distribuição *a posteriori* para as médias das pequenas áreas.

Este trabalho tem como objetivo apresentar e aplicar modelos hierárquicos para estimação de proporções em pequenas áreas. Apresentam-se as abordagens Clássica e Bayesiana. A metodologia estatística é aplicada a dados de educação do Município do Rio de Janeiro. Foram consideradas 34 regiões do estado como as pequenas áreas. Considerando-se como níveis hierárquicos as regiões e os alunos.

Na Seção 1 é feita uma introdução ao problema de estimação em pequenas áreas, como também uma breve revisão literária relacionada a este trabalho. Na Seção 2, apresenta-se o modelo Binomial para estimação de pequenas áreas via modelos hierárquicos. Na Seção 3, descrevem-se os métodos de estimação para os parâmetros do modelo hierárquico binomial sob a abordagem Clássica, tais como o MQL "Marginal Quasi-likelihood" e o PQL "Penalized Quasi-likelihood". Na Seção 4, apresenta-se a estimação sob a abordagem Bayesiana, descrevendo-se o método do Amostrador de Gibbs dentro das técnicas MCMC. Na Seção 5, apresentam-se os pacotes computacionais empregados. Na Seção 6, descreve-se a metodologia de estimação das proporções. Na Seção 7, apresenta-se a descrição dos dados, assim como os modelos que foram utilizados. Na

Seção 8, faz-se uma comparação dos resultados obtidos. Finalmente, na Seção 9, apresentam-se as conclusões e as recomendações finais.

2. Modelo hierárquico binomial para estimação de pequenas áreas

Modelos hierárquicos têm sido propostos para a estimação em pequenas áreas, usualmente supondo-se normalidade para as variáveis de interesse, por exemplo, Holt e Moura (1999) e You and Rao (2000). Porém, poucos trabalhos têm sido feitos para o caso de variáveis discretas em que o parâmetro de interesse é uma proporção. Alguns autores têm considerado o problema de estimar proporções em pequenas áreas usando métodos Bayesianos empíricos e completos. Nesta direção, pode-se citar o artigo Dempster e Tomberlin (1980) que propõem estimar proporções nas áreas locais baseados em modelos de regressão logística contendo efeitos fixos e aleatórios. Esta proposta foi desenvolvida por Farrell, MacGibbon e Tomberlin(1997), onde apresentaram um estudo de simulação para comparar o erro quadrático médio do previsor do modelo hierárquico e o modelo de regressão logística sem efeito aleatório, concluindo que o modelo hierárquico é mais eficiente.

Malec et. al.(1997) apresentam uma abordagem Bayesiana completa para estimar proporções usando Cadeias de Markov em integração de Monte Carlo (MCMC), em particular o amostrador de Gibbs, mostrando que a estimação hierárquica produz estimativas mais precisas para os parâmetros do modelo do que os estimadores de regressão logística convencional. Moura, Migon e Ferreira(2000) propuseram uma abordagem Bayesiana para a escolha de modelos binários competitivos para o problema de previsão de pequenas áreas.

Investigações sociais freqüentemente estudam a relação entre indivíduo e sociedade. Estas investigações assumem que indivíduos interagem no contexto social em que eles residem. Desta maneira os indivíduos são influenciados pelos grupos sociais. As propriedades destes grupos são por sua vez influenciados pelos indivíduos que os formam.

Geralmente os indivíduos e os grupos sociais são conceituados como um sistema hierárquico de indivíduos e grupos, onde indivíduos e grupos definem níveis separados neste sistema hierárquico.

Consideremos um modelo com dois níveis de hierarquia para a variável resposta y_{ij}, onde y_{ij} é uma variável binária, dada por:

$$y_{ij} = \begin{cases} 1 \text{ se a unidade ij possuir o atributo de interesse} \\ 0 \text{ caso contrário} \end{cases}$$

onde o índice ij representa os dois níveis hierárquicos que são estabelecidos, j denota a unidade amostral no último nível (nível mais desagregado) e i representa o subíndice correspondente às pequenas áreas.

Assim, $y_{ij} = 1$, indica a resposta da j-ésima unidade amostral do segundo nível pertencente à i-ésima unidade amostral do primeiro nível (correspondente às pequenas áreas). Supõem-se que a variável resposta y_{ij} tenha a seguinte distribuição de probabilidade:

$y_{ij} \mid \pi_{ij} \sim$ Bernoulli(π_{ij}) e independentes. A variável y_{ij} pode ser explicada a partir da probabilidade de resposta π_{ij}.

Para relacionar π_{ij} e o vetor de covariâveis dadas por $X^T = (1, x_1, \ldots, x_p)$, propõe-se o seguinte modelo hierárquico logístico:

$$\text{Logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = X_{ij}^{T}\beta_i \qquad (1)$$

$$\beta_i = Z_i\gamma + v_i .$$

onde:

β_i é um vetor de parâmetros, no qual todas as componentes podem conter efeitos aleatórios. X_{ij}^{T} representa o vetor de variáveis auxiliares associadas com os efeitos fixos, Z_i é um vetor de variáveis associadas com as pequenas áreas, γ é o vetor de parâmetros de efeitos fixos na regressão logística. A quantidade v_i têm média 0 e variância Φ. Cabe mencionar que variáveis auxiliares podem ser definidas em qualquer nível de hierarquia. Algumas destas variáveis podem ser medidas de uma forma natural. Por exemplo, se considerarmos como primeiro nível a escola, neste nível pode ser medido o tamanho da escola e a sua localização. Por outro lado, se considerarmos como segundo nível o aluno, poderíamos definir variáveis como média de aproveitamento escolar do aluno.

3. Estimação dos parâmetros sobre a abordagem clássica

Para estimar os parâmetros do modelo em (1) consideremos a função inversa da função logito: $f(\bullet) = \{1+exp[-(\bullet)]\}^{-1}$. Goldstein (1991) apresenta um algorítmo para estimação em modelos hierárquicos não lineares onde a resposta não é uma função linear dos parâmetros, tanto na parte fixa (γ) como na parte aleatória (v).

Primeiro o modelo é linearizado através de uma expansão em serie de Taylor de f em torno de uma função $H_{ij}^{(t)}$, onde há duas escolhas para $H_{ij}^{(t)}$:

$$\text{(a) } H_{ij}^{(t)} = X_{ij}^{T}Z_i\hat{\gamma}^{(t)} \qquad \text{(b)} \quad H_{ij}^{(t)} = X_{ij}^{T}Z_i\hat{\gamma}^{(t)} + X_{ij}^{T}v_i^{(t)} \qquad (2)$$

onde os símbolos $\hat{\gamma}^{(t)}$ e $\hat{v}_i^{(t)}$ denotam respectivamente as estimativas de γ e v_i na t-ésima iteração.

A escolha de (a) usa somente o preditor da parte fixa e a estimação é feita pelo método MQL ("Marginal quasi-likelihood") e de (b) usa o preditor tanto da parte fixa como da parte aleatória, cuja estimação é dada pelo método PQL ("Penalized quasi-likelihood").

Desta forma tem-se que:

$$f_{ij}(H_{ij}^{(t+1)}) = f_{ij}(H_{ij}^{(t)}) + X_{ij}^{T}Z_i(\hat{\gamma}^{(t+1)} - \hat{\gamma}^{(t)})\, f'_{ij}(H_{ij}^{(t)}) + X_{ij}^{T}\hat{v}_i^{(t)}\, f'_{ij}(H_{ij}^{(t)}) + (X_{ij}^{T}\hat{v}_i^{(t)})^2 f''_{ij}(H_{ij}^{(t)})/2 \qquad (3)$$

As variáveis explanatórias em f são transformadas usando as primeiras e segundas derivadas da função não linear:

$$f'_{ij}(H_{ij}^{(t)}) = f_{ij}(H_{ij}^{(t)})(1+\exp H_{ij}^{(t)})^{-1} \text{ e } f''_{ij}(H_{ij}^{(t)}) = f'_{ij}(H_{ij}^{(t)})(1-\exp H_{ij}^{(t)})(1+\exp H_{ij}^{(t)})^{-1}$$

O primeiro termo à direita da equação (3) é o valor da parte fixa de f na correspondente $(t+1)$-ésima iteração do algoritmo Iterative Generalized Least Squares - IGLS - descrito por Golsdtein (1986). Goldstein (1986) e Goldstein (1989) apresentou respectivamente os algoritmos Iterative Generalized Least Squares - IGLS - e Restricted Iterative Generalized Least Squares - RIGLS -, para estimar os parâmetros do modelo descrito na equação (1). O algoritmo RIGLS produz estimativas aproximadamente não viciadas dos parâmetros do modelo. Goldstein e Rasbash (1996), comparam através de um estudo de simulação os métodos de aproximação como Marginal quasi likelihood - MQL - e o Penalized quasi likelihood - PQL -, definidos por Breslow e Clayton (1993). As comparações foram feitas para estimações de primeira e segunda ordem, tanto para o MQL como para o PQL, utilizando em ambos o método iterativo IGLS. Goldstein e Rasbash (1996), através de um estudo de simulação, demonstraram que o procedimento de segunda ordem PQL produz estimativas mais precisas.

4. Estimação dos parâmetros sobre a abordagem bayesiana

Na abordagem Bayesiana, o modelo logístico com efeito aleatório para dois níveis de hierarquia pode ser escrito na seguinte forma:

$y_{ij} \mid \pi_{ij} \sim$ Bernoulli (π_{ij}) e são condicionalmente independentes dado π_{ij} .

$$\text{Logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = X_{ij}^{T}\beta_i, \quad i = 1, \ldots, r \tag{4}$$

Os β_i são condicionalmente independente dado γ, Φ:

$\beta_i / \gamma, \Phi \sim N(Z_i\gamma, \Phi)$,

As seguintes distribuições *a priori* independentes são atribuídas aos hiperparâmetros:

$\gamma \sim N(0,\Omega)$, e $\Phi^{-1} \sim W(\alpha, R)$, onde: Ω, α e R são parâmetros *a priori* conhecidos, W(α, R) denota a distribuição de Wishart com parâmetros α e R. Representando-se por D os valores observados de y_{ij}, e θ os parâmetros do modelo, a distribuição de probabilidade conjunta $p(D,\theta)$, pode ser escrita por : $p(D,\theta) = p(D/\theta)\, p(\theta)$, onde $p(\theta)$ é a distribuição *a priori* de θ. Tendo observado D, o teorema de Bayes fornece:

$$p(\theta/D) = \frac{p(\theta)\, p(D \mid \theta)}{\int p(\theta)\, p(D \mid \theta)\, d\theta} \tag{5}$$

Infelizmente para o modelo descrito acima, a distribuição dada pela equação (5) não pode ser obtida analiticamente. Uma forma de obter a distribuição *a posteriori* é usando aproximações numéricas como Cadeias de Markov em integração de Monte Carlo.

4.1 Amostrador de Gibbs

Gilks e Wild (1992) define o amostrador de Gibbs como um método de Monte Carlo que tem como objetivo principal estimar a distribuição *a posteriori* desejada. A grande vantagem deste método é sua fácil implementação e eficiência computacional. O método amostrador de Gibbs é discutido em detalhes por Geman e Geman (1984) no contexto de processos espaciais. O Amostrador de Gibbs trabalha através de um esquema de atualização Markoviano. Foi popularizado por meio do artigo de Gelfand e Smith (1990), onde foram apresentadas várias situações em que esse método pode ser utilizado. Zeger e Karim (1991) aplicaram o amostrador de Gibbs para resolver o problema de modelos lineares generalizados com efeitos aleatórios. Uma questão de grande relevância, quando da utilização do Amostrador de Gibbs, é o diagnóstico de convergência. Neste trabalho, aplicou-se o procedimento de diagnóstico de convergência dado por Geweke (1992).

5. Pacotes computacionais

Um dos pacotes computacionais que utilizam os métodos clássicos de estimação MQL e PQL descritos em Goldstein (1991) é o MLn. O pacote MLn é um sistema de análise inicialmente aplicado em Ciências Sociais, utiliza métodos como o IGLS e o RIGLS fornecendo estimadores consistentes para os parâmetros do modelo, podendo ser usado para previsão, ajuste de modelos de análise de variância e em modelos de multinível. Atualmente, está disponível uma versão do MLn para Windows: MlWin. Esta nova versão além da análise sobre a abordagem Clássica tem implantado uma análise sobre a abordagem Bayesiana. Outra técnica implantada é o de reamostragem por Bootstrap. Dentro da abordagem Bayesiana, Gamerman (1996), expressa que um dos maiores problemas a impedir o desenvolvimento da inferência Bayesiana sempre foi a dificuldade de sua implantação em problemas práticos como: (1) a especificação da distribuição *a priori* e (2) a convergência da *posteriori* resultante. A primeira fonte de dificuldade está sendo aos poucos eliminada pela disponibilidade de linguagens simbólicas, onde várias especificações para a distribuição *a priori* podem ser facilmente acomodadas em sistemas computacionais. A segunda fonte de dificuldade foi em grande parte eliminada pela introdução de métodos iterativos como o amostrador de Gibbs, que possibilita a análise de modelos bastante complexos através de sucessivas descomposições em distribuições condicionais completas. Assim, um sistema que resolve boa parte deste problema é o WinBUGS (Spiegelhelter, Tomas, A. e Best, 2000). O pacote WinBUGS consiste num conjunto de funções que permite a especificação de modelos e das distribuições de probabilidade para todos o seus componentes aleatórios (observações e parâmetros). Para cada conjunto de dados e modelo utilizado, o WinBUGS fornece os valores amostrados de cada parâmetro monitorado a cada n-iterações a partir de uma determinada iteração m. Ambos valores n e m, bem como os parâmetros a serem monitorados são especificados pelo usuário. Além disso, o programa fornece automaticamente alguns sumários decorrentes da amostra obtida, tais como média, mediana e os intervalos de confiança. A monitoração de convergência da saída do BUGS pode ser feita através do sistema CODA (Best, Cowles e Vines, 1995). Trata-se de um conjunto de funções

implantado em conexão com o pacote BUGS. Este sistema CODA tem disponíveis técnicas de convergência, entre estas o diagnóstico de convergência de Geweke (1992).

6. Estimação das proporções

6.1 Abordagem clássica

A proporção populacional na i-ésima pequena área (θ_i) para a característica de interesse *y*, pode ser escrita como: $\theta_i = N_i^{-1} \sum_{i=1}^{N_i} y_{ij}$, onde N_i é o número total de unidades na i-ésima pequena área.

O parâmetro θ_i pode ser decomposto pela soma de duas parcelas :

$$\theta_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} y_{ij} \right) \tag{6}$$

onde s_i representa o conjunto de unidades amostradas no segundo nível na i-ésima pequena área de tamanho n_i. Royal (1970) define o modelo de Superpopulação como "Totais e medias populacionais que poderiam ser gerados a partir de modelos de regressão utilizando variáveis auxiliares". Estas variáveis auxiliares constituem em nosso trabalho as covariáveis dos alunos.

O estimador Bayesiano empírico de θ_i é:

$$\hat{\theta}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} \hat{\pi}_{ij} \right) \tag{7}$$

onde a estimativa de π_{ij} para $j \notin s_i$ é dada por: $\hat{\pi}_{ij} = [1 + \exp(X_{ij}^T \hat{\beta}_i)]^{-1}$

e $\hat{\beta}_i = Z_i \hat{\gamma} + \hat{v}_i$, $\hat{\gamma} = (\hat{\gamma}_0, \ldots, \hat{\gamma}_p)$ é o vetor de coeficientes de regressão estimados a partir da amostra selecionada e $\hat{v}_i$ é o efeito aleatório estimado na i-ésima pequena área.

6. 2 Abordagem bayesiana

Na abordagem Bayesiana, a inferência sobre os θ_i é feita obtendo as distribuições *a posteriori* das quantidades desconhecidas na expressão (6) dadas as características de interesse observadas na amostra $y(s) = \{y_{ij}: i \in s, j \in s_i\}$. Consideremos o parâmetro θ_i dado na equação (6). *A posteriori* de θ_i é obtido sobre

repetidas amostras geradas pelo método MCMC, denotadas por $\theta_i^{(k)}$, onde k indica o k-ésimo elemento da amostra gerada. $\theta_i^{(k)}$ é obtido da seguinte maneira: $\theta_i^{(k)} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\notin s_i} y_{ij}^{(k)}\right)$.

Para obter as amostras da distribuição *a posteriori* $p(y_{ij}/y(s))$ *para* $i=1, \ldots, r;$ e $j = 1, \ldots, n_i$, primeiro obtemos uma amostra da distribuição *a posteriori* de π_{ij} a partir da função de ligação seguinte: $logit(\pi_{ij}^{(k)}) = X_{ij}^T \beta_i^{(k)}$. Onde $\beta_i^{(k)}$ é *(k)* é-sima amostra gerada da distribuição *a posteriori* de β_i/Y. Desta maneira pode-se gerar uma amostra de $p(y_{ij}/Y)$ a partir da relação: $y_{ij}^{(k)} \sim Bin(\pi_{ij}^{(k)})$. A média *a posteriori* de θ_i pode ser estimada por: $\hat{E}_{\theta_i} = \frac{1}{L}\sum_{k=1}^{L} \theta_i^{(k)}$. A variância *a posteriori* de θ_i pode ser estimada por: $\hat{V}_{\theta_i} = \frac{1}{L}\sum_{K=1}^{L} (\theta_i^{(k)} - \hat{E}_{\theta_i})^2$, onde L é o tamanho da amostra de Monte Carlo a ser considerada após convergência cadeia.

7. Descrição dos dados e modelos

Os dados utilizados foram extraídos de uma avaliação realizada nas escolas da rede municipal do Rio de Janeiro em 1996, utilizando-se as mesmas instruções do Sistema de Avaliação do Ensino Básico - SAEB - realizado em 1995. A população alvo é constituída de 15 288 alunos da oitava série do primeiro grau no Município do Rio de Janeiro, que realizaram o teste de matemática. O Município foi dividido em 34 regiões de acordo com a localização e demarcação geográfica dos bairros. O número total de escolas considerado foi de 392. Uma vez que a população de interesse é conhecida, foi possível obter os parâmetros da população. Estes parâmetros foram utilizados para serem avaliadas as estimativas amostrais produzidas por diferentes métodos. A variável - resposta Y é binaria (1: aluno com proficiência dentro do primeiro quartil, 0: para os outros casos). As covariáveis utilizadas foram: idade dos alunos *(x_1 e x_2)* categorizadas, onde x_1 (0: ≤ 14 anos, 1: entre 15-16 anos, 0: ≥ 17), x_2 (0: ≤ 14 anos, 0: entre 15-16 anos, 1: ≥ 17), sexo: x_3 (0: masculino, 1: feminino), escolaridade dos pais: x_4 (0: não superior, 1:superior), localização da escola:
x_5 (0: rural, 1: urbana).

Para cada uma das 34 regiões foram selecionadas 25% das escolas, e para cada escola selecionada foram selecionadas aleatoriamente 40% dos alunos, resultando em uma amostra de 1695 alunos. A partir da amostra selecionada e utilizando os modelos descritos abaixo foram estimados os parâmetros dos respectivos modelos .

7.1 Modelos sobre a abordagem clássica

7.1.1. Regressão logística

No modelo de regressão logística as variáveis explicativas x_1, x_2, x_3, x_4, são variáveis em nível de aluno, exceto a variável x_5 que é medida no nível de escola. O modelo pode ser escrito concisamente da seguinte forma:

$y_{ij} \sim Bernoulli(\pi_{ij})$.

$logit(\pi_{ij}) = \gamma_0 + \gamma_1 x_{1ij} + \gamma_2 x_{2ij} + \gamma_3 x_{3ij} + \gamma_4 x_{4ij} + \gamma_5 x_{5ij}$

7.1.2. Modelo hierárquico sobre a abordagem clássica

No modelo hierárquico logístico, dois níveis de hierarquia são considerados: alunos e área, onde o intercepto inclui o efeito aleatório da i-ésima pequena área v_{0i}. Este modelo foi estimado utilizando o método PQL de segunda ordem descrito por Goldstein (1991).

$y_{ij} \sim Bernoulli(\pi_{ij})$

$logit(\pi_{ij}) = \beta_{0i} + \gamma_1 x_{1ij} + \gamma_2 x_{2ij} + \gamma_3 x_{3ij} + \gamma_4 x_{4ij} + \gamma_5 x_{5ij}$

$\beta_{0i} = \gamma_0 + v_{0i}$,

onde $[v_{0i}] \sim N(0, \sigma_v^2)$ são os efeitos aleatórios associados às pequenas áreas.

7.2 Modelos sob a abordagem bayesiana

7.2.1. Modelo hierárquico sob a abordagem bayesiana

Sob esta abordagem, o modelo dado na equação (4) para dois níveis hierárquicos pode ser escrito da seguinte forma:

$y_{ij} / \pi_{ij} \approx$ *i.i.d.* $Bernoulli(\pi_{ij})$,

$$Logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_0 + \gamma_1 x_{1ij} + \gamma_2 x_{2ij} + \gamma_3 x_{3ij} + \gamma_4 x_{4ij} + \gamma_5 x_{5ij} + v_{0i}$$

Assume-se que a distribuição condicional de v_{0i} dado $\tau_v = 1/\sigma_v^2$ tem média 0 e precisão τ_v. *A priori* vagas assumidas para os hiperparâmetros foram: os parâmetros de regressão fixos têm distribuição Normal com média 0 e precisão 10^{-6}; o hiperparâmetro τ_0 tem distribuição Gama de parâmetros 10^{-3} e 10^{-3}. Consideramos 7 000 amostras geradas via MCMC, descartam-se as primeiras 2 000 amostras (aquecimento da cadeia). Para obter a média *a posteriori* de θ_i consideramos as 5 000 últimas amostras obtidas via MCMC. Para a convergência da cadeia, utilizamos o método descrito por Geweke (1992). Considerando como primeira amostra 1 000 unidades depois do aquecimento da cadeia, a segunda amostra considerada foi obtida nas 1 000 últimas amostras da seqüência da cadeia (6 001 até 7 000). Utilizando a estatística Z (normal) como teste de hipótese para a

comparação de médias tanto para os parâmetros de regressão fixos (γ) quanto para a variância dos efeitos aleatórios das regiões (σ_v^2).

7. 3 Interpretação dos parâmetros de regressão dos modelos

Na **Tabela 1**, observa-se que a média dos parâmetros γ_1, γ_2, γ_3, γ_4, para cada modelo é positiva. Sendo, γ_1 positiva, implica que alunos com idade menor ou igual a 14 anos têm menor probabilidade de apresentarem proficiência baixa comparado com alunos entre 15-16 anos. Analogamente a interpretação é similar para o parâmetro γ_2. Desta forma, pode-se afirmar que alunos acima da idade escolar adequada para a oitava série apresentam maior chance de obterem proficiência baixa. Com relação aos parâmetros γ_3 e γ_4, as interpretações são análogas, assim alunos do sexo masculino têm menor probabilidade de apresentarem uma proficiência baixa e alunos cujos pais não possuem escolaridade superior têm maior probabilidade de apresentarem baixa proficiência.

Na **Tabela 1**, observa-se que as estimativas dos parâmetros de regressão do modelo hierárquico clássico (Modelo 2), aplicando o método PQL de segunda ordem são relativamente similares ao modelo de regressão logística (Modelo 1). Cabe mencionar que a variância para o nível mais desagregado foi padronizada a partir da seguinte relação: $e_{0ij}x_0^*$, onde: $x_0^* = [\pi_{ij}(1-\pi_{ij})/n_{ij}]^{0.5}$. Portanto na **Tabela 1** abaixo $\sigma_a^2 = V(e_{0ij}x_0^*)$. Assim quando o modelo é binomial espera-se que o valor estimado da variância seja próximo de 1. Nesta tabela apresentam-se também as estimativas dos parâmetros de regressão aplicando o método MCMC a partir de 5000 amostras (Modelo 3). As estimativas dos parâmetros são relativamente similares ao modelo hierárquico clássico.

Tabela 1 - Resumo das estimativas para os parâmetros de regressão nos modelos clássicos: logísticos e hierárquicos

Parâmetro	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5	σ_v^2	σ_a^2
Modelo 1	**-1,915 (0,192)**	**0,672 (0,148)**	**1,819 (0,183)**	**0,553 (0,127)**	**-0,411 (0,177)**	**-0,335 (0,154)**	**----**	**1,002 (0,034)**
Modelo 2	**-2,024 (0,372)**	**0,682 (0,158)**	**1,902 (0,202)**	**0,591 (0,138)**	**-0,346 (0,192)**	**-0,331 (0,370)**	**0,336 (0,124)**	**0,970 (0.034)**
Modelo 3	**-1,988 (0,390)**	**0,679 (0,148)**	**1,898 (0,189)**	**0,590 (0,134)**	**-0,359 (0,182)**	**-0,370 (0,388)**	**0,321 (0,119)**	**----**

Na **Tabela 2**, apresenta-se o diagnóstico de convergência de Geweke, a partir de duas amostras de 1000 unidades cada uma (de 2001 até 3000 e de 6001 até 7000). O valor da estatística de teste não deve ser grande caso a convergência foi atingida. Observa-se que a há indícios de convergência para todos os parâmetros.

Tabela 2 . Diagnóstico de convergência de Geweke via CODA no WinBugs

Parâmetro	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5	τ^2
Valor do Teste	**1,868**	**0,869**	**0,602**	**-1,267**	**-1,496**	**-1,736**	**1,026**
p-valor	**0.06**	**0.38**	**0.55**	**0.20**	**0.14**	**0.08**	**0.30**

8. Comparação das Previsões

Obtivemos a estimação da proporção de alunos com proficiência baixa (θ_i) em cada região, tanto para o modelo de regressão logística, como para o modelo hierárquico que utiliza o método de estimação PQL de segunda ordem.

Considerando-se o modelo hierárquico foram obtidas estimativas mais precisas, como pode ser observado no **Gráfico 3**.

Gráfico 2 - Proporção verdadeira e estimativas obtidas através da regressão logística

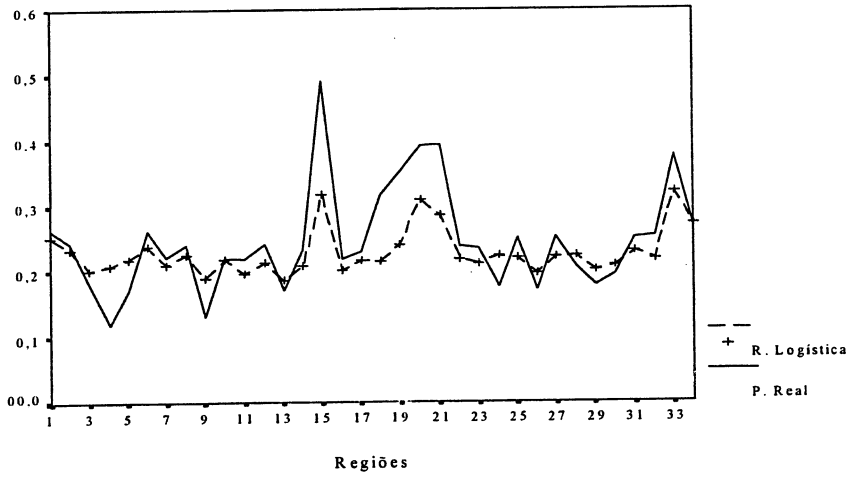


Gráfico 3 - Comparação dos métodos PQL e MCMC.

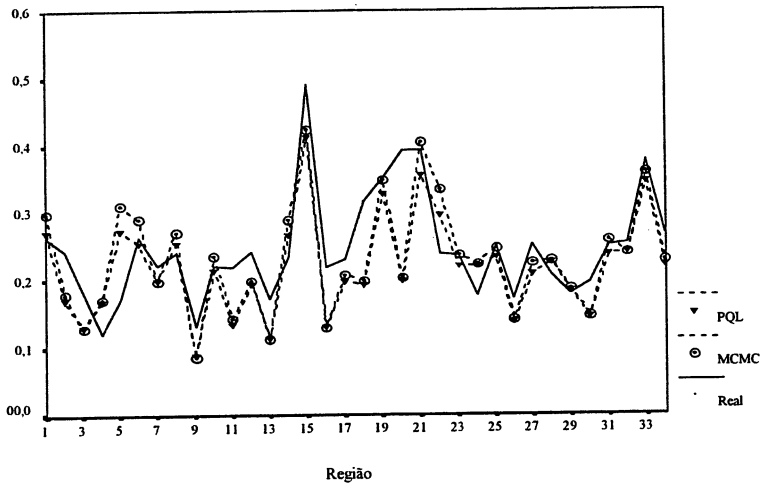
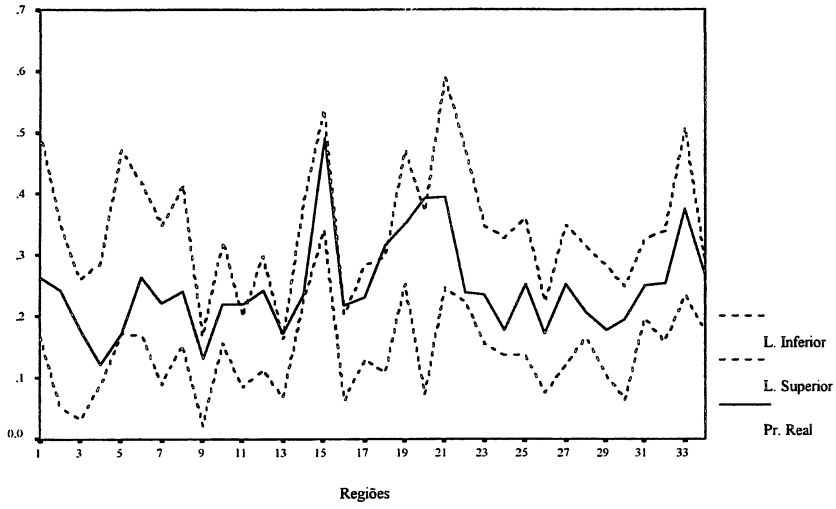
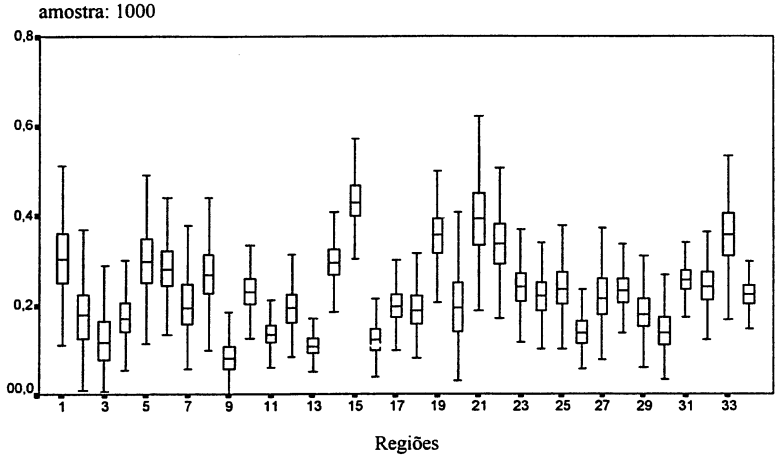


Gráfico 4 - Credibilidade (95%) para a proporção média de alunos com baixa proficiência



No **Gráfico 4**, apresentamos o parâmetro populacional θ_i para cada região e os limites de 95% de credibilidade θ_i obtidos via MCMC. Neste gráfico, observamos que a proporção real de alunos com proficiência menor que o primeiro quartil encontrou-se dentro destes limites de credibilidade.

Gráfico 5 - Box-plot de 1000 amostras via MCMC



No **Gráfico 5**, apresentamos as distribuições (*box-plots)* das últimas 1000 amostras para cada região, geradas via MCMC, onde as regiões 15, 21 e 33 apresentaram maior proporção de alunos com proficiência baixa.

9. Conclusões e recomendações

Para variáveis binárias, as estimações em pequenas áreas são melhoradas usando modelos hierárquicos. Introduzimos dois níveis hierárquicos: definidos por aluno e região. Neste trabalho, obtivemos valores próximos na estimação da proporção de alunos com proficiência baixa tanto na abordagem Clássica como na abordagem

Bayesiana. Contudo a abordagem Bayesiana para previsão em pequenas áreas é mais completa, pois permite adicionar informação *a priori* e obter as distribuições dos parâmetros desconhecidos.

Finalmente, é claro que neste trabalho, como em muitos outros, é pouco provável obter resultados satisfatórios para todas as pequenas áreas.

A medida de incerteza para estimadores de modelos hierárquicos podem depender da estrutura dos dados. A metodologia pode ser estendida a fim de se considerar a modelagem e estrutura espacial da população. Dados obtidos em pequenas áreas freqüentemente exibem uma estrutura espacial como acontece no Município do Rio de Janeiro, onde alunos com proficiência baixa concentram-se mais freqüentemente na zona oeste.

Apêndice: condicionais completas dos parâmetros do modelo

Neste Apêndice apresentam-se as distribuições condicionais completas de todos os parâmetros para o modelo hierárquico. Para maiores detalhes veja Castañeda (1999).

A distribuição condicional completa de γ:

$p(\gamma/Y, \beta_i, \Phi)$ é multivariada normal $N(m_\gamma^*, C_\gamma^*)$, onde:

$$m_\gamma^* = C_\gamma^* \left(C_\gamma^{-1} m_\gamma + \sum^r Z_i^T \Phi^{-1} \beta_i \right)$$

$$C_\gamma^{*-1} = C_\gamma^{-1} + \sum_{i=1}^{r} Z_i^T \Phi^{-1} Z_i$$

A distribuição condicional completa $p(\Phi \mid y, \beta, \gamma)$ é Wishart Invertida: $IW\ (n_\Phi^*, S_\Phi^*)$, onde:

$$n_\Phi^* = \alpha + r \quad \text{e} \quad (S_\Phi^*)^{-1} = R^{-1} + \sum_{i=1}^{r} (\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T$$

α e R são os respectivos parâmetros da distribuição *a priori* (Wishart) para Φ^{-1}.

A densidade condicional completa de $p(\beta_i \mid Y, \gamma, \Phi)$ é proporcional a:

$$\prod_{j=1}^{n_i} [p(y_{ij} \mid \beta_i)] p(\beta_i \mid \gamma, \Phi) \alpha \frac{\exp\left\{ \beta_i \sum_{j=1}^{n_i} x_{ij}^T y_{ij} \right\}}{\prod_{j=1}^{n_i} \{1 + \exp(x_{ij}' \beta_i)\}} \exp\{ -\frac{1}{2} (\beta_i - Z_i\gamma)^T \Phi^{-1} (\beta_i - Z_i\gamma) \}$$

Para gerar observações da densidade acima é utilizado um algoritmo de aceitação/rejeição definido como amostragem de rejeição.

Referências bibliográficas

BATTESE, G. E., HARTER, R. M. AND FULLER, W. A. (1988). An Error Component Model For Prediction of Country Crop Areas Using Survey and Satellite Data. *Journal of American Statistical Association,,* 83, 28-36.

BEST, N.G.,COWLES, M.K. AND VINES, S.K.(1995) CODA: *Convergence Diagnostics and Output Analysis Software for Gibbs Sampler Output: Version 03.* Technical Report, Biostatistics Unit_MRC, Cambridge, UK.

BRESLOW, N. E. AND CLAYTON, D, G. (1993). Aproximate Inference In Generalized Linear Mixed Models. *Journal of American Statistical Association,*. 88, 9-25.

BRACKSTONE, G.J.(1987). Small Area Data: Policy issues and Technical Chalenges. *Small Area Statistics.* Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh., Wiley, New York, 3-20.

CASTAÑEDA D.F.N.(1999). Estimação em Pequenas áreas para dados discretos via Modelos Hierárquicos. Dissertação de Mestrado, Departamento de Métodos Estatísticos, IM-UFRJ.

COCHRAN W.G. (1977) Sampling Techniques, 3ra Edição. Wiley, New York.

DATTA, G. S. AND GHOSH, M. (1991). Bayesian Prediction In Linear Models: Applications To Small Area Estimation. *Annals of Statistics,*13, 262-271.

DEMPSTER, A. P. AND TOMBERLIN, T. J. (1980). The Analysis Of Census Undercount From a Post Enumeration Survey. *Proc. Conference on Census Undercount,* Arligton, Va. , 88-89.

FARRELL, P. J., MACGIBBON, B. AND TOMBERLIN, T. J. (1997).Empirical Bayes Estimators of Small Area Proportions in Multistage Designs. *Statistica Sinica,* 7, 1065-1083.

GAMERMAN, D.(1996). Simulação Estatocástica Via Cadeias de Markov. *12º Sinape - Caxambu. Abe - Associação Brasileira De Estatística. São Paulo.*

GEMAN, S. AND GEMAN, D.(1984) "Stochastics Relaxation, Gibbs Distributions and The Bayesian Restoration of Imagens". *Ieee Transactions On Pattern Analysis And Machine Intelligence,* 6, 721-741.

GELFAND, A. E., E SMITH. A. F. M., (1990) "Sampling Based Approaches To Calculating Marginal Densities". *Journal of American Statistical Association,* 85, 398-409.

GEWEKE, J. (1992). Evaluating The Accuracy Of Sampling Based Approaches To The Calculation of Posterior Moments. *Bayesian Statistics.* 4. (Eds: J M Bernardo Et. Al.), Pp. 625 - 631. Oxford: University Press.

GILKS, W.R. E WILD, P.(1992). Adaptative Rejection Sampling for Gibbs Sampling. *Applied Statistics, 41, 337-348.*

GOLDSTEIN, H. (1986). Multilevel Mixed Linear Models Analysis Using Iterative Generalized Least Squares. *Biometrika,* 73, 43-56.

GOLDSTEIN, H. (1991). Nonlinear Multilevel Models With Application To Discrete Response Data. *Biometrika,* 78, 45-51.

GOLDSTEIN, H AND RASBASH, J. (1996). Improved Approximation For Multilevel Models With Binary Responses. *Journal. of the Royal. Statistical. Society,* Ser A, 159, 505-513.

GILKS, W.R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1996). Markov Chain Montecarlo In Practice, London: *Chapman & Hall.*

HOLT, D. E MOURA, F. (1993) Smallmall Area Estimation Using Multi-Level Models. Proceedings of The Section on Survey Research Methods. *Journal of American Statistical Association,* 1, 21-30

MALEC, D., SEDRANSK, J., MORIARITY, C. L. AND LE CLERE, F. B. (1997). Small Area Inference For Binary Variables In The National Hearlth Interview Survey. *Journal of the American Statistical Association* , 92, 815-826.

MOURA, F. (1994) Small Area Estimation Using Multilevel Models. Tese de Doutorado University of Southampton, UK.

MOURA, F. E HOLT, D. (1999) Small Area Estimation Using Multinivel Models. Survey Methodology. Vol 25, 1, 73-80.

MOURA, FA.S. , MIGON, H.S. E FERREIRA, M. A. R.(2000). Small Area Estimation for Bianry Data via Hierarchical Models. *Statistics in Transition, 4, 665-677.*

PRASAD , N.G.N. AND RAO, J.N.K.(1990). The Estimation of the Mean Square Error of Small Area Predictors. *Journal of the American Statistical Association, 85, 163-171.*

ROYALL, R. M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. *Biometrika*, 74, 1-12.

SPIEGELHELTER, D.J. , TOMAS, A. AND BEST, N.G.(2000). Winbugs Version 1.3. User Manual MRC Biostatistics Unit.

YOU, Y. AND RAO, J. N. K.(1999) Hierarchical Bayes Small Area Estimation Using Multi-Level Models. *Proceedings of Small Area Estimation Satellite Conference, Riga, Latvia*, 171-185.

ZEGER, S. L. AND KARIM, M. R.(1991). Generalized Linear Models with Random Effects: a Gibbs sampling approach. *Journal of American Statistical Association*, 90, 921-927.

ABSTRACT

Sampling Surveys are usually designed for providing estimates at a relative aggregated region level (ex: states). However, there is interest in obtaining information for small areas, where the respective sample sizes are usually small. Therefore, the usual direct survey estimators, as described in Cochran (1977), yield unacceptable large standard errors. In this situation, model-based approaches are recommended. In this work, Classical and Bayesian model-based approaches are presented for estimating small area binary data through a logistic hierarchical model. Both Bayesian and Classical approaches are applied to a Brazilian Educational Data. Finally, a critical analysis is made about the two kinds of approaches.

Key words: Small Area Estimation, Logistic Hierarchical Regression, MCMC.

Testes para comparação de séries temporais: uma aplicação a séries de temperatura e salinidade da água, medidas em profundidades diferentes

Clélia Maria de Castro Toloi *

Gladys Elena Salcedo Echeverry **

RESUMO

Na análise de séries temporais, muitas vezes, é de interesse verificar se duas séries, ou trechos de uma mesma série, estão sendo gerados pelo mesmo processo estocástico. No caso de processos estacionários de segunda ordem, as hipóteses de interesse são verificar se elas apresentam igual estrutura de autocovariância (autocorrelação) ou igualdade das funções de densidades espectrais. Neste trabalho, apresentamos vários testes, alguns para séries multivariadas, com esse objetivo. Uma aplicação com séries reais é dada.

1 . Introdução

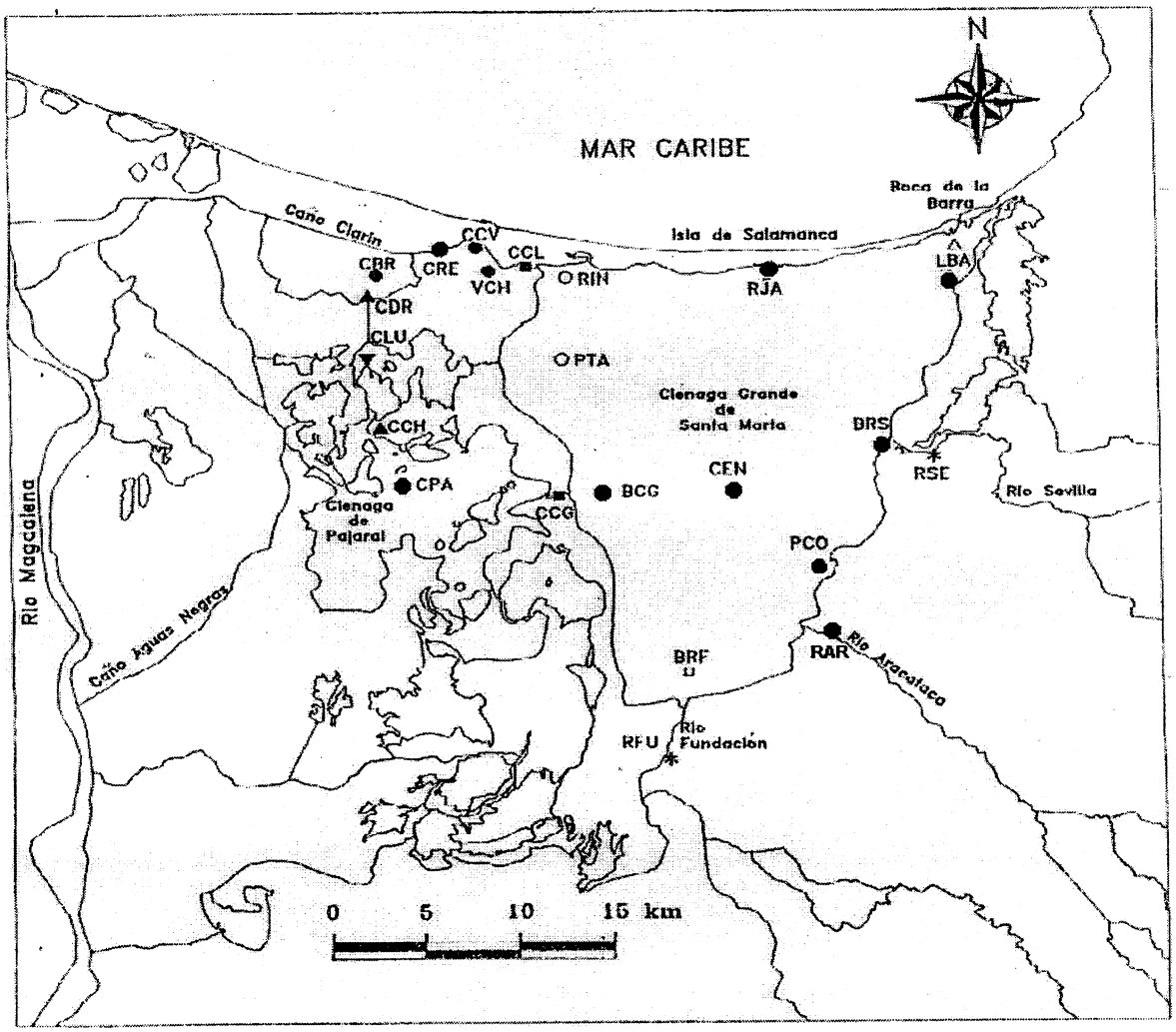
Na costa caribenha colombiana, encontra-se o maior ecosistema lagoa-estuarino, o país chamado "La Ciénaga Grande de Santa Marta" (veja Figura 1).

Este ecosistema é de muita importância para o país, já que é uma das principais fontes de fornecimento de peixes e mariscos para as cidades do litoral norte colombiano e, ainda, para mercados do interior do país. Desde 1956, aproximadamente, ele entrou em um sério deterioramento ambiental até hoje, ainda, refletido na morte massiva de mangue e na diminuição da abundância de peixes, aves e moluscos.

* Endereço para correspondência: Deptº de Estatística, IME-USP - Caixa Postal 66281 - CEP 05315-970 - São Paulo - SP.
** Departamento de Matemática e Computação - Universidad del Quindío-Colombia A. A. 460.

Nos diferentes estudos feitos sobre o ecosistema, pesquisadores têm tomado medições de diferentes variáveis físicas, químicas e biológicas, sendo que algumas destas variáveis como salinidade, temperatura e oxigênio têm sido medidas mensalmente em três profundidades de água: superfície, fundo e coluna.

Figura 1 - Delta exterior del Rio Magdalena



De algumas análises feitas usando estas séries, surgiu-nos a suspeita que as séries nas três profundidades poderiam estar sendo geradas pelo mesmo processo estocástico e conseqüentemente não precisariam ser medidas nas três profundidades diferentes.

Neste trabalho, apresentamos diferentes metodologias usadas para testar a igualdade de séries temporais univariadas e multivariadas. Na **seção 2,** apresentamos um pouco da teoria básica utilizada nos diferentes procedimentos. Na **seção 3,** apresentamos os diferentes testes usados para comparar séries univariadas e multivariadas. Na **seção 4,** aplicamos estes testes para verificar a conjectura de que salinidade e temperatura

medidas na superfície e no fundo da estação Boca de Caño Grande - BCG - da CGSM, estão sendo geradas pelo mesmo processo e, finalmente, na **seção 5** apresentamos algumas conclusões.

2 . Transformada de *Fourier* de processos estacionários

Seja $X(n), n = 0,1,\ldots,$ um processo estacionário real. A *transformada finita de Fourier* dos valores $X(0), X(1),\ldots, X(T-1)$ é dada por

$$d_x^{(T)}(\lambda) = \sum_{t=0}^{T-1} X(t) e^{-i\lambda t}, \quad -\infty < \lambda < \infty. \tag{1}$$

Se, em particular, $\lambda = \frac{2\pi s}{T}$, $s = 0,\ldots,T-1$, então os valores $d_x^{(T)}(\lambda) = d_x^{(T)}\left(\frac{2\pi s}{T}\right)$, $s = 0,\ldots,T-1$, constituem a *transformada discreta de Fourier* de $X(t)$, $t = 0,\ldots,T-1$.

Teorema 2.1: *Suponha* $\mathbf{X(t)} = (X_1(t),\ldots,X_r(t))'$ $t = 0,\pm1,\ldots$ *um processo r-variado, estritamente estacionário com função cumulante de ordem k satisfazendo*

$$\sum_{u_1,\ldots,u_{k-1}=-\infty}^{\infty} |C_{\mathbf{x}}(u_1,\ldots,u_{k-1})| < \infty.$$

Seja $s_j(T)$ *um inteiro com* $\lambda_j(T) = \frac{2\pi s_j(T)}{T} \to \lambda_j$ quando $T \to \infty$ *para* $j = 1,2,\ldots,J$, *se* $2\lambda_j(T)$, $\lambda_j(T) \pm \lambda_k(T) \quad 0(\mathrm{mod}\, 2\pi)$, $1 \le j < k \le j$ 1,2,..., $d_x^{(T)}(\lambda_j(T)) \xrightarrow{D} N^c(0, 2\pi T f_x(\lambda_j))$ *onde* f_x *representa a função de densidade espectral da série X(t). Além disso,* $d_x^{(T)}(\lambda_j(T))$ e $d_x^{(T)}(\lambda_k(T))$ *são assintóticamente independentes.*

O periodograma de uma série real estacionária $X(t), t = 0,\ldots,T-1$ é definido por

$$I_x^{(T)}(\lambda) = (2\pi T)^{-1} \left| d_x^{(T)}(\lambda) \right|^2, \qquad \lambda \quad 0 \pmod{2\pi}. \tag{2}$$

Sob as suposições do Teorema 2.1, demonstra-se que $I_x^{(T)}(\lambda)$ $I_x^{(T)}(\lambda_j(T)) \xrightarrow{D} f_x(\lambda_j)\chi_2^2/2$, isto é, o periodograma $I_x^{(T)}(\lambda)$ é um estimador assintoticamente não viciado de $f_x(\lambda)$ embora inconsistente.

Se $\mathbf{X}(t), t = 0,\pm1,\ldots,T-1$ é uma série r-variada com média $\mu_{\mathbf{x}}$ e matriz de densidade espectral $\mathbf{f}_{\mathbf{x}}(\lambda)$ a *transformada finita de Fourier* de $\mathbf{X}(t)$ é dada por

$$d_{\mathbf{x}}^{(T)}(\lambda) = \left[d_j^{(T)}(\lambda) \right] = \left[\sum_{t=0}^{T-1} X_j(t) e^{-i\lambda t} \right], \quad -\infty < \lambda < \infty, \quad j = 1,\ldots,r.$$

Teorema 2.2: *Se a série r-variada* $\mathbf{X}(t)$ *satisfaz as condições do Teorema* 2.1, *então*

$$d_{\mathbf{x}}^{(T)}(\lambda) \xrightarrow{D} N_r^c(0, 2\pi T f_{\mathbf{x}}(\lambda)) \quad se \quad \lambda \quad 0 \pmod{\pi}.$$

Um estimador de $\mathbf{f}_{\mathbf{x}}{}^{(T)}(\lambda)$, no caso $\lambda 0, \pm 2\pi, \ldots$, é dado pela estatística

$$I_{\mathbf{x}}^{(T)}(\lambda) = \left[I_{ij}^{(T)}(\lambda)\right] = \left[(2\pi T)^{-1} d_i^{(T)}(\lambda)\overline{d_j^{(T)}(\lambda)}\right] \tag{3}$$

com $\lambda = \frac{2\pi s}{T}, s = 1, \ldots, \left[\frac{T-1}{2}\right]$, $i, j, = 1, \ldots, r$. O estimador $I_{\mathbf{x}}^{(T)}(\lambda)$ também é assintoticamente não viciado embora inconsistente para $\mathbf{f}_{\mathbf{x}}$.

Sob as condições do Teorema 2.2, prova-se que $I_{\mathbf{x}}^{(T)}(\lambda_j(T)) \xrightarrow{D} W_1^c(1, \mathbf{f}_{\mathbf{x}}(\lambda_j))$.

Além disso, $I_{\mathbf{x}}^{(T)}(\lambda_j(T))$ e $I_{\mathbf{x}}^{(T)}(\lambda_k(T))$ são assintoticamente independentes.

Um estimador para $\mathbf{f}_{\mathbf{x}}$, com melhores propriedades, é dado por

$$\mathbf{f}_{\mathbf{x}}{}^{(T)}(\lambda) = 2m+1)^{-1} \sum_{s=-m}^{m} I_{\mathbf{x}}{}^{(T)}\left(\frac{2\pi[s(T)+s]}{T}\right) \quad \text{se} \quad \lambda \quad 0 \pmod{\pi} \tag{4}$$

e é denominado *estimador suavizado da matriz de densidade espectral.*

Se r = 1, o estimador $f_x^{(T)}(\lambda)$ é denominado *periodograma suavizado.*

Teorema 2.3: *Sob as condições do Teorema 2.2,* $\mathbf{f}_{\mathbf{x}}^{(T)}(\lambda) \xrightarrow{D} (2m+1)^{-1} W_r^c(2m+1, \mathbf{f}_{\mathbf{x}}(\lambda))$ *se* $\lambda \not\equiv 0 \pmod{\pi}$ *onde* $\mathbf{f}_x^{(T)}(\lambda_j)$, $j = 1, \ldots, J$, *são assintoticamente independentes se* $\lambda_j \pm \lambda_k \not\equiv 0 \pmod{2\pi}$ *para* $1 \le j < k \le J$. *Quando* r = 1, $f_x^{(T)}(\lambda_j(T)) \xrightarrow{D} f_x(\lambda_j)\chi^2_{4m+2}$ /(4*m*+2).

Para maiores detalhes consultar Brillinger(1981) e Priestley(1981).

3. Métodos de comparação de séries temporais

Considere duas séries *r*-variadas, finitas e estacionárias

$$\{\mathbf{X}^{(1)}(t); t = 1, \ldots, n\} \quad \text{e} \quad \{\mathbf{X}^{(2)}(t); t = 1, \ldots, n\}$$

com $\mathbf{X}^{(i)}(t) = [X_1^{(i)}(t), \ldots, X_r^{(i)}(t)]'$ um vetor que contém *r* séries estacionárias, $i = 1,2$.

Nosso interesse é saber se as duas séries multivariadas foram geradas por um mesmo processo estacionário $\{\mathbf{X}(t)\}$.

3.1 Comparação de duas séries univariadas

O caso mais simples acontece quando r = 1 e as duas séries são independentes, ou seja, temos duas séries univariadas independentes $\{\mathbf{X}_{1t}\}$ e $\{\mathbf{X}_{2t}\}$ e queremos saber se elas são geradas pelo mesmo processo univariado $\{\mathbf{X}(t)\}$.

Assumindo que as duas séries univariadas $\{\mathbf{X}_{it}\}, i=1,2$ são Gaussianas com $f_i(\lambda)$, $\gamma_i(j)$ e $\rho_i(j)$, $i=1,2$, suas funções de densidade espectral, autocovariância e autocorrelação, respectivamente, o problema se reduz a testar uma das seguintes hipóteses:

$$1. \quad \begin{array}{ll} H_{01}: f_1(\lambda) = f_2(\lambda) & \textit{para todo} \quad 0<\lambda<\pi \\ H_{A1}: f_1(\lambda) \neq f_2(\lambda) & \textit{para a lg um} \quad 0<\lambda<\pi \end{array} \tag{5}$$

$$2. \quad \begin{array}{ll} H_{02}: \gamma_1(j) = \gamma_2(j) & \textit{para todo} \quad j=0,\pm1,\ldots \\ H_{A2}: \gamma_1(j) \neq \gamma_2(j) & \textit{para a lg um} \quad j=0,\pm1,\ldots \end{array} \tag{6}$$

$$3. \quad \begin{array}{ll} H_{03}: \rho_1(j) = \rho_2(j) & \textit{para todo} \quad j=0,\pm1,\pm2,\ldots \\ H_{A3}: \rho_1(j) \neq \rho_2(j) & \textit{para a lg um} \quad j=0,\pm1,\pm2,\ldots \end{array} \tag{7}$$

A seguir apresentamos os testes para verificar estas hipóteses.

Os dois testes foram sugeridos por Coates e Diggle(1986).

3.1.1 Teste das somas acumuladas (SA-Coates e Diggle)

Seja $I_i^{(T)}(\lambda)$ o periodograma da série $\{\mathbf{X}_{it}\}$, $i=1,2$. Sabe-se que, assintoticamente, $I_i^{(T)}(\lambda) \sim f_i(\lambda)\chi_2^2/2, \lambda \neq 0,\pi$. Definindo as razões espectrais

$$J(\lambda) = \frac{I_1^{(T)}(\lambda)}{I_2^{(T)}(\lambda)} \qquad \text{e} \qquad U(\lambda) = \frac{f_1(\lambda)}{f_2(\lambda)}, \qquad 0<\lambda<\pi$$

com $\{\mathbf{X}_{1t}\}$ e $\{\mathbf{X}_{2t}\}$ independentes temos que, assintoticamente,

$$J(\lambda) = \frac{I_1^{(T)}(\lambda)}{I_2^{(T)}(\lambda)} \sim U(\lambda)F_{2,2},$$

onde F é a distribuição de Fisher-Snedecor.

Pode-se demonstrar, veja Coates e Diggle(1986), que os valores.

$$z_i = 1n(1 + J^{-1}(\lambda_i)) \sim U(\lambda_i)\exp(1)$$

com $\lambda_i = \frac{2\pi i}{T}$, $i = 1,\ldots,m$, onde $m = \left\lfloor \frac{T-1}{2} \right\rfloor$ e $\exp(1)$ denota a distribuição exponencial de média 1.

Sob a hipótese H_{01}, dada por (5), $U(\lambda_i) = 1$ e, assintóticamente, $z_i \sim \exp(1)$. Assim $c_j = \sum_{i=1}^{j} z_i$ constituem os pontos de um processo de Poisson e, conseqüentemente, $\left\langle \frac{c_j}{c_m} \right\rangle, j = 1,\ldots,m$ é o vetor das estatísticas de ordem da distribuição uniforme no intervalo (0,1).

O primeiro teste proposto por Coates e Diggle(1986), consiste em construir as estatísticas $\left\langle \frac{c_j}{c_m} \right\rangle, j = 1,\ldots,m$ e usar a estatística de Kolmogorov-Smirnov para testar afastamentos da distribuição $U(0,1)$, ou equivalentemente, pode-se calcular simplesmente as estatísticas z_i e aplicar o teste de Kolmogorov-Smirnov para testar afastamentos da distribuição exponencial de média 1.

3.1.2 Teste da razão de verossimilhança (RV-Coates e Diggle)

Demonstra-se que

$$1nJ(\lambda) \sim \textit{logística}\{1nU(\lambda),1\}, \tag{8}$$

resultado que permite a construção de um teste da razão de verossimilhança generalizada, dentro da estrutura da distribuição logística.

Adotando um modelo quadrático para $1nU(\lambda)$

$$1nU(\lambda) = \alpha + \beta\lambda + \gamma\lambda^2 \tag{9}$$

temos que,

$$H_{01} : f_1(\lambda) = f_2(\lambda),\ 0 < \lambda < \pi \quad \Leftrightarrow \quad H_{01} : \alpha = \beta = \gamma = 0.$$

A escolha do modelo quadrático deve-se à sua flexibilidade aos tipos de quebra de comportamento de $1nU(\lambda)$, sob H_{01}. Seja $t_i = 1nJ(\lambda_i), i = 1,\ldots,m$, então de (8) e (9) temos que, assintóticamente, $t_i \sim \textit{logística}(1nU(\lambda_i),1)$ e a função de densidade de $t_i, i = 1,\ldots,m$ é dada por

$$f_T(t_i) = \frac{e^{-(t_i - 1nU(\lambda_i))}}{\left\{1 + e^{-(t_i - 1nU(\lambda_i))}\right\}^2} \qquad i = 1,\ldots,m.$$

Conseqüentemente, a função de verossimilhança dos t_i é dada por

$$L(\Theta/t) = \prod_{i=1}^{m} f_T(t_i).$$

Aplicando o logaritmo temos

$$ln(L(\Theta/t) = \sum_{i=1}^{m}(-t_i + lnU(\lambda_i)) - 2\sum_{i=1}^{m} ln(1 + e^{(-t_i + lnU(\lambda_i)})$$

$$= \sum_{i=1}^{m}(-t_i + \alpha + \beta\lambda_i + \gamma\lambda_i^2) - 2\sum_{i=1}^{m} ln(1 + e^{(-t_i + \alpha + \beta\lambda_i + \gamma\lambda_i^2)})$$

e os estimadores de máxima verossimilhança das componentes do vetor $\Theta = (\alpha, \beta, \gamma)$ podem ser obtidos numericamente através do algoritmo de Newton-Raphson.

O segundo teste proposto por Coates e Diggle(1986) para testar a hipótese H_{01} dada por (5), baseia-se na distribuição assintótica da razão de verossimilhanças, ou seja, baseia-se no fato que

$$R_v^{'} = -2ln\frac{\sup_{\Omega 01} L(\Theta)/t)}{\sup_{\Omega} L(\Theta)/t)} \sim \chi_3^2.$$

Logo, rejeita-se H_{01} para valores grandes de $R_v^{'}$.

Os dois testes anteriores requerem que as duas séries tenham o mesmo tamanho.

3.1.3 Teste de igualdade das funções de autocovariância (Mélard e Roy)

Para testar a hipótese H_{02} dada por (6), Mélard e Roy(1984) propuseram o teste descrito a seguir.

Seja $\{X_{kt}, t = 1,\ldots,n_k\}$ séries estacionárias. Considere então, para a *k*-ésima série, o vetor $\mathbf{c}_k$ das primeiras $J+1$ autocovariâncias estimadas

$$\mathbf{c}_k = [c_k = (0), c_k = (1), c_k = (2),\ldots, c_k = (J)]'$$

onde para cada $j = 0,1,\ldots,J$ o estimador $c_k(j)$ é dado por

$$c_k(-j) = c_k(j) = \frac{1}{n_k}\sum_{t=1}^{n_k - j}(X_{kt} - \overline{X_k})(X_{k,t+j} - \overline{X_k}) \qquad (10)$$

e $\overline{X_k} = \frac{1}{n_k}\sum_{t=1}^{n_k} X_{kt}$ é a média amostral da *k*-ésima série.

A matriz de variâncias e covariâncias do vetor $\mathbf{c}_k$ é dada por

$$\Sigma_k = \lim_{n_k \to \infty} n_k E[(\mathbf{c}_k - \gamma_{k)}(\mathbf{c}_k - \gamma_k)'].$$

e, assintoticamente,

$$\mathbf{c}_k \sim N_{J+1}(\gamma_k, n_k^{-1}\Sigma_k).$$

Além disso, sob a suposição que as duas realizações são independentes,

$$\mathbf{c}_1 - \mathbf{c}_2 \sim N_{J+1}(\gamma_1 - \gamma_2, n_1^{-1}\Sigma_1 + n_2^{-1}\Sigma_2),$$

e, sob H_{02}, $\Sigma_1 = \Sigma_2 = \Sigma$ e

$$\mathbf{c}_1 - \mathbf{c}_2 \sim \mathbf{N}_{J+1}(\mathbf{0}, (n_1 + n_2)(n_1 n_2)^{-1}\Sigma).$$

Colocando $n = n_1$ e supondo que $\frac{n_1}{n_2}$ é constante para $n \to \infty$ e, além disso, denotando

$$\mathbf{Z}^{(n)} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} (\mathbf{c}_1 - \mathbf{c}_2)$$

temos,

$$\mathbf{Z}^{(n)} \xrightarrow{D} N_{J+1}(\mathbf{0}, \Sigma) \qquad \text{e} \qquad \mathbf{Z}^{(n)'}\Sigma^{-1}\mathbf{Z}^{(n)} \xrightarrow{D} \chi^2_{J+1}.$$

A proposta de Mélard e Roy(1984) consiste em testar a hipótese H_{02} dada por (6) utilizando a estatística

$$\mathbf{Q}^{(n)} = \mathbf{Z}^{(n)'}(\hat{\Sigma}^{(n)})^{-1}\mathbf{Z}^{(n)},$$

onde $(\hat{\Sigma}^{(n)})^{-1}$ é um estimador convergente em probabilidade para Σ^{-1}. Utilizando o Teorema 5.1 de Billingsley(1968) temos que

$$\mathbf{Q}^{(n)} \xrightarrow{D} \chi^2_{J+1}.$$

De acordo com Anderson (1971) podemos escrever o elemento (l,i), $l,i = 0,1,\ldots,J$, de Σ na forma

$$\sigma_{li} = \theta_{i-l} + \theta_{i+l}$$

$$\text{com} \qquad \theta_i = \sum_{j=-\infty}^{\infty} \gamma(j)\gamma(i+j). \tag{11}$$

Mélard e Roy(1984) propõem um novo estimador para θ_i da forma

$$\tilde{\theta}_i = \sum_{j=-n+1}^{n-i-1} w(jb_n)w((i+j)b_n)c(i+j) \tag{12}$$

com as mesmas hipóteses que Robinson(1977) exceto que $w^2(\cdot)$ é uma função integrável e a função $w_j = w(jb_n))$, j inteiro, uma função positiva. Loève(1978) mostra que a função $w_j c(j)$ é também uma função de autocovariância de forma que ao substituir $\gamma(j)$ por $w_j c(j)$ em (11) obtém-se uma matriz $\hat{\Sigma}$ positiva definida como era requerida.

Finalmente pode-se demonstrar que $\tilde{\theta}_i$ converge para θ_i na norma L_1.

Para maiores detalhes consultar Mélard e Roy(1984). Neste artigo, a função w utilizada é a janela de Bartlett dada por

$$w_j(b_{n,H}) = \begin{cases} 1 - \dfrac{|j|}{b_{n,H}} & |j| \leq b_{n,H} \\ 0 & |j| > b_{n,H}. \end{cases}$$

3.1.4 Teste de igualdade das funções de autocorrelação (Quenouille)

Baseado nas suposições:

1. Se o ajuste de um modelo auto-regressivo para uma série $\{X_t\}$ for adequado, de modo que os resíduos sejam independentes, as autocorrelações parciais $v_j\, j = 0,1,\ldots,$ destes resíduos são assintoticamente independentes e distribuídas normalmente com variâncias assintóticas $\frac{1}{n-j}$, onde n é o tamanho da série; e
2. O resultado anterior não é sensível a imprecisões no ajuste, Quenouille(1958) propôs um teste para a hipótese H_{03} dada por (7).

O procedimento de Quenouille para testar diferenças entre dois conjuntos de autocorrelações é dado pelos seguintes passos:

1. Obter a função de autocorrelação $\hat{\rho}_1(j)$ e $\hat{\rho}_2(j)$, $j = 0,1,\ldots,J$, das séries $\{X_{1t}\}$ e $\{X_{2t}\}$, respectivamente;
2. Obter a média ponderada de $\hat{\rho}_1(j)$ e $\hat{\rho}_2(j)$, da forma $\hat{\rho}(j) = \frac{n_1\hat{\rho}_1(j)+n_2\hat{\rho}_2(j)}{n_1+n_2}$, obtendo uma função de autocorrelação comum as duas séries; n_1 e n_2 são os tamanhos das séries $\{X_{1t}\}$ e $\{X_{2t}\}$, respectivamente;
3. Calcular a função de autocorrelação parcial estimada comum, $\hat{\Phi}(k)$, a partir de $\hat{\rho}(j)$;
4. Identificar a ordem auto-regressiva p através de $\hat{\Phi}(k)$;.
5. Estimar os p coeficientes do modelo $AR(p)$, resolvendo as equações de Yule-Walker;
6. Filtrar cada série com estes coeficientes, isto é, ajustar a cada série o modelo $AR(p)$ com os coeficientes comuns encontrados no item 5 e obter as duas séries residuais $\{\hat{a}_{1t}\}$ e $\{\hat{a}_{2t}\}$;
7. Calcular as facp, v_j e v'_j das duas séries residuais $\{\hat{a}_{1t}\}$ e $\{\hat{a}_{2t}\}$, respectivamente; e
8. Testar se $\frac{v_j - v'_j}{\sqrt{\frac{1}{n_1-j}+\frac{1}{n_2-j}}}$ tem distribuição aproximadamente $N(0,1)$, ou equivalentemente, testar se

$$SQ = \sum_{j=1}^{J} \frac{(v_j - v'_j)^2}{\frac{1}{n_1 - j} + \frac{1}{n_2 - j}} \sim \chi^2_J .$$

A hipótese H_{03} dada por (7) é rejeitada, a um nível de significância α, se $SQ > C_\alpha$ onde C_α é tal que $P(\chi^2_J > C_\alpha) = \alpha$.

Salcedo(1999) fez algumas simulações com o objetivo de comparar o poder dos testes univariados mencionados anteriormente.

3.2 Comparação de duas séries multivariadas

Consideremos novamente as duas séries r-variadas,

$$\{\mathbf{X}^{(1)}(t) = [X_1^{(1)}(t), X_2^{(1)}(t), \ldots, X_r^{(1)}(t)], \quad t \in \mathbf{Z}\}$$

$$\{\mathbf{X}^{(2)}(t) = [X_1^{(2)}(t), X_2^{(2)}(t), \ldots, X_r^{(2)}(t)], \quad t \in \mathbf{Z}\}.$$

Assume-se que este processo é estacionário pelo menos no sentido fraco e que sua função de covariância é absolutamente integrável. Isto garante a existência da função matricial de densidade espectral $\mathbf{f}(\lambda)$ dada por

$$\mathbf{f}(\lambda) = \begin{bmatrix} \mathbf{f}_{1,1}(\lambda) & \vdots & \mathbf{f}_{1,2}(\lambda) \\ \cdots & \cdots & \cdots \\ \mathbf{f}_{2,1}(\lambda) & \vdots & \mathbf{f}_{2,2}(\lambda) \end{bmatrix}_{(2r)x(2r)} .$$

onde cada matriz $\left|\mathbf{f}_{i,j}\right|_{rxr}\ i, j = 1,2$ é a matriz de densidades espectrais e densidades espectrais cruzadas. Para saber se as duas séries multivariadas estacionárias $\{\mathbf{X}^{(1)}(t)\}$ e $\{\mathbf{X}^{(2)}(t)\}$ são geradas pelo mesmo processo estacionário, Carmona e Wang(1996) propuseram dois testes no domínio das freqüências, um para o caso de séries independentes e outro para o caso de séries dependentes e que, basicamente, são uma generalização da idéia do teste das somas acumuladas proposto por Coates e Diggle(1986) para o caso univariado.

O objetivo é comparar as matrizes de densidade espectral, ou seja, o teste de interesse é

$$H_0 : \mathbf{f}_{1,1}(\lambda) \equiv \mathbf{f}_{2,2}(\lambda) \qquad 0 < \lambda < \pi \tag{13}$$

$$H_A : \mathbf{f}_{1,1}(\lambda) \not\equiv \mathbf{f}_{2,2}(\lambda) \qquad \text{para algum} \quad 0 < \lambda < \pi .$$

Sem perda de generalidade, assume-se que o número de observações T é um múltiplo de L, $T = 2Lm$. Considera-se, também, o estimador suavizado da matriz de densidade espectral $\mathbf{f}$ da forma

$$\hat{\mathbf{f}}(\lambda_j) = \frac{1}{L}\sum_{l=1}^{L} d_X^{(T)}(\lambda_{j,l}) d_X^{(T)}(\lambda_{j,l})^* \tag{14}$$

onde $d_X^{(T)}(\lambda_{j,l})$ é a transformada de Fourier discreta da série 2*r*-variada $\{\mathbf{X}^{(1)}\}$ e $\{\lambda_{j,l} : l = 1,\ldots L\}$ é um conjunto de L freqüências que convergem para a freqüência λ_j, $j = 1,\ldots,m$.

Sabe-se que, $\hat{\mathbf{f}}(\lambda_1),\ldots,\hat{\mathbf{f}}(\lambda_m)$, quando $T \to \infty$ e L fixo, convergem em distribuição para

$$\left[\hat{\mathbf{f}}(\lambda_1),\ldots,\hat{\mathbf{f}}(\lambda_m)\right] \xrightarrow{D} \left[L^{-1}W_{2r}^{c}(L,\mathbf{f}(\lambda_1)),\ldots,L^{-1}W_{2r}^{c}(L,\mathbf{f}(\lambda_m))\right]$$

independentes. Também,

$$\left[\hat{\mathbf{f}}_{i,j}(\lambda_1),\ldots,\hat{\mathbf{f}}_{i,j}(\lambda_m)\right] \xrightarrow{D} \left[L^{-1}W_{2r}^{c}(L,\mathbf{f}_{i,j}(\lambda_1)),\ldots,L^{-1}W_{r}^{c}(L,\mathbf{f}_{i,j}(\lambda_m))\right], \qquad i,j = 1,2$$

independentes.

3.2.1 Comparação de séries no caso de independência (Carmona e Wang)

Quando $\{\mathbf{X}^{(1)}(t)\}$ e $\{\mathbf{X}^{(2)}(t)\}$ são independentes, temos que

$$\mathbf{f}_{1,2}(\lambda) \equiv \mathbf{f}_{2,1}(\lambda) \equiv \mathbf{0}$$

e, neste caso, a matriz de densidade espectral $\mathbf{f}(\lambda)$ é uma matriz bloco diagonal.

Da independência das séries segue-se que $\hat{\mathbf{f}}_{1,1}$ e $\hat{\mathbf{f}}_{2,2}$ são independentes. Logo as m matrizes de dimensão rxr, $M_1(\lambda_j)$ e $M_2(\lambda_j)$, definidas por

$$M_1(\lambda_j) \equiv \mathbf{f}_{1,1}^{-1/2}(\lambda_j)\hat{\mathbf{f}}_{1,1}(\lambda_j)\mathbf{f}_{1,1}^{-1/2}(\lambda_j)$$

e $$M_2(\lambda_j) \equiv \mathbf{f}_{2,2}^{-1/2}(\lambda_j)\hat{\mathbf{f}}_{2,2}(\lambda_j)\mathbf{f}_{2,2}^{-1/2}(\lambda_j)$$

$j = 1,\ldots,m$, formam amostras independentes de matrizes independentes com distribuição $W_r^c(L,\mathbf{I})$ onde $\mathbf{I}$ representa a matriz identidade de dimensão r.

Sob a hipótese H_0 dada por (13), a distribuição da variável aleatória

$$\begin{aligned} T(\lambda_j) &= traço\left[M_1(\lambda_j)M_2^{-1}(\lambda_j)\right] \\ &= traço\left[\hat{\mathbf{f}}_{1,1}(\lambda_j)\hat{\mathbf{f}}_{2,2}^{-1/2}(\lambda_j)\right], \end{aligned} \tag{15}$$

independe das matrizes de densidade espectral $\hat{\mathbf{f}}_{1,1}$ e $\hat{\mathbf{f}}_{2,2}$. Se $F_{r,L}$ representa a distribuição da estatística $T(\lambda_j)$, então esta distribuição independe da freqüência λ_j, pois M_1 e M_2 têm distribuição $W_r^c(L,\mathbf{I})$.

O teste proposto por Carmona e Wang(1996), para testar a hipótese H_0, dada por (13), baseia-se na estatística $T(\lambda_j)$. Devido ao fato de não conhecermos a distribuição $F_{r,L}$, utiliza-se a técnica Monte Carlo para fazer o teste.

O procedimento do teste é o seguinte:

1. A partir das séries $\{\mathbf{X}_{1t}\}$ e $\{\mathbf{X}_{2t}\}, t=1,\ldots,T$, estime as matrizes de densidades espectrais $\hat{\mathbf{f}}_{1,1}(\lambda_j)$ e $\hat{\mathbf{f}}_{2,2}(\lambda_j)$, $j=1,\ldots,m$;
2. Calcule as estatísticas $T(\lambda_1),\ldots,T(\lambda_m)$, onde $T(\lambda_j)=traço\left[\hat{\mathbf{f}}_{1,1}(\lambda_j)\hat{\mathbf{f}}_{2,2}^{-1}(\lambda_j)\right]$, $j=1,\ldots,m$;
3. Gere independentemente as matrizes $M_1(\lambda_j)$ e $M_2(\lambda_j)$ $j=1,\ldots,m$ da distribuição $W_r^c(L,\mathbf{I})$;
4. Calcule as estatísticas do tipo Monte Carlo, $T^*(\lambda_1),\ldots,T^*(\lambda_m)$, onde $T^*(\lambda_j)=traço\left[M_1(\lambda_j)M_2^{-1}(\lambda_j)\right]$, $j=1,\ldots,m$; e
5. Aplique um teste de Kolmogorov-Smirnov para as duas amostras de traços $T(\lambda_1),\ldots,T(\lambda_m)$ e $T^*(\lambda_1),\ldots,T^*(\lambda_m)$ e rejeite a hipótese nula em nível de significância α se o nível descritivo do teste é menor que α. Daqui em diante este teste é denominado T_{KS};

Observação:

Os autores também sugerem usar a estatística $D(\lambda_j)=Det[M_1(\lambda_j)M_2^{-1}(\lambda_j)]$ onde *Det* indica o determinante de $M_1(\lambda_j)M_2^{-1}(\lambda_j)$. A vantagem de utilizar o traço é devido a maior simplicidade dos cálculos envolvidos.

3.2.2 Comparação de séries no caso de dependência (Carmona e Wang)

Neste caso $\mathbf{f}_{2,1}(\lambda_j)\not\equiv\mathbf{0}$ mas agora, por conveniência, supõe-se que a matriz de coerência é constante para toda freqüência λ. Assim, definimos a matriz $\mathbf{F}$ de dimensão rxr, tal que

$$\mathbf{F}\equiv\hat{\mathbf{f}}_{2,2}^{-1/2}(\lambda)\hat{\mathbf{f}}_{2,1}(\lambda)\hat{\mathbf{f}}_{1,1}^{-1/2}(\lambda) \text{ para toda freqüência } \lambda. \qquad (16)$$

Como foi especificado anteriormente, a matriz de densidade espectral dos dois processos é dada por

$$\mathbf{f}(\lambda)=\begin{bmatrix}\mathbf{f}_{1,1}(\lambda) & \vdots & \mathbf{f}_{1,2}(\lambda)\\ \cdots & \cdots & \cdots\\ \mathbf{f}_{2,1}(\lambda) & \vdots & \mathbf{f}_{2,2}(\lambda)\end{bmatrix}_{(2r)x(2r)},$$

satisfazendo

$$\mathbf{f}_{1,2}(\lambda)=\mathbf{f}_{2,1}^{*}(\lambda)=\overline{\mathbf{f}_{1,2}(\lambda)'} \quad \text{e} \quad \overline{\mathbf{f}_{1,2}(\lambda)'}\not\equiv\mathbf{0}.$$

Sob a hipótese H_0, dada por (13), e a suposição da existência da matriz $\mathbf{F}$, dada por (16), a correspondente matriz de coerência é

$$\hat{\mathbf{R}}(\lambda)=\begin{bmatrix}\mathbf{I}_r & \vdots & \hat{\mathbf{F}}^{*}\\ \cdots & \cdots & \cdots\\ \hat{\mathbf{F}} & \vdots & \mathbf{I}_r\end{bmatrix}_{(2r)x(2r)},$$

para todo λ, com $\mathbf{I}_r$ a matriz identidade de dimensão r.

Uma estimativa da matriz $\mathbf{F}$ que leva em consideração pequenas variações das matrizes de coerência nas diferentes freqüências é dada por

$$\hat{\mathbf{F}}=\frac{1}{m}\sum_{j=1}^{m}\hat{\mathbf{f}}_{2,2}^{-1}/2(\lambda_j)\hat{\mathbf{f}}_{2,1}(\lambda_j)\hat{\mathbf{f}}_{1,1}^{-1/2}(\lambda_j). \tag{17}$$

Substituindo $\mathbf{F}$ por $\hat{\mathbf{F}}$ podemos estimar a matriz de coerência

$$\hat{\mathbf{R}}(\lambda)=\begin{bmatrix}\mathbf{I}_r & \vdots & \hat{\mathbf{F}}^{*}\\ \cdots & \cdots & \cdots\\ \hat{\mathbf{F}} & \vdots & \mathbf{I}_r\end{bmatrix}_{(2r)x(2r)}. \tag{18}$$

Para testar a hipótese (13) para o caso de séries dependentes, Carmona e Wang(1996) propõem usar o mesmo teste T_{KS} usado no caso de independência, com a diferença que, no caso de dependência, a distribuição sob H_0 de M_1 e M_2 é desconhecida.

Os autores propõem o seguinte procedimento:

1. A partir das séries $\{\mathbf{X}_{1t}\}$ e $\{\mathbf{X}_{2t}\}$ estime as matrizes $2r$-dimensionais de densidades espectrais, $\hat{f} = (\lambda_j), j = 1,\ldots,m$;

2. Calcule as estatísticas $T(\lambda_1),\ldots,T(\lambda_m)$, onde $T(\lambda_j) = traço\left[\hat{\mathbf{f}}_{1,1}(\lambda_j \hat{\mathbf{f}}_{2,2}^{-1}(\lambda_j)\right]$, $j = 1,\ldots,m$;

3. Gere uma série $2r$-variada $\{\mathbf{Y}(t)\} = \{\mathbf{Y}_{1t}, \mathbf{Y}_{2t}\}$ de um processo Gaussiano com média zero e matriz de densidade espectral $\hat{\mathbf{R}}(\lambda)$,

$$\hat{\mathbf{R}}(\lambda) = \begin{bmatrix} \mathbf{I_r} & \vdots & \hat{\mathbf{F}}^* \\ \ldots & \ldots & \ldots \\ \hat{\mathbf{F}} & \vdots & \mathbf{I_r} \end{bmatrix}_{(2r)x(2r)} ;$$

4. A partir das séries $\{\mathbf{Y}_{1t}\}$ e $\{\mathbf{Y}_{2t}\}$ estime as matrizes $2r$-dimensionais de densidades espectrais, $\hat{\mathbf{g}}(\lambda_j)$, onde $\hat{\mathbf{g}}$ tem a forma

$$\hat{\mathbf{g}} = \begin{bmatrix} \hat{\mathbf{g}}_{1,1} & \vdots & \hat{\mathbf{g}}_{1,2} \\ \ldots & \ldots & \ldots \\ \hat{\mathbf{g}}_{2,1} & \vdots & \hat{\mathbf{g}}_{2,2} \end{bmatrix}_{(2r)x(2r)} ;$$

5. Calcule as estatísticas $T^*(\lambda_1),\ldots,T^*(\lambda_m)$, onde $T^*(\lambda_j) = traço\left[\hat{\mathbf{g}}_{1,1}(\lambda_j \hat{\mathbf{g}}_{2,2}^{-1}(\lambda_j)\right]$, $j = 1,\ldots,m$;

6. Calcule $T_{KS_{obs}}$, a estatística de Kolmogorov-Smirnov para as duas amostras de traços $T(\lambda_1),\ldots,T(\lambda_m)$ e $T^*(\lambda_1),\ldots,T^*(\lambda_m)$;

7. Gere B amostras independentes *bootstrap* $2r$-variadas, $\{\mathbf{Y}^{(b)}(t), t = 1,\ldots,T\}$ para $b = 1,\ldots,B$;

8. Para cada amostra $\{\mathbf{Y}^{(b)}(t), t = 1,\ldots,T\}$, estime as matrizes espectrais $2r$-dimensionais denominadas $\hat{\mathbf{g}}_b(\lambda_j)$, $b = 1,\ldots,B$, $j = 1,\ldots,m$;

9. Para cada amostra $\{\mathbf{Y}^{(b)}(t), t = 1, \ldots, T\}$, calcule as correspondentes estatísticas $T^b(\lambda_1), \ldots, T^b(\lambda_m)$ onde $T^b(\lambda_j) = traço\left[\hat{\mathbf{g}}_{1,1_b}(\lambda_j)\hat{\mathbf{g}}^{-1}_{2,2_b}(\lambda_j)\right]$; e

10. Calcule as B estatísticas T^b_{KS} de Kolmogorov-Smirnov para as duas amostras de traços, $T^*(\lambda_1), \ldots, T^*(\lambda_m)$ e $T^b(\lambda_1), \ldots, T^b(\lambda_m)$ e construa um histograma que será utilizado para calcular o nível descritivo do teste da hipótese (13), através da expressão

$$\alpha' = P(T^b_{KS} > T_{KS_{obs}}) = \frac{\#(T^b_{KS} > T_{KS_{obs}}}{B}.$$

Rejeite a hipótese nula em nível descritivo α se $\alpha' < \alpha$.

Um outro procedimento sugerido pelos autores é o seguinte:

1. Repita os passos 1 e 2 do procedimento anterior;
2. Gere uma série $2r$-variada $\{\mathbf{Y}(t)\} = \{\mathbf{Y}_{1t}, \mathbf{Y}_{2t}\}$ de um processo Gaussiano com média zero e matriz de densidade espectral $\hat{\mathbf{R}}(\lambda)$,

$$\hat{\mathbf{R}}(\lambda) = \begin{bmatrix} \mathbf{I_r} & \vdots & \hat{\mathbf{F}}^* \\ \cdots & \cdots & \cdots \\ \hat{\mathbf{F}} & \vdots & \mathbf{I_r} \end{bmatrix}_{(2r)x(2r)};$$

3. A partir das séries $\{\mathbf{Y}_{1t}\}$ e $\{\mathbf{Y}_{2t}\}$ estime as matrizes $2r$-dimensionais de densidades espectrais, $\hat{\mathbf{g}}(\lambda_j)$, onde $\hat{\mathbf{g}}$ tem a forma

$$\hat{\mathbf{g}} = \begin{bmatrix} \hat{\mathbf{g}}_{1,1} & \vdots & \hat{\mathbf{g}}_{1,2} \\ \cdots & \cdots & \cdots \\ \hat{\mathbf{g}}_{2,1} & \vdots & \hat{\mathbf{g}}_{2,2} \end{bmatrix}_{(2r)x(2r)};$$

4. Calcule as estatísticas $T^*(\lambda_1), \ldots, T^*(\lambda_m)$, onde $T^*(\lambda_j) = traço\left[\hat{\mathbf{g}}_{1,1}(\lambda_j)\hat{\mathbf{g}}^{-1}_{2,2}(\lambda_j)\right]$, $j = 1, \ldots, m$; e

5. Aplique um teste de Kolmogorov-Smirnov para as duas amostras de traços $T(\lambda_1),\ldots,T(\lambda_m)$ e $T^*(\lambda_1),\ldots,T^*(\lambda_m)$ e rejeite a hipótese H_0, dada por (13), em nível de significância α, se o nível descritivo do teste é menor que α.

Carmona e Wang (1996) apresentam algumas simulações para comparar o desempenho da estatística T_{KS} no caso de séries independentes e séries dependentes.

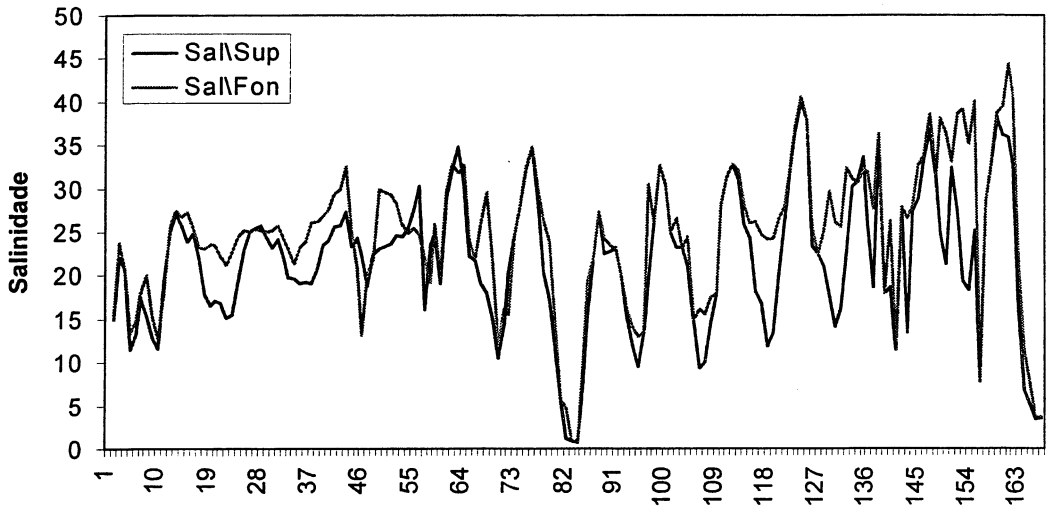
4. Aplicação

Para aplicar a teoria discutida na **seção 3**, selecionamos as variáveis temperatura e salinidade medidas no fundo e na superfície, na estação de monitoramento denominada Boca de Caño Grande - BCG -, da Ciénaga Grande de Santa Marta" (Figura 1).

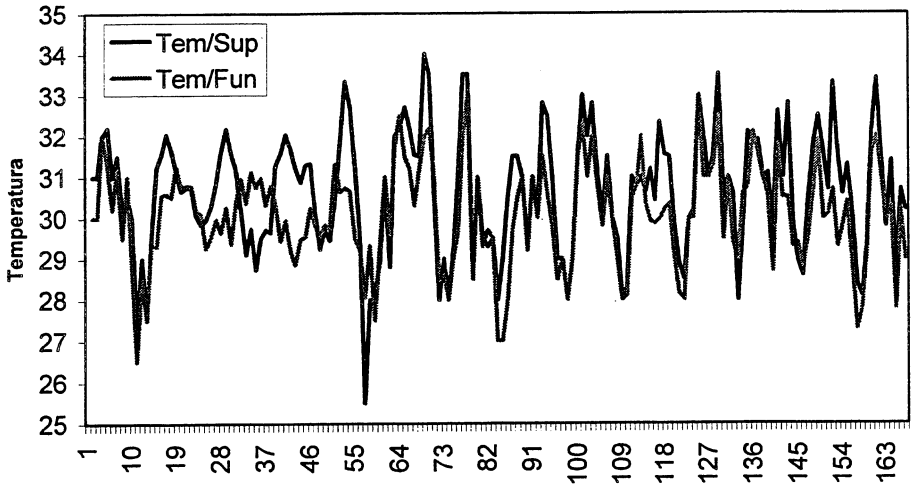
4.1 Aplicação dos testes univariados

A **Figura 2** apresenta as duas séries de salinidade e temperatura medidas na superfície e no fundo na estação BCG, registradas desde março de 1982 a dezembro de 1995, cada uma com um total de 166 observações.

Figura 2 - Séries de salinidade em BCG (a) e temperatura em BCG (b)



(a)



(b)

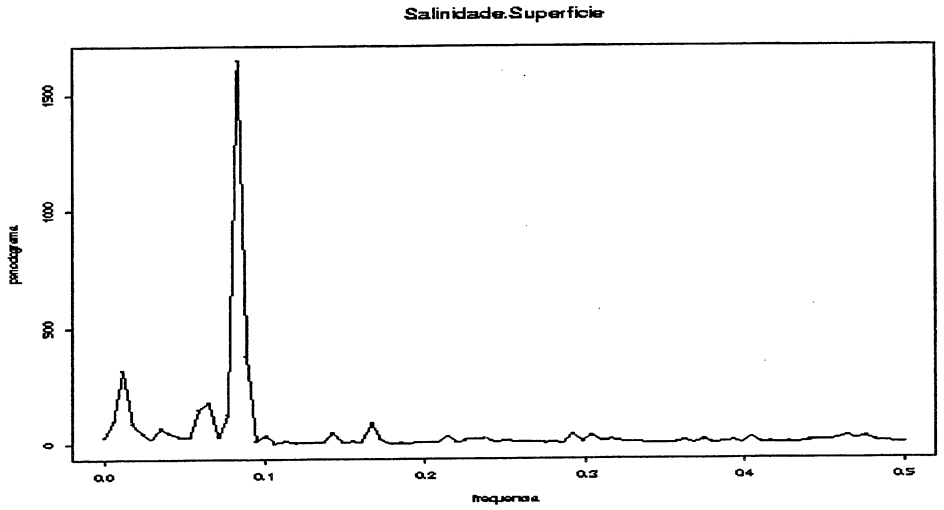
4.1.1 Análise para salinidade

Inicialmente, foi feito um teste de igualdade de médias para as séries pareadas e obtivemos o valor $Z_{obs} = 8{,}9362$, que por ser maior que 1,96 nos leva a rejeitar a hipótese de que as duas séries de salinidade têm a mesma média em nível de significância de 0,05.

Na **Figura 2**(a), pode-se observar que as séries são estáveis em média mas não em variância e que as séries apresentam um comportamento sazonal.

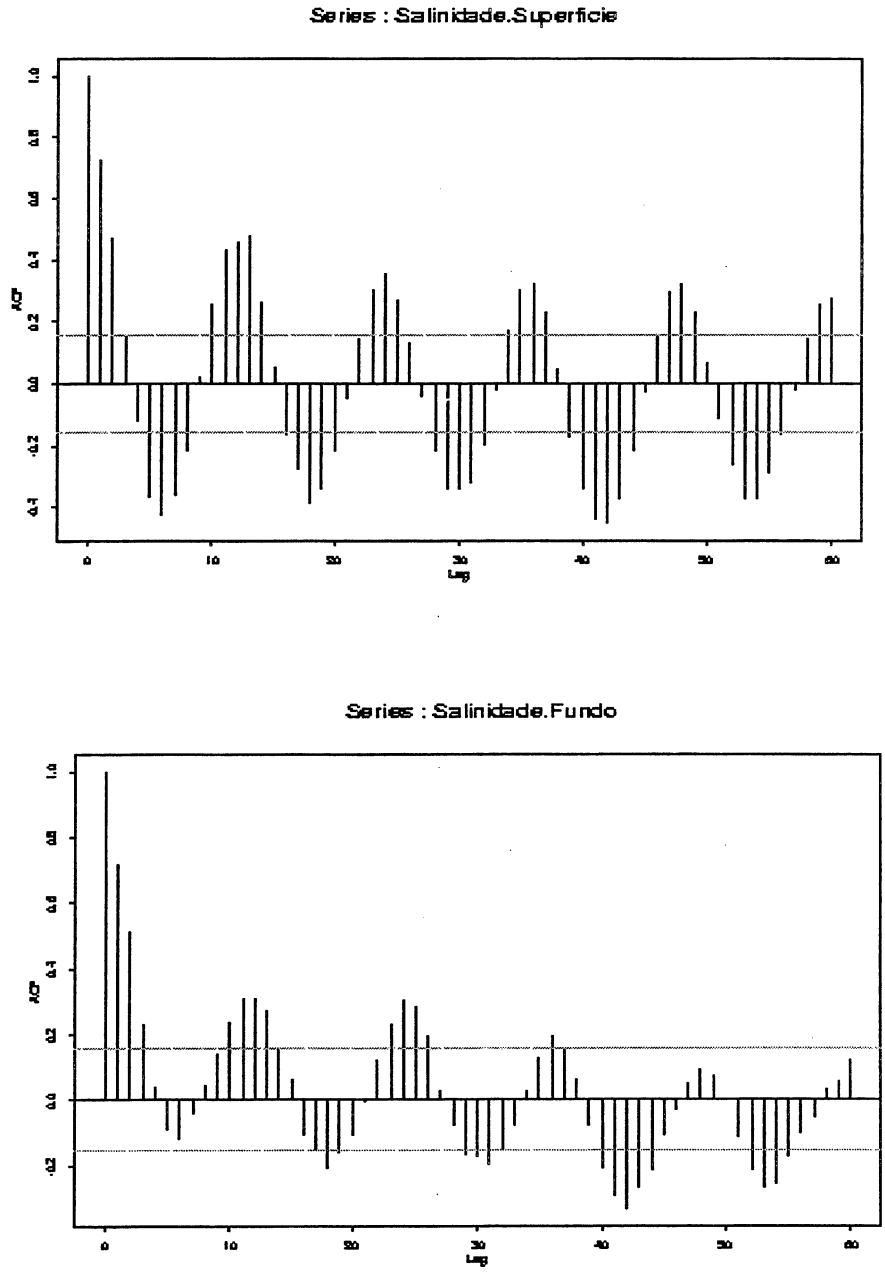
Na **Figura 3,** apresentamos os periodogramas das séries de salinidade na superfície e no fundo, respectivamente. O teste de periodicidade de *Fisher* estabelece que as duas séries apresentam um pico significante na freqüência 0,083333 o que confirma que as séries têm uma componente periódica de 12 meses.

Figura 3 - Periodogramas da salinidade na superfície e no fundo.



A **Figura 4** apresenta as funções de autocorrelação das duas séries de salinidade. Nestas duas figuras também observa-se um comportamento sazonal de ordem 12.

Figura 4 - FAC das séries de salinidade na superfície e no fundo



Para tornar as séries estacionárias foi necessário aplicar o logaritmo natural a cada uma das duas séries e eliminar a sazonalidade determinística através do operador $(1-B^{12})$. Feitas estas transformações, prosseguimos a comparação das salinidades usando agora as duas séries estacionárias.

Nas **Figuras 5** e **6** aparecem os periodogramas e as funções de autocorrelação das duas séries estacionárias de salinidade.

Figura 5 - Periodogramas das séries estacionárias de salinidade na superfície e fundo

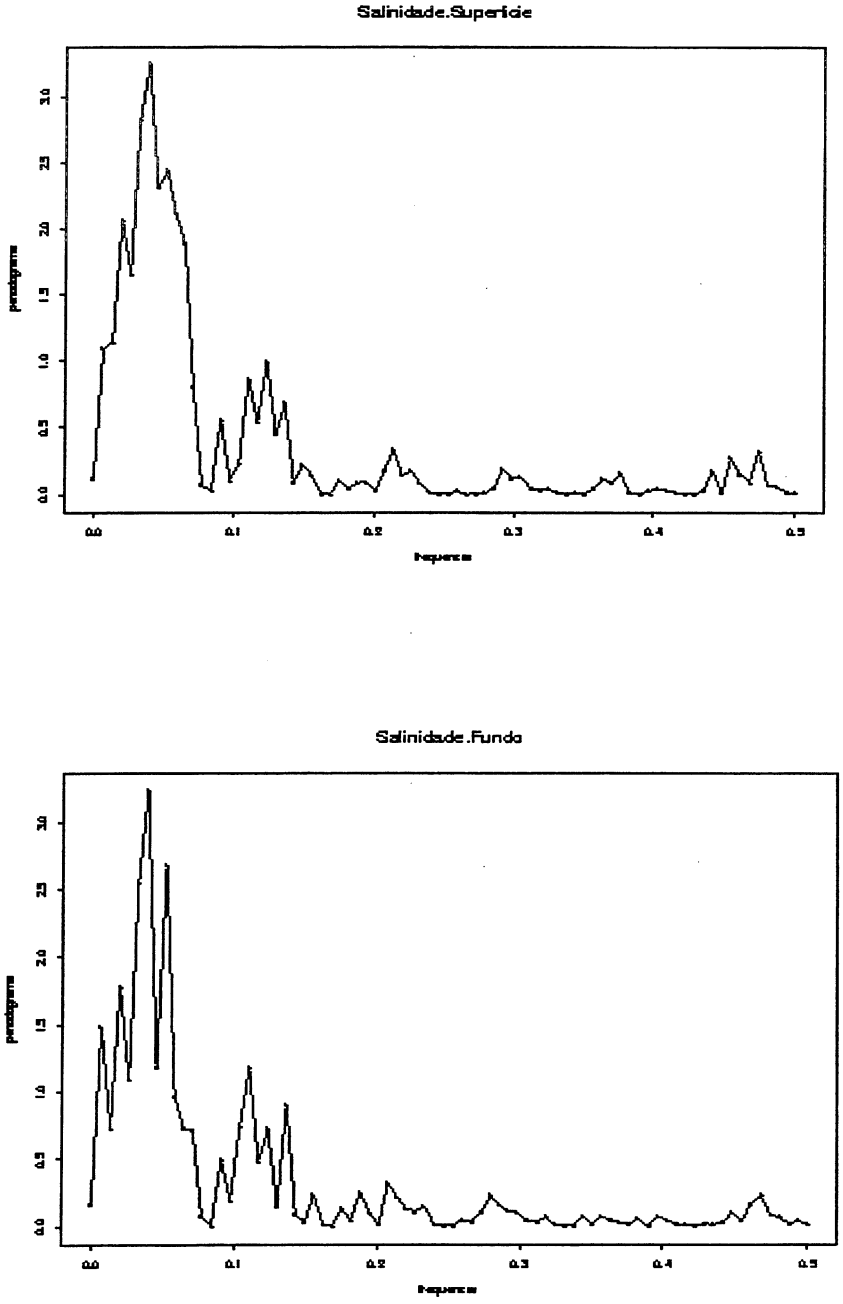
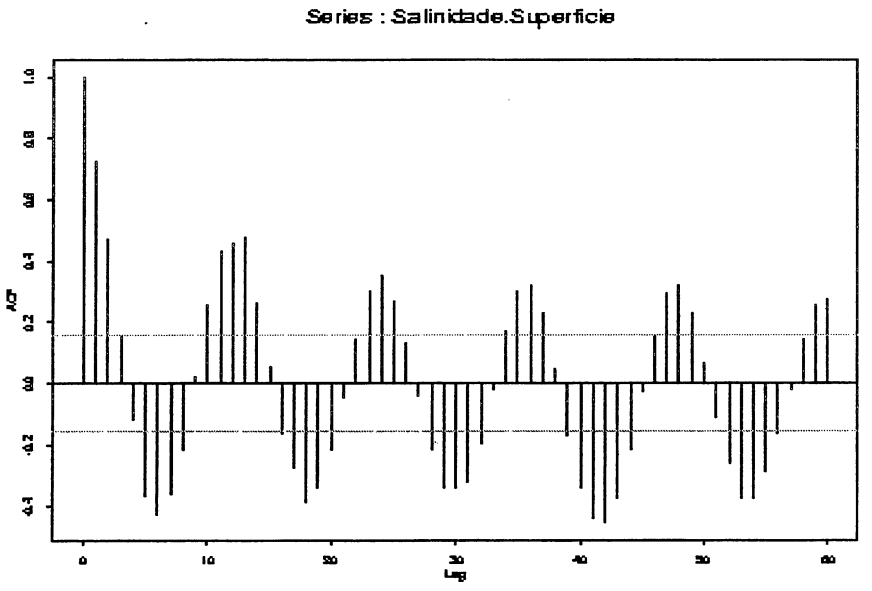
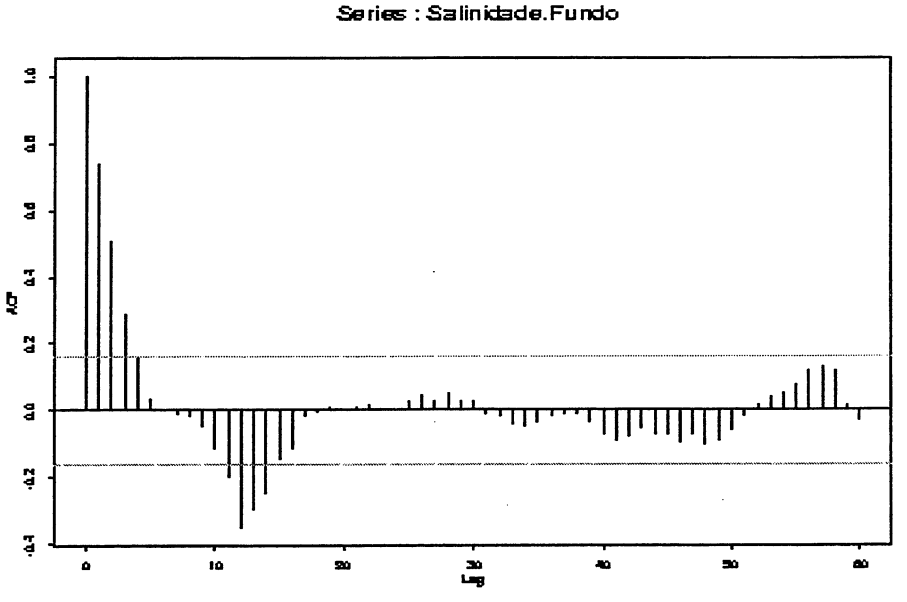


Figura 6 - FAC das séries estacionárias de salinidade na superfície e no fundo





A **Tabela 1** contém os resultados (níveis descritivos) obtidos na aplicação de cada teste.

Tabela 1 - Resultados dos testes univariados

Teste	Nível Descritivo
Quenouille	0,8788
Mélard e Roy	0,9988
Coates e Diggle (AS)	0,4165
Coates e Diggle (RV)	0,5234

Assim, em nível de significância de 0,05 os quatro testes estão aceitando a hipótese de que as duas séries de salinidade, na superfície e no fundo, na estação de monitoramento BCG são geradas por dois processos com a mesma estrutura de dependência de segunda ordem, porém, com médias diferentes.

Uma estimativa das médias dos processos é dada por $\overline{x}_{sup} = 21,7566$ e $\overline{x}_{fundo} = 24,7135$.

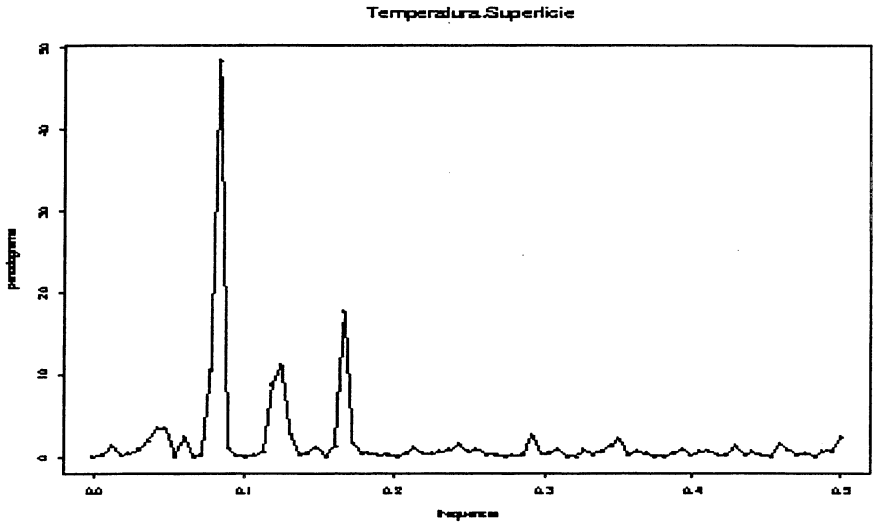
4.1.2 Análise para temperatura

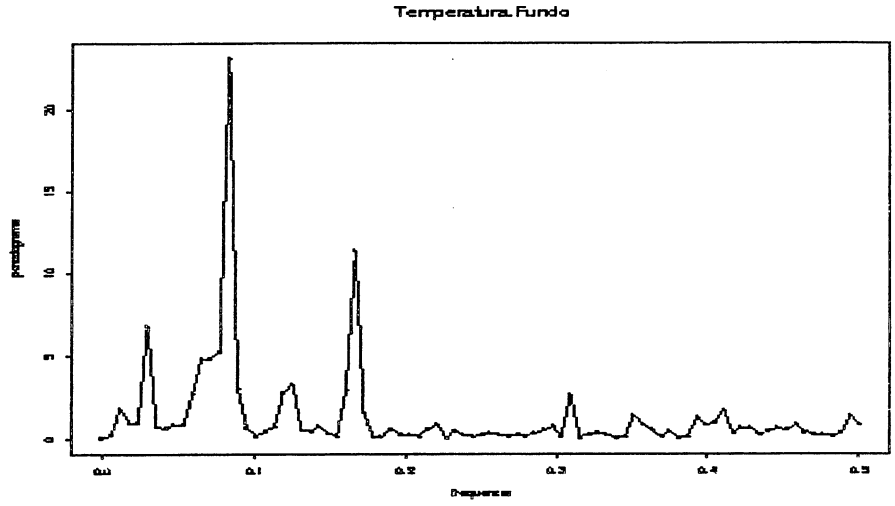
Também foi feito, inicialmente, um teste de igualdade de médias para as séries pareadas e obtivemos o valor $Z_{obs} = 8,49$, que por ser maior que 1,96, nos levar a rejeitar a hipótese de que as duas séries de temperatura têm a mesma média em nível de significância de 0,05.

Na **Figura 2**(b), observa-se que as séries são estacionárias em média e em variância e que apresentam um comportamento sazonal.

Na **Figura 7,** aparecem os periodogramas das séries de temperatura na superfície e no fundo, respectivamente. Um teste de *Fisher* estabelece que as duas séries apresentam um pico significante na freqüência 0,083333 o que também confirma que as séries têm uma componente periódica de 12 meses.

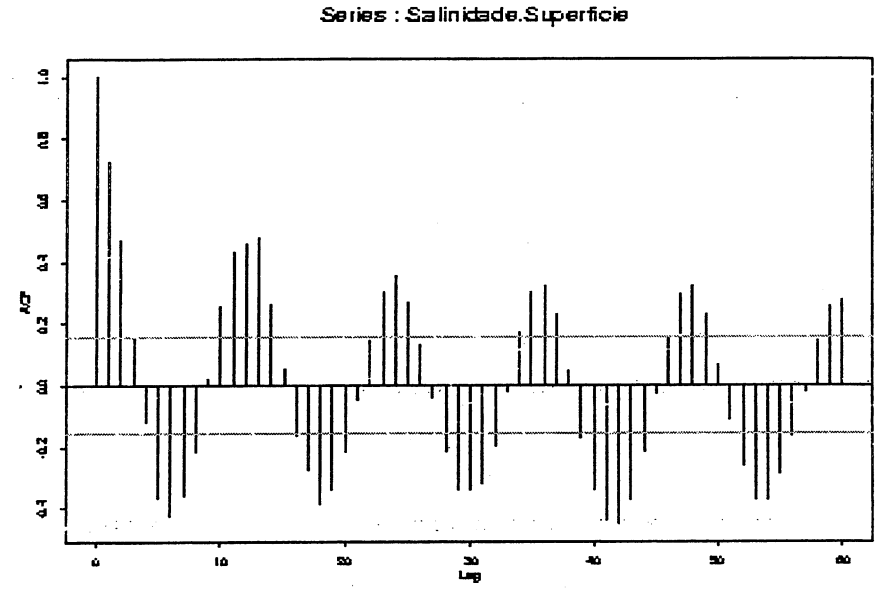
Figura 7 - Periodogramas da séries de temperatura na superfície e no fundo

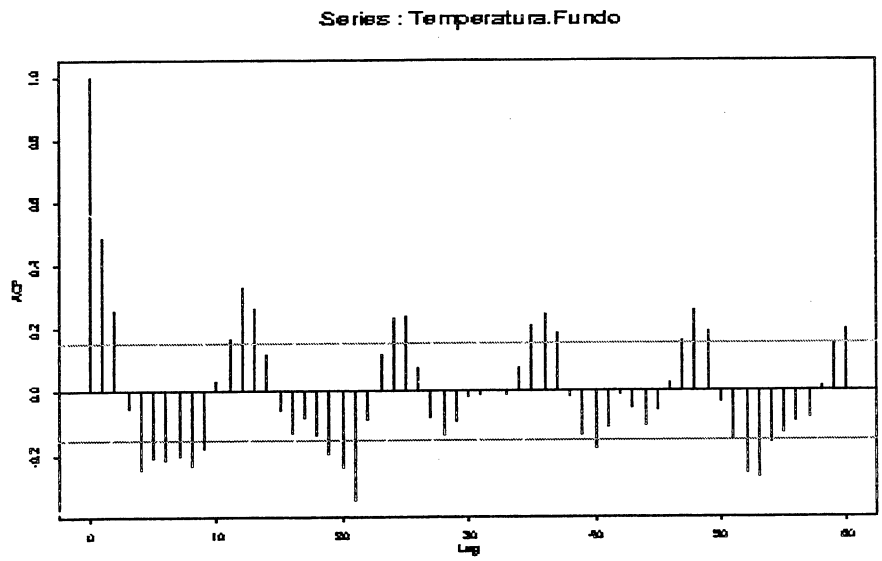




A **Figura 8** apresenta as funções de autocorrelação das duas séries de temperatura. Nestas duas figuras também observa-se um comportamento sazonal de período 12.

Figura 8 - FAC das séries de temperatura na superfície e no fundo

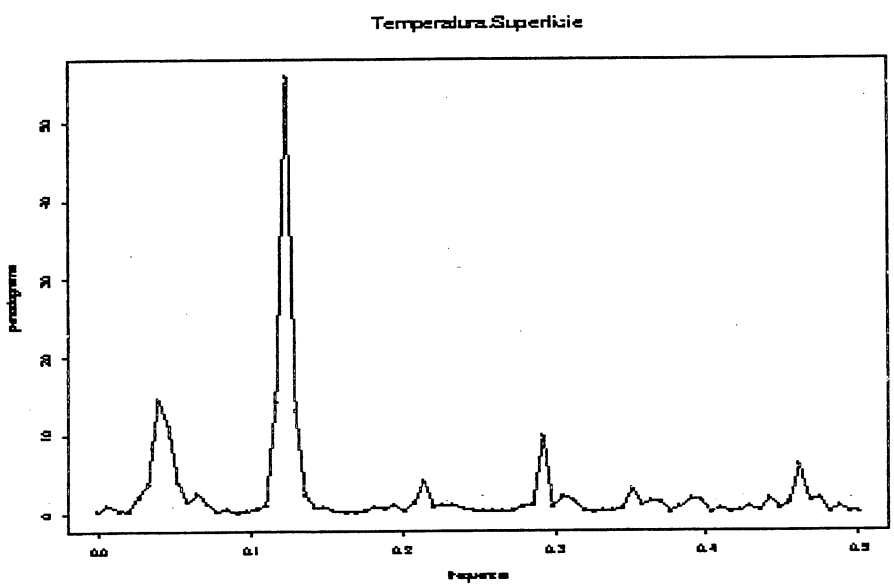




A componente sazonal foi eliminada aplicando o operador $(1 - B^{12})$. Prosseguimos a comparação das temperaturas utilizando as duas séries estacionárias.

As **Figuras 9** e **10** apresentam os periodogramas e as funções de autocorrelação das duas séries estacionárias de temperatura.

Figura 9 - Periodograma das diferenças sazonais da temperatura na superfície e no fundo



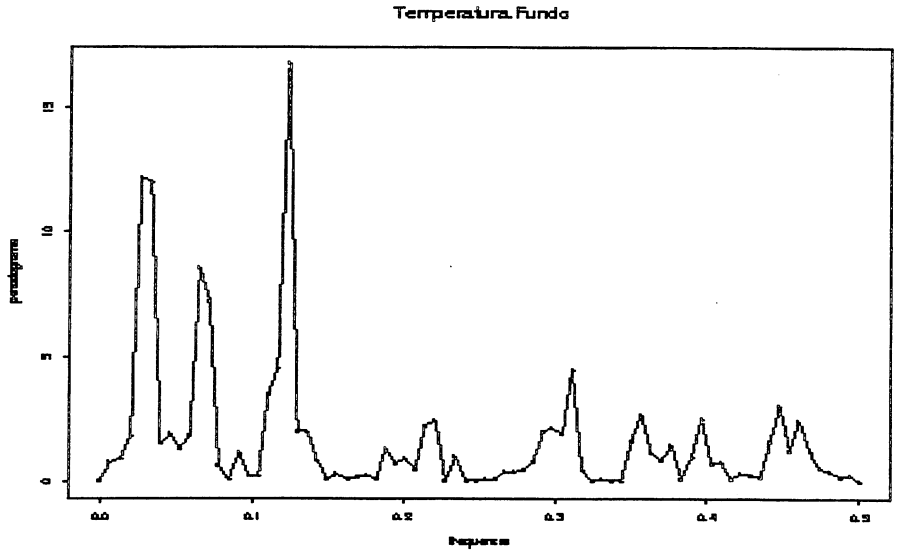
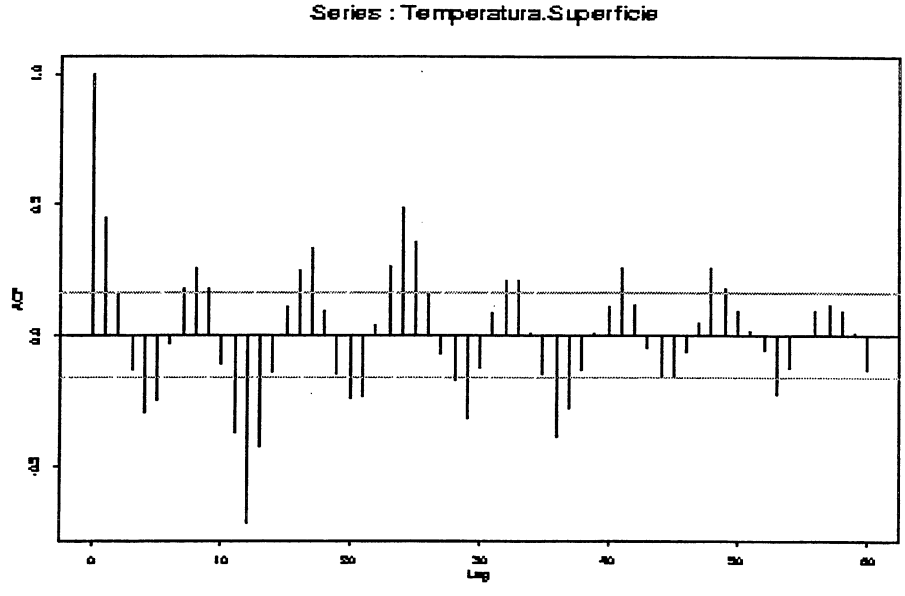
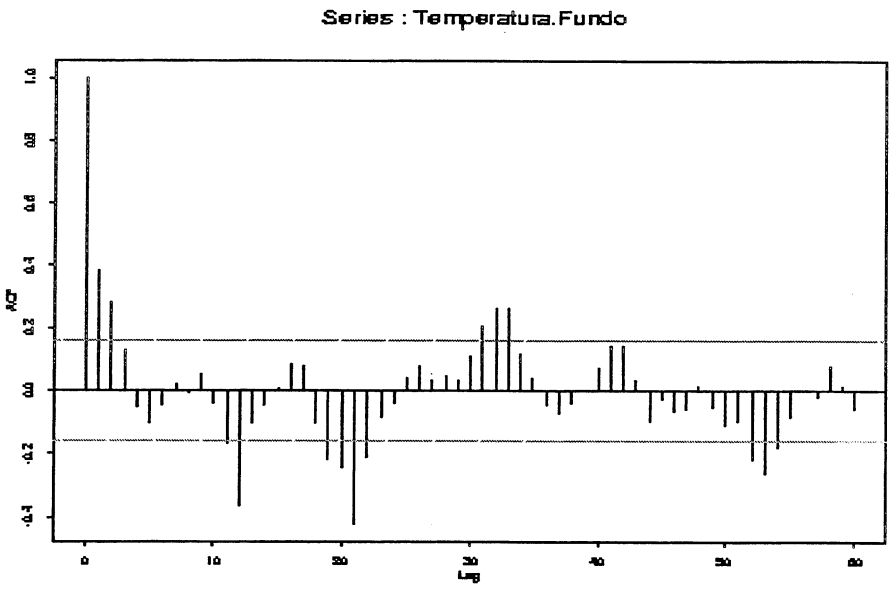


Figura 10 - FAC das diferenças sazonais da temperatura na superfície e no fundo





A **Tabela 2** contém os resultados (níveis descritivos) obtidos na aplicação de cada teste univariado.

Tabela 2 - Resultados dos testes univariados

Teste	Nível Descritivo
Quenouille	**0,7954**
Mélard e Roy	**0,0220**
Coates e Diggle (AS)	**0,2840**
Coates e Diggle (RV)	**0,3173**

Assim, em nível de significância de 0,05, os testes de Quenouille e de Coates e Diggle aceitam a hipótese de que as duas séries de temperatura, na superfície e no fundo, na estação de monitoramento BCG são geradas por processos com a mesma estrutura de dependência (porém com médias diferentes), mas o teste de Mélard e Roy rejeita esta hipótese; no entanto, a hipótese é aceita em nível de significância de 0,01. As estimativas das médias da temperatura são $\bar{x}_{\text{sup}} = 30.6293$ e $\bar{x}_{fundo} = 30.0654$.

4.2 Aplicação dos testes multivariados

Consideraremos, aqui, a comparação das séries bivariadas $\{X_1^{(1)}(t), X_2^{(1)}(t)\}$ e $\{X_1^{(2)}(t), X_2^{(2)}(t)\}$ onde $\{X_1^{(1)}(t)\}$ e $\{X_2^{(1)}(t)\}$ representam a salinidade e a temperatura na superfície e $\{X_1^{(2)}(t)\}$ e $\{X_2^{(2)}(t)\}$ representam a salinidade e a temperatura no fundo. A comparação é feita utilizando os testes de

Carmona e Wang com o objetivo de verificar se a série bivariada $\{X_1^{(1)}(t)\}$ e $\{X_2^{(1)}(t)\}$ tem o mesmo comportamento que $\{X_1^{(2)}(t), X_2^{(2)}(t)\}$.

Inicialmente, fazemos um teste baseado em comparações pareadas (Johnson e Wichern (1998)) para verificar se os dois vetores de médias da salinidade e temperatura, na superfície e no fundo, são ou não iguais.

O procedimento é descrito a seguir:

1. Considerar as séries bivariadas na superfície e no fundo, respectivamente, na forma vetorial dada por

$$\begin{bmatrix} \{X_1^{(1)}(t)\} \\ \{X_2^{(1)}(t)\} \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} \{X_1^{(2)}(t)\} \\ \{X_2^{(2)}(t)\} \end{bmatrix};$$

2. Construir o vetor das diferenças $\mathbf{D}$ dado por

$$\mathbf{D} = \begin{bmatrix} \{X_1^{(1)}(t) - \{X_1^{(2)}(t) \\ \{X_2^{(1)}(t) - \{X_2^{(2)}(t) \end{bmatrix};$$

Assume-se que este vetor tem distribuição $\mathbf{N_p}(\mu_\mathbf{D}, \mathbf{S_D})$, com $p = 2$; e

3. Testar a hipótese $\mathbf{H}_0 : \mu_\mathbf{D} = \mathbf{0}$ contra a alternativa $\mathbf{H}_A : \mu_\mathbf{D} \neq \mathbf{0}$. Rejeita-se a hipótese $\mathbf{H}_0$ em nível de significância α, se o valor observado

$$T^2 = n\overline{\mathbf{D}}'\mathbf{S}_\mathbf{D}^{-1}\overline{\mathbf{D}} > \frac{(n-1)}{(n-p)} F_{p,n-p}(\alpha),$$

onde n representa o tamanho das séries, $\overline{\mathbf{D}}$ o vetor das médias das diferenças, $\mathbf{S_D}$ a matriz de variâncias e covariâncias das diferenças e $F_{p,n-p}(\alpha)$ o $100\,\alpha$-ésimo percentil da distribuição **F** com p graus de liberdade no numerador e $n - p$ graus de liberdade no denominador.

A rejeição de $\mathbf{H}_0$ implica que os vetores média das duas séries bivariadas são diferentes.

Esse procedimento foi aplicado a um conjunto de 166 observações de cada uma das séries fornecendo os resultados:

$$\overline{\mathbf{D}} = \begin{bmatrix} -2{,}95691 \\ 0{,}56382 \end{bmatrix}, \mathbf{S_D} = \begin{bmatrix} 18{,}1749 & -0{,}5258 \\ -0{,}5258 & 0{,}8464 \end{bmatrix},$$

$$T^2 = 166[-2,95691 \quad 0,56382]\begin{bmatrix} 18,1749 & -0,5258 \\ -0,5258 & 0,8464 \end{bmatrix}\begin{bmatrix} -2,95691 \\ 0,56382 \end{bmatrix} = 125,539$$

e

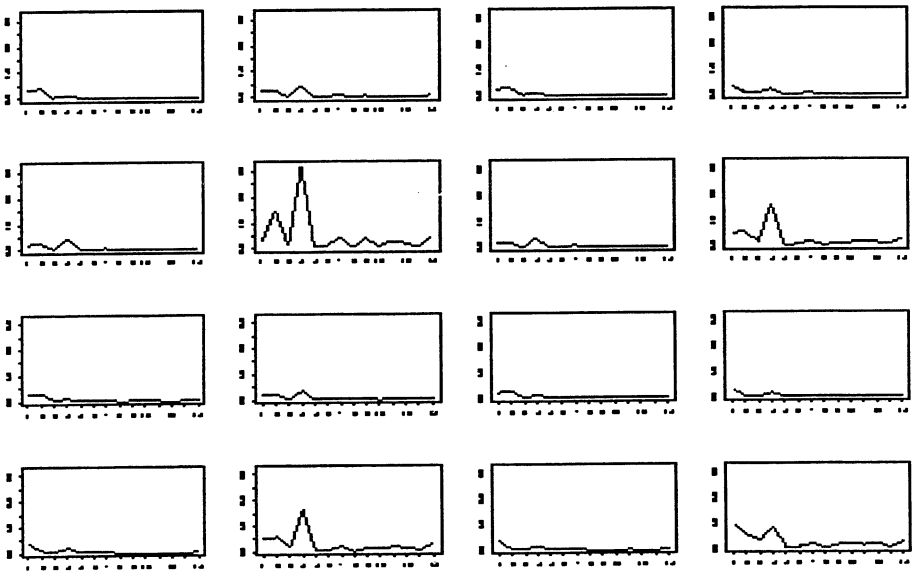
$$\frac{165x2}{164} F_{2,164}(0,95) = 6,139 .$$

Como T^2 = 125,5 > 6,139, rejeitamos a hipótese H_0 e concluímos que os vetores média das duas séries bivariadas em diferentes profundidades de água são diferentes.

Prosseguimos a análise com o objetivo de verificar se a estrutura de dependência das séries bivariadas é a mesma.

A **Figura 11** representa os módulos das entradas da estimativa da matriz de densidade espectral suavizada de tamanho 4x4, $\mathbf{f}_{\mathbf{x}}(\lambda_j)$, das séries bivariadas $\{X_1^{(1)}(t), X_1^{(2)}(t)\}$ e $\{X_1^{(2)}(t), X_2^{(2)}(t)\}$. Na suavização desta matriz, usamos L=5 valores de periodogramas.

Figura 11 - Módulo das entradas da matriz de densidade espectral estimada



4.2.1 Testes de Carmona e Wang

a) Comparação das séries no caso de independência

Usando a estatística traço, dada por (5.14) e assumindo independência entre as séries $\{X_1^{(1)}(t), X_2^{(1)}(t)\}$ e $\{X_1^{(2)}(t), X_2^{(2)}(t)\}$, obtemos os seguintes resultados:

1. O vetor de estatísticas traço, $\mathbf{T}$, de tamanho (1x14) e que corresponde aos traços dos produtos das matrizes de densidade espectral estimada das séries bivariadas $\{X_1^{(1)}(t), X_2^{(1)}(t)\}$ e $\{X_1^{(2)}(t), X_2^{(2)}(t)\}$ é dado por

T= [1,8178, 3,9121, 1,4572, 4,5471, 2,0761, 2,3065, 4,20972, 5,8119, 3,0606, 2,9770, 4,4579, 2,9175, 2,6182, 5,5006];

2. O vetor de estatísticas traço, **T***, de tamanho (1x14), formado pelos traços do tipo Monte Carlo é dado por **T*** = [3,6611, 4,3280, 3,3670, 2,0479, 1,8812, 2,1415, 2,1605, 1,8679, 3,1620, 2,4345, 3,9932, 4,0515, 2,8674, 2,7795]; e

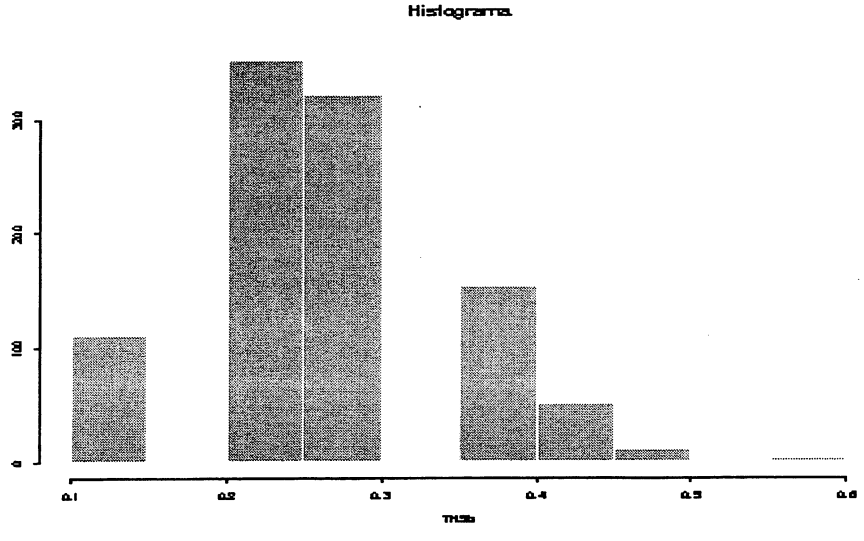
3. O teste de Kolmogorov-Smirnov aplicado às duas amostras **T** e **T***, forneceu um nível descritivo de 0,54074, que nos leva à aceitação da hipótese de igualdade das matrizes de densidade espectrais das séries bivariadas.

b) Comparação das séries no caso de dependência

Neste caso aplicamos os dois procedimentos sugeridos.

1. Quando assumimos dependência entre as séries $\{X_1^{(1)}(t), X_2^{(1)}(t)\}$ e $\{X_1^{(2)}(t), X_2^{(2)}(t)\}$ e utilizando o primeiro procedimento descrito na seção 3.5.2, geramos 1000 réplicas *bootstrap* que forneceram as estatísticas $T_{KS}^b, b = 1, \ldots, 1000$, cujo histograma está representado na **Figura 12**. O valor observado $T_{KSobs}^b = 0{,}42857$ junto com o histograma fornecem um nível descritivo de 0,064 que nos leva à aceitação da hipótese de igualdade das matrizes de densidade espectrais das séries bivariadas.

Figura 12 - Histograma das estatísticas *bootstrap*



2. Utilizando o segundo procedimento descrito na seção 3.5.2, obtemos o vetor de traços, do tipo Monte Carlo, dado por **T*** = [2,878252, 2,820556, 2,697550, 3,401028, 1,413149, 1,511414, 1,179629, 1,891015, 4,085084, 2,675872, 1,771545, 2,644145, 4,395269, 2,038589].
   O teste de Kolmogorov-Smirnov aplicado às duas amostras **T** e **T***, forneceu um nível descritivo de 0,1106, que também nos leva à aceitação da hipótese de igualdade das matrizes de densidade espectrais das séries bivariadas.

Em resumo, a suposição de independência não interfere no resultado da aplicação do teste de Carmona e Wang. Assim, em nível de significância de 5\%, podemos concluir que apesar das séries bivariadas de salinidade e temperatura apresentarem médias diferentes na superfície e fundo da lagoa, elas têm a mesma estrutura de dependência de segunda ordem.

5. Conclusões

Em nossa análise dos dados de temperatura e salinidade verificamos que:

1. Foi detectado, em nível de 0,05, uma diferença entre as médias de salinidade na superfície e no fundo da lagoa. A ocorrência de uma salinidade média maior no fundo da lagoa já era esperada. Também foi detectada uma diferença entre as médias de temperatura na superfície e no fundo da lagoa, neste caso, a temperatura média na superfície é maior do que no fundo, o que também era esperado;
2. Os testes univariados, quando utilizados para verificar se salinidade na superfície e fundo da lagoa têm a mesma estrutura de dependência, forneceram o mesmo resultado em nível de significância de 5\%, isto é, todos aceitaram que as duas séries, corrigidas pela média, têm o mesmo comportamento. O mesmo acontece às duas séries de temperatura, se considerarmos em nível de significância de 0,01;
3. Um teste bivariado de comparação de médias, também, constatou, em nível de significância de 0,05, que as séries na superfície e no fundo têm médias diferentes;
4. Como mencionado na seção anterior, quando utilizamos técnicas multivariadas para verificar se a série bivariada de salinidade e temperatura na superfície da água têm a mesma estrutura de dependência que a série salinidade e temperatura no fundo da água, verificamos que a suposição de dependência, ou não, entre as séries bivariadas, não altera o resultado de que as duas séries, corrigidas pela média, têm o mesmo comportamento de segunda ordem; e
5. Como conclusão final podemos sugerir que se tome uma única medida (na superfície) das séries de salinidade e temperatura e se for de interesse obter uma amostra de cada uma delas no fundo, basta fazer a respectiva correção da média.

Para maiores detalhes, veja Salcedo (1999).

Referências bibliográficas

ANDERSON, T. W. *The Statistical Analysis of Time Series.* John Wiley, New York, 1971.

BILLINGSLEY, P. *Convergence of Probability Measures.* John Wiley, New York, 1968.

BRILLINGER, D. R. *Time Series: Data Analysis and Theory*, Holden -Day, San Francisco, 1981.

CARMONA, R. A. AND WANG, A. Comparison Tests for the Spectra of Dependent Multivariate Time Series. *Stochastic Modelling in Physical Oceanography*, 39 (1):69-88, 1996.

COATES, D. S. AND DIGGLE, P. J. Tests for Comparing two Estimated Spectral Densities. *Journal of Time Series Analysis*, 7 (1):7 -20, 1986.

LOÉVE, M. *Probability Theory II.* Springer-Verlag, New York, 1978.

MÉLARD, G. AND ROY, R. Sur un test d'egalité des autocovariances de deux séries chronologiques. *La Revue Canadienne de Statistique*, 12 (4):333--342, 1984.

PRIESTLEY, M. B. *Spectral Analysis and Time Series*, volume 1,2. Academic Press, New York, 1981.

QUENOUILLE, M. H. The Comparison of Correlations in Time-Series. *Journal Royal Statist. Soc. Ser.* B, 20(1):158-164, 1958.

ROBINSON, P. M. Estimating Variances and Covariances of sample autocorrelations and autocovariances. *Australian Journal of Statistics*, (19):236-240, 1977.

SALCEDO, G. E. *Métodos de Comparação de Séries Temporais.* Dissertação, Departamento de Estatística, USP, 1999.

ABSTRACT

In time series analysis sometime is interesting to verify if two series or two parts of the same series are realizations of the same stationary process. Considering the class of the second order stationary process, statistical test procedures are presented to compare the autocovariance, the autocorrelation and the spectral functions of univariate and multivariate time series. An application with real series is given.

Análise de intervenção em séries temporais: aplicações em transporte urbano

Adriano Ferreti Borgatto*

Thelma Sáfadi*

RESUMO

Este trabalho analisa o comportamento do transporte urbano na cidade de São Paulo, utilizando séries temporais, sendo incluído na análise efeitos de intervenção com o objetivo de gerar previsão mais precisa. As séries utilizadas foram o número médio de passageiros por ônibus, o número de assaltos nos ônibus e o número de acidentes com os ônibus. As identificações dos modelos de Box e Jenkins foram feitas através da função de autocorrelação e função de autocorrelação parcial, acrescentando-se parâmetros para os fenômenos independentes da série, denominados intervenção. A verificação da adequabilidade dos modelos foram analisadas através do teste de Box e Pierce, critério de Akaike, quadrado médio residual e do erro quadrático médio estimado através da previsão.

Palavras-chave: análise de intervenção, modelos SARIMA e transporte urbano.

1. Introdução

Em estudos relacionados ao transporte urbano, sendo o principal interesse analisar o desenvolvimento histórico das empresas responsáveis pelo transporte, é apropriado utilizar técnicas de séries temporais. Existem fatores que influenciam a mudança de comportamento de uma série temporal, sendo que para analisá-los é necessário acrescentar técnicas de análise de intervenção. O seu comportamento consiste em uma inclinação ou mudança de nível decorrente na série num determinado instante do tempo.

A cidade de São Paulo possui uma população em torno de 10 milhões de habitantes, e a região metropolitana, contendo 38 municípios, acomoda mais de 7 milhões de habitantes. Cerca de 55% das viagens diárias motorizadas na região metropolitana são feitas por transportes coletivos. Na cidade os ônibus atendem cerca de 70% das viagens de transporte coletivo.

Todas as linhas de ônibus são operadas por empresas privadas, sob a supervisão da São Paulo Transporte S.A. – SPTrans, empresa municipal de planejamento e gerenciamento do transporte coletivo.

Atualmente 54 empresas operam, com uma frota aproximada de 10 mil ônibus, 800 linhas, utilizadas por quase 4 milhões de passageiros/dia.

* Endereço para correspondência: Deptº de Ciências Exatas da Univ. Federal de Lavras. Lavras - MG. CEP 37200-000 - E-mail: borgatto@zipmail.com.br , safadi@ufla.br

As informações obtidas, como referência citadas no texto, foram fornecidas pelo sistema de transporte coletivo do Município de São Paulo e na cronologia do transporte público de São Paulo, na página http://www.sptrans.com.br.

Os principais eventos que ocorreram com o transporte coletivo em São Paulo no período de 1983 a 1999, são:

a) 1983 – Início da integração ônibus-ferrovia, em que o transporte público volta a ser alvo da atenção da administração municipal;
b) 1984 – Entra em funcionamento a primeira linha operadora a gás metano;
c) 1991 – Em Janeiro é assinada uma lei que determina a total substituição, no prazo de dez anos, da frota de ônibus urbanos movidos a diesel por ônibus a gás natural.
   Em junho, tem início a operação da primeira linha com entradas pela porta dianteira.
   Em julho surge uma lei que municipaliza o transporte coletivo dos ônibus, determina licitações para cobrir 42 lotes de áreas de operação, substituindo as 23 áreas exclusivas de operação de ônibus;
d) 1993 – A nova administração assume e encontra a CMTC e o sistema municipalizado em condição precária. O número de passageiros volta a ter um peso significativo na remuneração das empresas contratadas;
e) 1994 – O Sistema de Transporte Coletivo por Ônibus passa a ser operado por 47empresas privadas; e
f) 1995 – A partir do segundo semestre do ano, é adotada a sistemática de coleta automática de dados operacionais do sistema de transporte urbano por ônibus, a Fiscalização Eletrônica; é implementado o Plano de Recuperação do Desempenho Operacional dos Corredores de Transporte Coletivo.

Estes eventos podem influenciar o comportamento de séries de transporte urbano ao longo do tempo: a primeira série a ser considerada é o "número de passageiros por ônibus"; a segunda constata o "número de assaltos nos ônibus"; e a terceira registra o "número de acidentes com os ônibus".

Existem na literatura aplicações utilizando as técnicas de análise de intervenção. A utilidade desta análise pode-se manifestar em várias áreas: ciências sociais, economia, sociologia, meio ambiente, entre outras.

Pino (1980) aplica análise de intervenção para avaliar o impacto de variações climáticas e medidas de política agrícola sobre séries de produção e produtividade de leite no Estado de São Paulo e sobre séries de produção e preço de café no Brasil.

Bhattacharrya e Layton (1979) utilizaram a análise de intervenção para avaliar o efeito da legislação do uso de cinto de segurança em automóveis, sobre o número de mortes por acidentes rodoviários, no Estado de Queensland (Austrália).

O objetivo deste trabalho é gerar previsões para as séries de transporte urbano, aplicando as metodologias de análise de intervenção e séries temporais a séries de interesse prático. As análises serão feitas utilizando os softwares SAS©(1999) e STATISTICA©(1995).

2. Análise de intervenção

Neste trabalho irão ser aplicadas técnicas de séries temporais para ajustar modelos às séries de transporte urbano. Os modelos ajustados para as séries foram propostos utilizando as técnicas de Box e Jenkins (1970).

Estas técnicas utilizadas por Box e Jenkins, baseiam-se nas idéias propostas por Wold (1938), e consistem em identificação do modelo, estimação dos parâmetros e verificação da validade do modelo. Essa sofisticada técnica fornece previsões com embasamento probabilístico e com o menor erro de estimativa, baseada na construção de modelos estimados de forma iterativa. Contudo, a estimação com tais modelos tem algumas restrições, tais como estacionaridade e tamanho da amostra, que deve ser de no mínimo 50 observações. Entretanto, antes da aplicação desta técnica é necessário verificar o efeito das componentes de tendência e sazonalidade nas séries.

A tendência se caracteriza como uma inclinação que ocorre na série e a sazonalidade é uma periodicidade que ocorre num intervalo máximo de 12 meses.

O termo intervenção foi introduzido por Glass (1972), baseado em Box e Tiao (1965), que já utilizavam esta metodologia, mas não com o termo intervenção.

Existe na literatura uma extensa bibliografia que trata da análise de intervenção, tais como Pankratz (1991), Bowerman (1993), Box et al. (1994), entre outras.

O modelo de intervenção que será utilizado neste trabalho tem a forma geral

$$z_t = \sum_{i=1}^{k} v_i(B)\, x_{i,t} + \eta_t \qquad (1)$$

sendo

- z_t é a variável resposta;
- B é o operador de retardo, tal que $Bx_t = x_{t-1}$ e $B^m x_t = x_{t-m}$;
- $v_i(B)$ é a função de transferência para a i-ésima intervenção;
- $x_{i,t}$ é variável indicadora de intervenção; e
- η_t é o resíduo representado através dos modelos de Box e Jenkins,

em que cada $v_i(B)$ é da forma $\frac{w_i(B)}{\delta_i(B)}$, sendo $w_i(B)$ os parâmetros que determinam o efeito de intervenção e $\delta_i(B)$ o efeito da intervenção até atingir um novo nível na série temporal.

Na análise de intervenção, é necessário indicar seus efeitos através de variáveis binárias, que estão representadas no modelo por $x_{i,t}$.

Existem várias formas de uma intervenção afetar a série temporal. As observações podem ser afetadas por uma mudança de nível ou por uma inclinação da série. A inclinação ocorre quando a intervenção é complexa e seu efeito é gradativo.

A mudança de nível ou da inclinação da série temporal pode ocorrer de forma abrupta ou gradual, afetando a série no instante em que ocorre a intervenção, ou depois de algum período de iniciada a intervenção (defasagem), e seu efeito pode ser temporário ou permanente.

Após a ocorrência da intervenção pode haver uma mudança na variabilidade da série, bem como um efeito de evolução, onde a série cai inicialmente e retoma o crescimento, até atingir um novo nível. Este efeito aparece, por exemplo, em estudos de sobrevivência de uma espécie após uma mutação adaptativa (Glass, Wilsson e Gottman, 1975).

O efeito de intervenção é determinado através da estrutura da função de transferência. Se já for conhecida a forma da função de transferência no modelo e seus parâmetros estimados, conhece-se o tipo de efeito de intervenção.

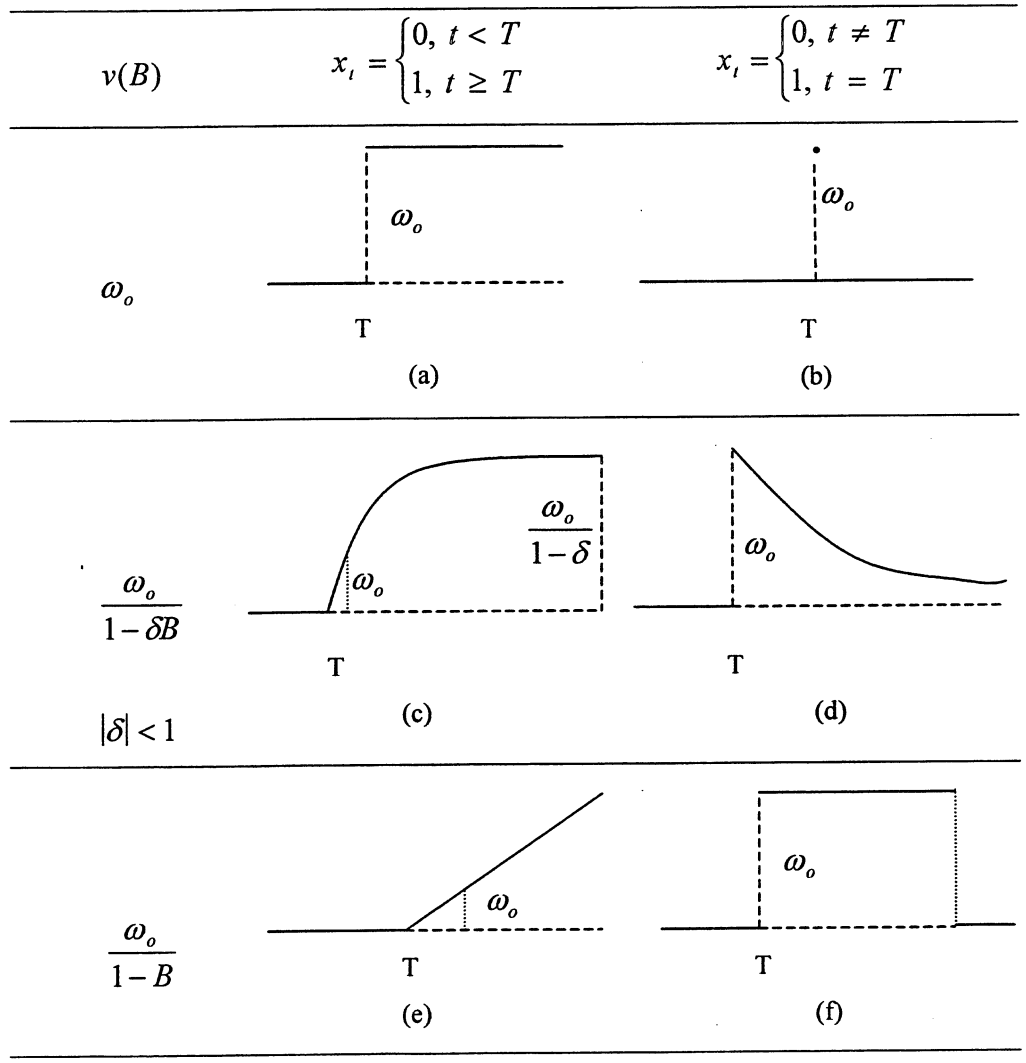
No caso de uma série temporal, com fenômeno conhecido, identifica-se o que está ocasionando o efeito de intervenção, facilitando a identificação da estrutura da função de transferência.

Na **Figura 1** estão apresentados os tipos mais comuns de função de transferência e os tipos de efeito de intervenção.

Para cada tipo de função de transferência, ajusta-se um modelo do tipo $v_i(B)$.

Para verificar o efeito de intervenção na série temporal, aplicam-se alguns testes estatísticos paramétricos ou não-paramétricos. Neste trabalho o efeito de intervenção será testado utilizando o teste t de student (Box e Tiao, 1965).

Figura 1 - Estrutura da função de transferência



3. Resultados

As séries analisadas neste trabalho são o "número de passageiros por ônibus", o "número de assaltos nos ônibus" e o "número de acidentes com os ônibus". *A priori*, considerou-se a época da privatização da CMTC em 1993 como possível efeito de intervenção nas três séries.

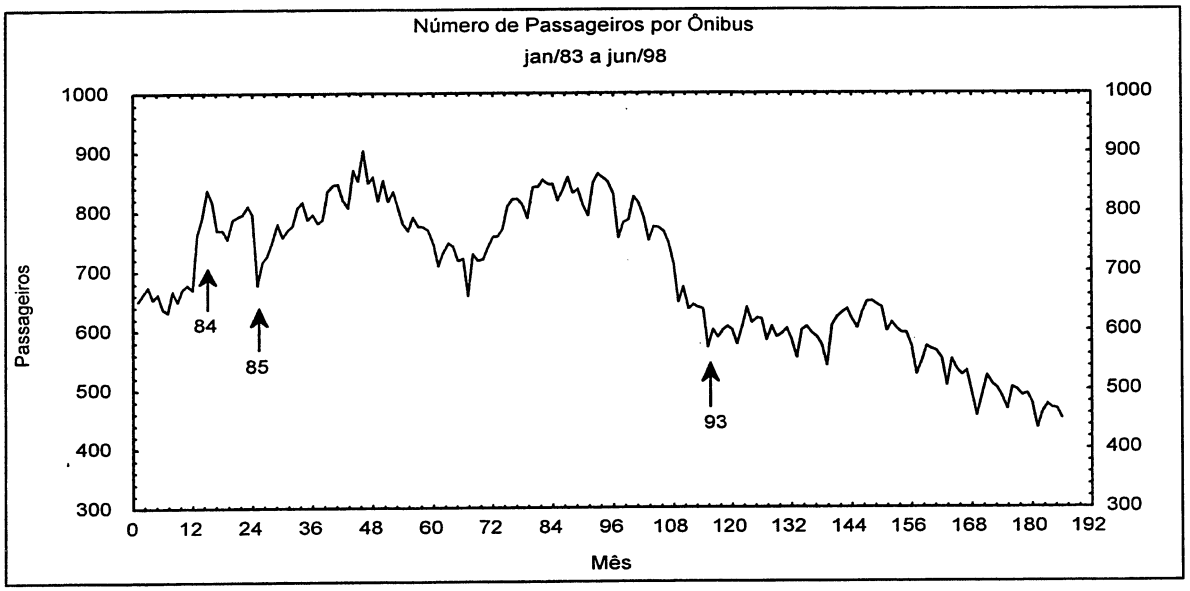
Para cada série foram propostos quatro modelos, sendo dois sem intervenção e dois com intervenção. Os critérios adotados para a escolha do melhor modelo para cada série foram o critério de Akaike (Littell et all,1996), o quadrado médio residual , o erro quadrático médio e o teste de Box e Pierce (1970).

Para prever e analisar as séries de transporte urbano, aplicam-se as metodologias de análise de intervenção e de séries temporais, utilizando os modelos de Box e Jenkins.

Aplicação 1: número de passageiros por ônibus

Esta série indica o número médio mensal de passageiros por ônibus que circulam dentro de São Paulo, no período de janeiro de 1983 a dezembro de 1998, sendo as observações de julho de 1998 a dezembro de 1998 excluídas da análise para serem confrontadas com os valores de previsão.

Figura 2 - Série do número médio mensal de passageiros por ônibus no período de janeiro de 1983 a junho de 1998



Através da **Figura 2** observam-se três mudanças de níveis que podem ser consideradas como intervenção, sendo a primeira no instante 13 que será representado por $x_{1,t}$, em janeiro de 1984; a segunda no instante 25 representado por $x_{2,t}$, em janeiro de 1985, considerando-se também *a priori* como possível efeito de intervenção, a privatização da CMTC, ocorrida no instante 115 em janeiro de 1993, sendo representado por $x_{3,t}$.

A partir de 1983 a gestão do serviço de transporte público voltou a ser alvo da atenção da administração municipal, melhorando a qualidade do transporte urbano, com o intuito de aumentar o número de passageiros nos ônibus.

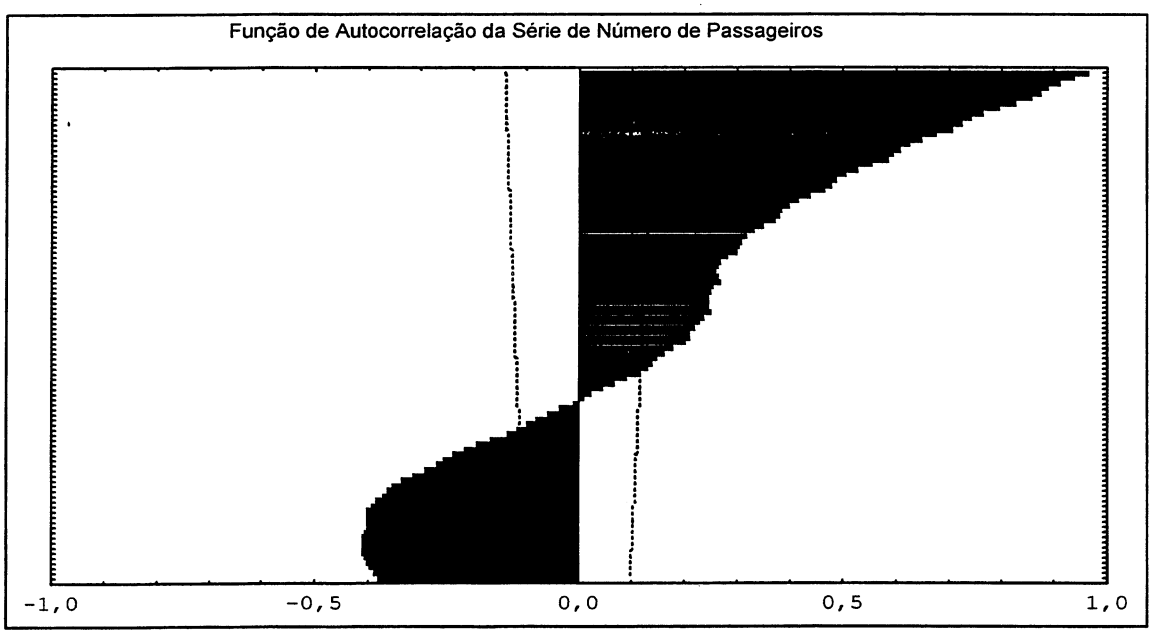
No segundo semestre de 1984 há um acréscimo no número de ônibus em circulação. Neste período entra em circulação a linha operadora a gás metano, ocasionando a queda no número de passageiros por ônibus, mas a série só apresenta efeito da queda de passageiros em janeiro de 1985.

No período da privatização, em janeiro de 1993, representado pelo instante 115, verifica-se que existe um decréscimo no número de passageiros por ônibus, o qual pode ter ocorrido pela inclusão de várias empresas privadas, as quais entram em circulação na cidade, aumentando o número das frotas de ônibus.

Para uma análise mais detalhada dos dados, é necessário verificar o efeito das componentes de tendência e (ou) sazonalidade na série a fim de obter a série estacionária, para que se ajustem modelos e analise os possíveis efeitos de intervenção.

Através do gráfico da função de autocorrelação , **Figura 3** , observou-se a presença de tendência e sazonalidade na série. Assim, foram feitas diferenças de ordem 1 para tirar a tendência e de ordem 12 para tirar a sazonalidade.

Figura 3 - Função de autocorrelação da série de número médio mensal de passageiros



A **Figura 4** apresenta as funções de autocorrelação e autocorrelação parcial da série sem tendência e sazonalidade. Baseado nestes gráficos, observa-se a existência de correlação significativa nos instantes múltiplos de 12, levando à escolha do modelo sazonal SARIMA.

Através da análise visual da FAC e da FACP, propõe-se dois modelos sem intervenção:

o modelo SARIMA$(0,1,0)\text{x}(0,1,1)_{12}$

$$(1-B)(1-B^{12})z_t = (1-\Theta_1 B^{12})a_t \qquad (2)$$

e o modelo SARIMA$(0,1,0)\text{x}(2,1,0)_{12}$.

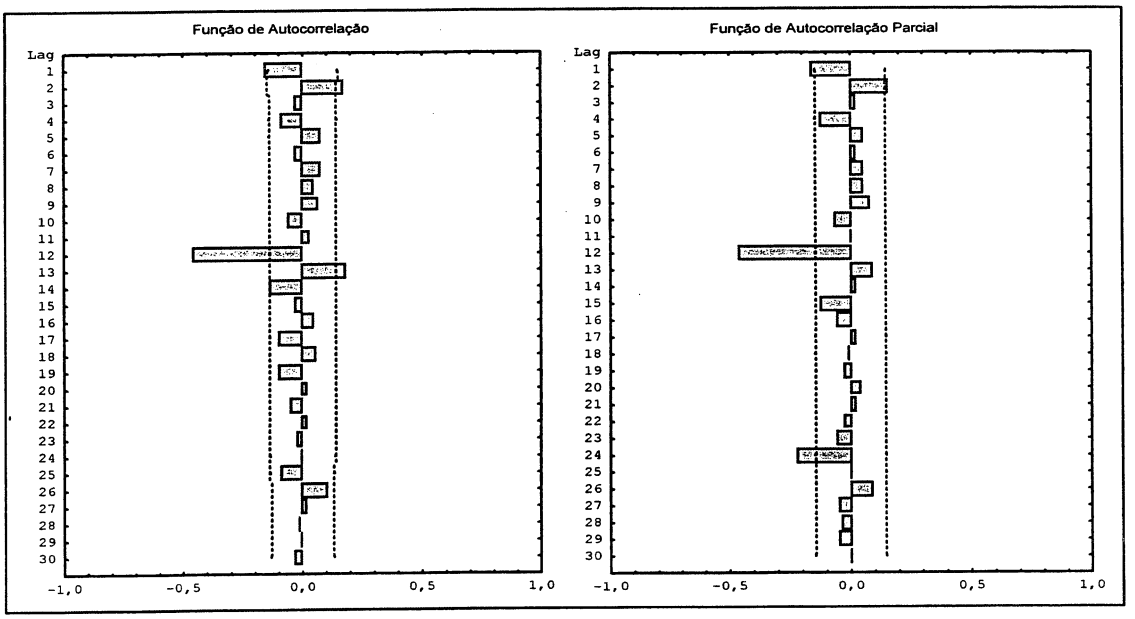
$$(1-\Phi_1 B^{12}-\Phi_2 B^{24})(1-B)(1-B^{12})z_t = a_t \qquad (3)$$

onde a_t é o resíduo (ruído branco) e $\Phi(B) = (1-\Phi B^{12}-\ldots-\Phi_P B^{12P})$ e

$\Theta(B) = (1-\Theta_1 B^{12}-\ldots-\Theta_Q B^{12Q})$ são os polinômios de graus P (autorregressivo sazonal) e Q (médias-móveis sazonal), respectivamente.

É importante esclarecer que poderiam ser propostos mais modelos para o ajuste desta série temporal.

Figura 4 - Função de autocorrelação e função de autocorrelação parcial para a série diferenciada do número médio mensal de passageiros por ônibus



Para ajustar os modelos de intervenção, considera-se para o resíduo n_t os modelos (2) e (3). Estimando os efeitos de intervenção, observa-se que o efeito da privatização na série ocorrido em janeiro de 1993 $(x_{3,t})$ não é significativo, e o efeito de intervenção ocorrido em janeiro de 1984 $(x_{1,t})$ não é possível estimar por estar no início da série, pois foi feita a diferença de ordem 12 para se tirar a sazonalidade. O efeito de intervenção em janeiro de 1985 $(x_{2,t})$ é significativo, sendo a função de transferência da forma $v_i(B) = w_0$, determinado pela Figura 1.b).

Com referência nas variáveis binárias descritas por $x_{2,t}$, apresenta-se a intervenção

$$x_{2,t} = \begin{cases} 1, & T = 25 \\ 0, & T \neq 25 \end{cases}$$

Propõem-se, então, os seguintes modelos de intervenção

$$z_t = w_0 x_{2,t} + \frac{(1-\theta_1 B^2)}{(1-B)(1-B^{12})(1-\Phi_1 B^{12}-\Phi_2 B^{24})} a_t \qquad (4)$$

$$z_t = w_0 x_{2,t} + \frac{(1-\Theta_1 B^{12})}{(1-B)(1-B^{12})(1-\phi_1 B)} a_t \qquad (5)$$

Aplicando-se os critérios de Akaike, do quadrado médio residual e o teste de Box e Pierce para a escolha dos modelos mais adequados sem intervenção e com intervenção, escolhe-se os modelos (2) e (5), apresentados nas **Tabelas 1 e 2.**

A **Tabela 1** apresenta a estimação do modelo sem intervenção tendo todas as estimativas significativas. A estimação do modelo com intervenção, na **Tabela 2**, apresenta que no período de janeiro de 1985 houve um decréscimo médio de 137 passageiros por ônibus. Em ambas as tabelas foi considerado o nível de 5% de significância. Através do teste de Box e Pierce, verificou-se que os resíduos dos dois modelos são independentes e identicamente distribuídos com média zero e variância constante, sendo portanto ruído branco.

Tabela 1 - Estimativas dos parâmetros do modelo sem intervenção

Modelo: SARIMA(0,1,0)x(0,1,1)$_{12}$				
Parâmetro	**Estimativa**	**Erro-padrão**	**Teste t**	**p-value**
Θ_1	0,6293	0,0594	10,59	<0,001
Quadrado Médio Residual			673,96	
AIC			1618,73	
Teste de Box e Pierce			Q=22,38 < χ_{22}=33,93 (ruído branco)	

Tabela 2 - Estimativas dos parâmetros do modelo com intervenção

Modelo: SARIMA(1,1,0)x(0,1,1)$_{12}$ w_0= Janeiro 1985				
Parâmetro	**Estimativa**	**Erro-padrão**	**Teste t**	**p-value**
ϕ_1	-0,1843	0,0756	-2,44	0,012
Θ_1	0,8428	0,0447	18,87	< 0,001
w_0	-136,9334	10,6937	-12,81	< 0,001
Quadrado Médio Residual			405,43	
AIC			1532,78	
Teste de Box e Pierce			Q=20,59 < χ_{21}=32,68 (ruído branco)	

As previsões realizadas para o melhor modelo com intervenção (5) encontram-se na **Tabela 3** e para o modelo sem intervenção (2) estão apresentadas na **Tabela 4**, estas são estimadas, utilizando o método da máxima verossimilhança, no período de julho a dezembro de 1998, considerando o intervalo de confiança de 95%.

Tabela 3 - Previsão para o modelo SARIMA(1,1,0)x(0,1,1)$_{12}$ w_0= janeiro 1985

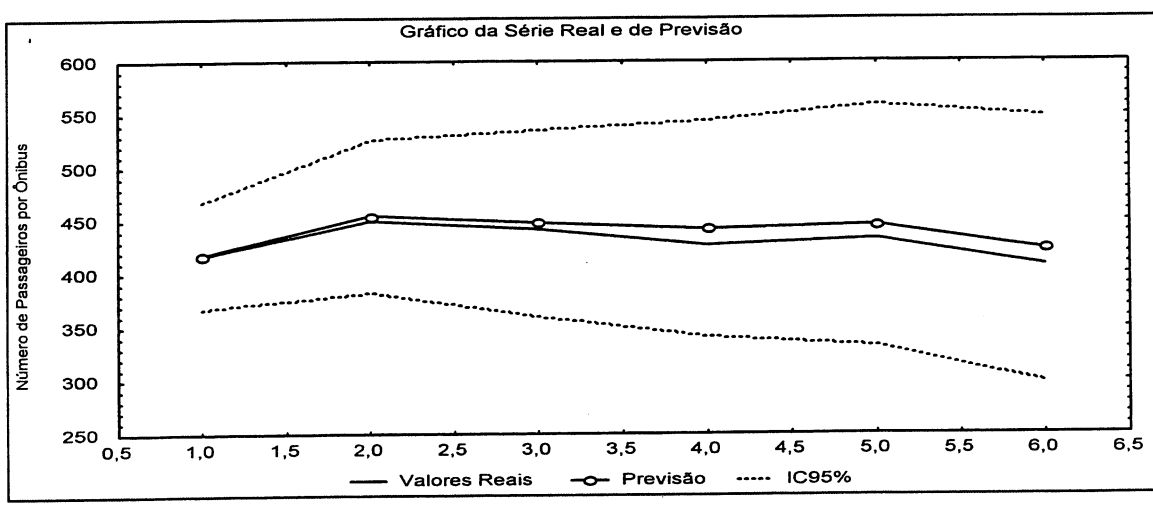
Data	**Valor Real**	**Previsão**	**LI:95%**	**LS:95%**
Julho 1998	417	420	381	460
Agosto 1998	450	458	407	509
Setembro 1998	442	452	391	513
Outubro 1998	427	453	384	523
Novembro 1998	433	454	377	531
Dezembro 1998	408	437	353	520
Erro Quadrático Médio			2131	

Tabela 4 - Previsão para o modelo SARIMA(0,1,0)x(0,1,1)$_{12}$

Data	Valor Real	Previsão	LI:95%	LS:95%
Julho 1998	417	418	367	468
Agosto 1998	450	455	383	527
Setembro 1998	442	448	360	536
Outubro 1998	427	442	341	544
Novembro 1998	433	446	332	559
Dezembro 1998	408	423	298	548
Erro Quadrático Médio			681	

Através do Erro Quadrático Médio - EQM -, observa-se que o modelo sem intervenção fornece melhores previsões, ao modelo com intervenção, sendo para esta série não indicado utilizar a intervenção.

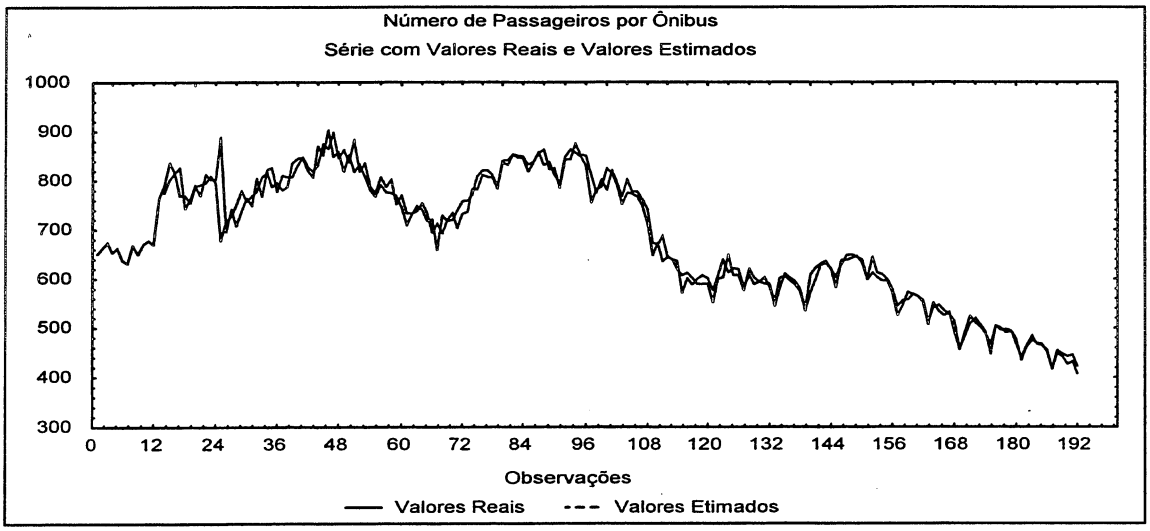
Figura 5 - Série real e de previsão (modelo 5) com seus respectivos I.C. 95%, no período de julho de 1998 a dezembro de 1998



A **Figura 5** apresenta os valores reais, de previsão (**Tabela 4**) e seus respectivos intervalos de confiança de 95% para a série do número médio de passageiros por ônibus, no período de julho de 1998 a dezembro de 1998. Observa-se, também, que todos valores reais estão dentro do intervalo de confiança dos valores preditos.

A **Figura 6** apresenta os valores reais e os valores estimados da série, considerando o modelo sem intervenção (2), no período de janeiro de 1983 a dezembro de 1998. Analisando o comportamento dos valores estimados no modelo, observa-se uma similaridade com os valores reais da série.

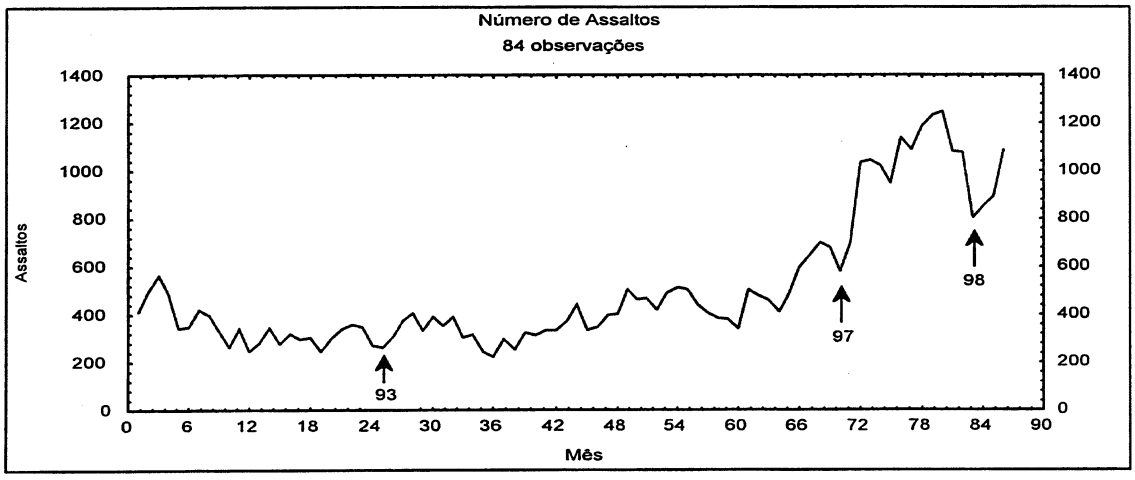
Figura 6 - Série real e estimada no período de janeiro de 1983 a dezembro de 1998



Aplicação 2: Número de assaltos nos ônibus

Esta série fornece o número de assaltos que ocorreram nos ônibus urbanos de São Paulo, no período de janeiro de 1992 a agosto de 1999, sendo as observações de março a agosto de 1999 retiradas da série para serem comparadas com os valores preditos.

Figura 7 - Série do número de assaltos mensais no período de janeiro de 1992 a agosto de 1999



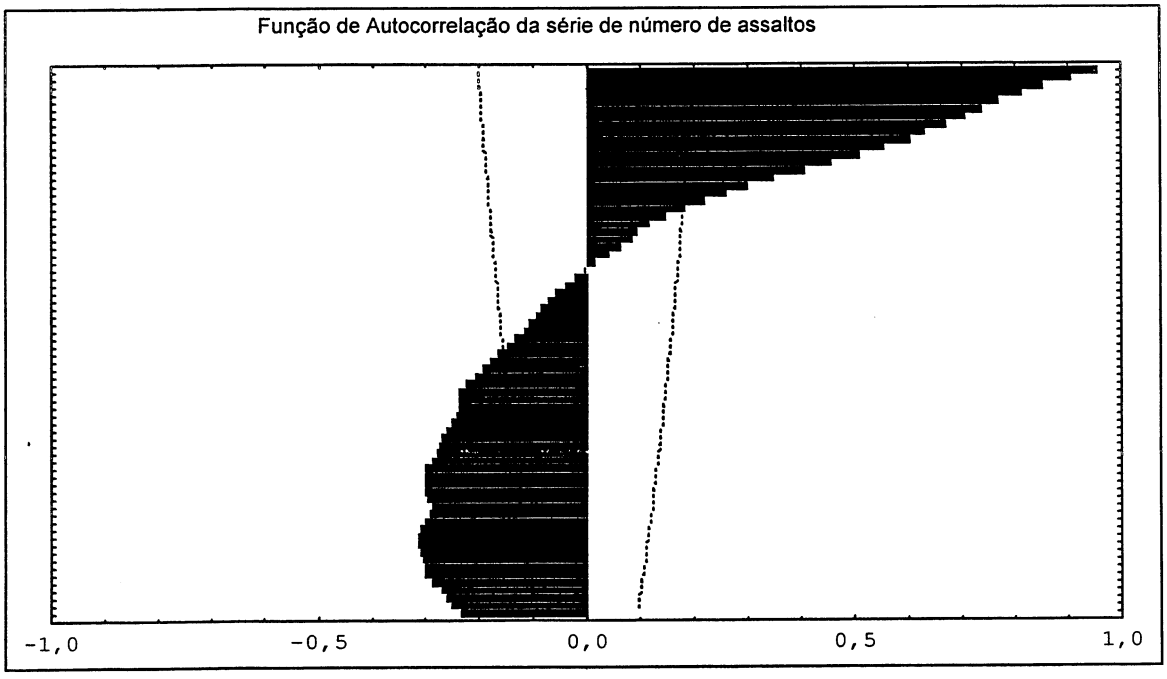
Através da **Figura 7** considera-se a inclinação existente na 72ª observação, dezembro de 1997, como uma possível intervenção. Neste período houve aumento no número de assaltos nos ônibus, sendo constatado somente a introdução dos "perueiros" como uma opção de meio de transporte.

No período referente a 83ª observação, novembro de 1998, ocorreu uma queda no número de assaltos, sendo proposto como segundo efeito de intervenção. Como não se constata nenhum evento que possa ter

reduzido o número de assaltos nesta época, não foi possível identificar o fenômeno que ocasionou a queda no número de assaltos.

Na época de privatização, em janeiro de 1993, a qual poderia ter elevado o número de assaltos nos ônibus, por causa do aumento das frotas de ônibus reduzindo o número médio de passageiros, possivelmente não exista efeito de intervenção.

Figura 8 - Função de autocorrelação da série de número de assaltos mensais

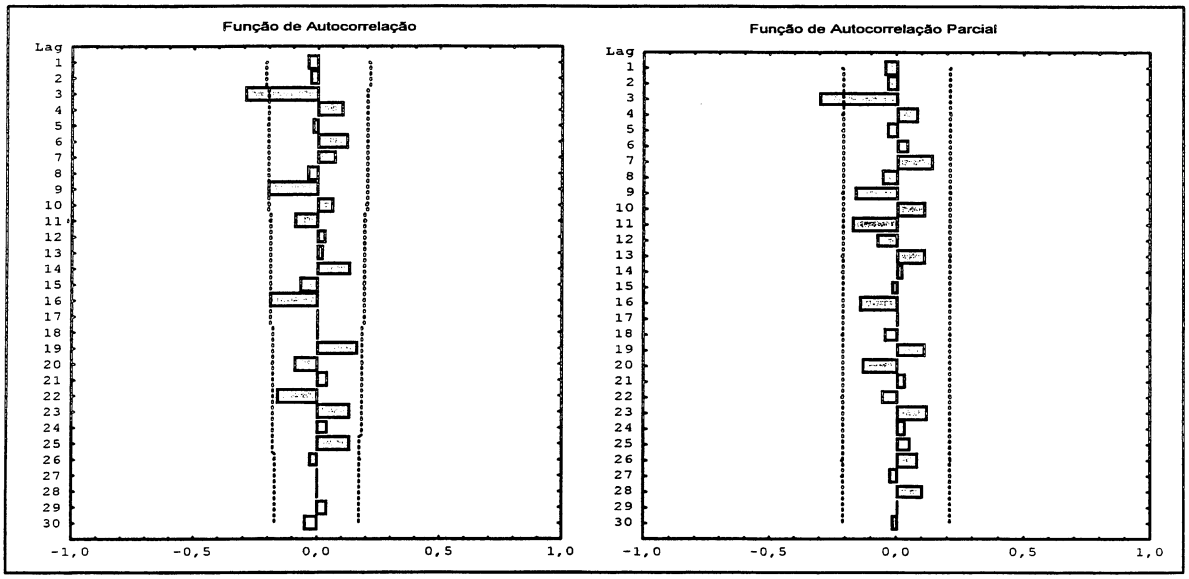


A série temporal apresenta uma componente de tendência a qual pode ser observada no gráfico da função de autocorrelação , **Figura 8**.

A FAC e a FACP para a série diferenciada, isto é, sem o efeito de tendência **(Figura 9)**, fornecem informações para ajustar os modelos para série.

Baseando-se na **Figura 9** propõe-se dois modelos sem intervenção. A FAC apresenta correlação no instante 3, sendo interessante compor um modelo médias-móveis de ordem 3, e a FACP também apresenta correlação no instante 3, sendo proposto um modelo auto-regressivo de ordem 3.

Figura 9 -Função de autocorrelação e função de autocorrelação parcial para a série diferenciada do número de assaltos mensais



Considerando o efeito da FACP, com correlação no instante 3, indica-se o modelo ARIMA(3,1,0).

$$(1-\phi_3 B^3)(1-B)\eta_t = a_t \tag{6}$$

Verificando a FAC, com correlação no instante 3, indica-se o modelo ARIMA(0,1,3)

$$(1-B)\eta_t = (1-\theta_3 B^3)a_t \tag{7}$$

Estimando-se os dois modelos de intervenção propostos a seguir, observou-se efeitos de intervenção significativo apenas em dezembro de 1997 e novembro de 1998, os quais serão representados por $x_{1,t}$ e $x_{2,t}$, respectivamente.

A intervenção ocorrida em dezembro de 1997, representada por $x_{1,t}$, tem efeito gradual e duração permanente. Em novembro de 1998, representada por $x_{2,t}$, sofre efeito abrupto e duração temporária. As funções de transferência $(v_i(B))$, representadas através das variáveis binárias $x_{1,t}$ e $x_{2,t}$ estão apresentadas a seguir, respectivamente. Através do item e) da **Figura 1**, observa-se o efeito de intervenção da variável binária $x_{1,t}$, e o item b) mostra o efeito de $x_{2,t}$.

$$v(B) = \frac{w_{1,0}}{1-B} \qquad \text{v(B)} = \text{w}_{2,0}$$

Os efeitos de intervenção representados pelas variáveis binárias $x_{1,t}$ e $x_{2,t}$, são dados por

$$x_{1,t} = \begin{cases} 1, & \text{T} \geq 72 \\ 0, & \text{T} < 72 \end{cases} \qquad x_{2,t} = \begin{cases} 1, & \text{T} = 83 \\ 0, & \text{T} \neq 83 \end{cases}$$

Os modelos de intervenção, descritos a seguir, têm seus respectivos resíduos descritos através dos modelos sem intervenção (6) e (7), respectivamente.

$$z_t = \frac{w_{1,0}}{1-B} x_{1,t} + w_{2,0} x_{2,t} + \frac{1}{(1-B)(1-\phi_3 B^3)} a_t \tag{8}$$

$$z_t = \frac{w_{1,0}}{1-B} x_{1,t} + w_{2,0} x_{2,t} + \frac{(1-\theta_3 B^3)}{(1-B)} a_t \qquad (9)$$

Através dos critérios de Akaike e do quadrado médio residual, escolhe-se o melhor modelo sem intervenção e o melhor com intervenção para fazer previsão. Considerando as estimativas dos modelos propõe-se os modelos (6) e (9).

A **Tabela 5** apresenta estimativa significativa no modelo sem intervenção. A **Tabela 6** fornece a estimativa do modelo ARIMA e as estimativas de intervenção, sendo que no mês de dezembro de 1997 houve um aumento médio de 42 assaltos e em novembro de 1998 um decréscimo de 289 assaltos. Em ambas as tabelas a significância testada foi de 5%.

Tabela 5 - Estimativas dos parâmetros do modelo sem intervenção

Modelo: ARIMA(3,1,0)				
Parâmetro	**Estimativa**	**Erro-padrão**	**Teste t**	**p-value**
ϕ_3	-0,3060	0,1076	-2,85	0,005
Quadrado Médio Residual			6269,83	
AIC			985,71	
Teste de Box e Pierce			Q=22,07 < χ_{22}=33,93 (ruído branco)	

Tabela 6 - Estimativas dos parâmetros do modelo com intervenção

Modelo: ARIMA(0,1,3)	$w_{1,0}$=Dezembro 1997 e $w_{2,0}$=Novembro 1998			
Parâmetro	**Estimativa**	**Erro-padrão**	**Teste t**	**p-value**
θ_3	0,3085	0,1112	2,77	0,007
$w_{1,0}$	41,8636	15,3972	2,72	0,008
$w_{2,0}$	-288,6983	76,1489	-3,79	< 0,001
Quadrado Médio Residual			4490,63	
AIC			959,27	
Teste de Box e Pierce			Q=21,33 < χ_{22}=33,93 (ruído branco)	

Através do teste de Box e Pierce (**Tabelas 5 e 6**), verificou-se que o resíduo dos respectivos modelos são ruído branco, considerando-se, portanto, que estes modelos estão se ajustando adequadamente à série do número de assaltos nos ônibus.

O melhor modelo sem intervenção, representado por (6), e o melhor modelo com intervenção dado por (9) são submetidos à previsão para que se possa analisar, através do erro quadrático médio, qual dos dois modelos fornece melhor previsão no período de março a agosto de 1999.

Na **Tabela 7**, observa-se os valores preditos e seus respectivos intervalos de confiança de 95% do modelo com intervenção, e a **Tabela 8** fornece os valores do modelo sem intervenção.

Tabela 7 - Previsão para o modelo ARIMA(0,1,3) $w_{1,0}$=Dezembro 1997 e $w_{2,0}$=Novembro 1998

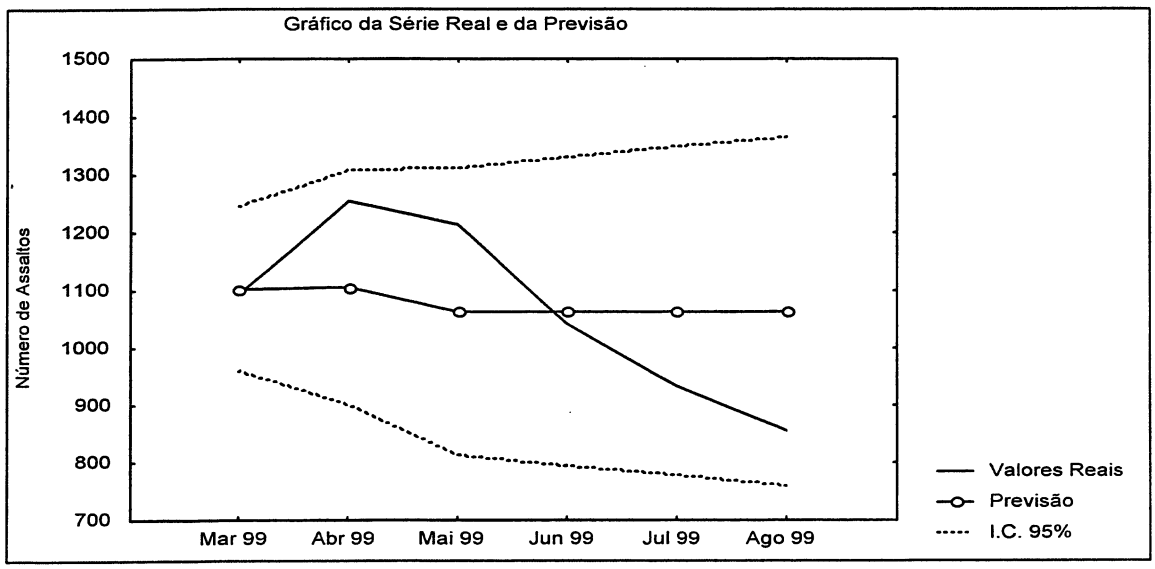
Data	Valor Real	Previsão	LI:95%	LS:95%
Março 1999	1094	1084	953	1216
Abril 1999	1255	1070	884	1256
Maio 1999	1214	1018	790	1245
Junho 1999	1042	1018	772	1264
Julho 1999	933	1018	755	1281
Agosto 1999	856	1018	739	1296
Erro Quadrático Médio			106708	

Tabela 8 - Previsão para o modelo ARIMA(3,1,0)

Data	Valor Real	Previsão	LI:95%	LS:95%
Março 1999	1094	1071	916	1227
Abril 1999	1255	1058	839	1278
Maio 1999	1214	1000	731	1269
Junho 1999	1042	1004	715	1294
Julho 1999	933	1008	699	1371
Agosto 1999	856	1026	699	1353
Erro Quadrático Médio			121103	

Através do EQM, observa-se que o modelo com intervenção fornece melhores previsões ao modelo sem intervenção, sendo o modelo com intervenção proposto para o ajuste da série do número de assaltos.

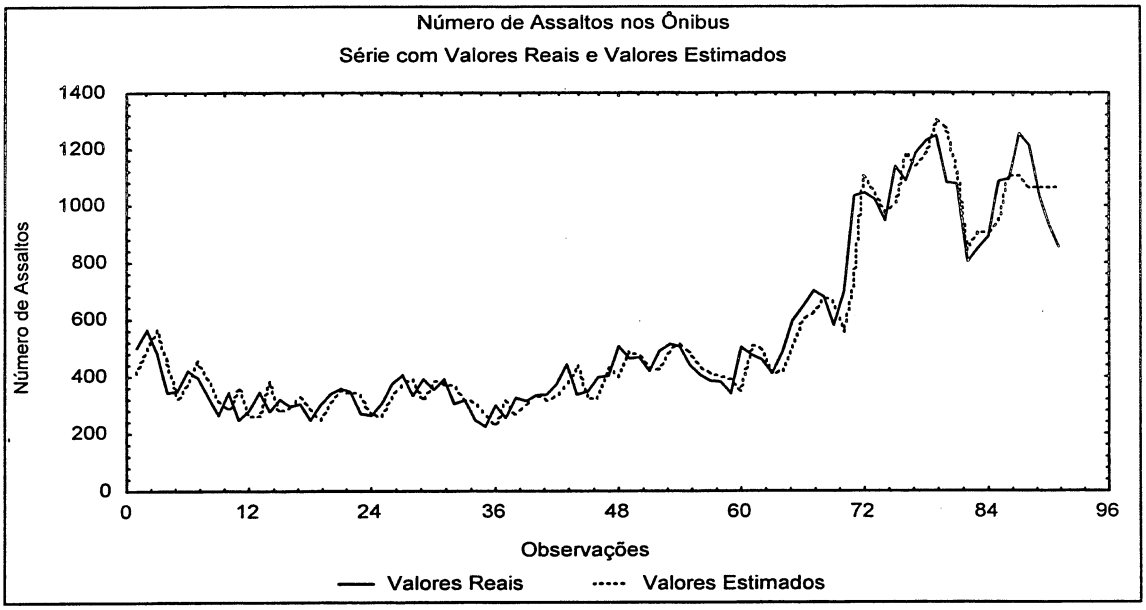
Figura 10 - Série real e de previsão (modelo 9) com seus respectivos IC 95%, no período de março de 1999 a agosto de 1999



Através da **Figura 10,** observa-se que os valores preditos do modelo com intervenção não seguiram o comportamento dos valores reais, mas os valores reais apresentados no período de março a agosto de 1999 encontram-se todos dentro do intervalo de confiança de 95%.

A **Figura 11** apresenta os valores estimados a partir do modelo com intervenção, no período de janeiro de 1992 à agosto de 1999, observando que o ajuste do modelo seguiu o comportamento da série original até fevereiro de 1999.

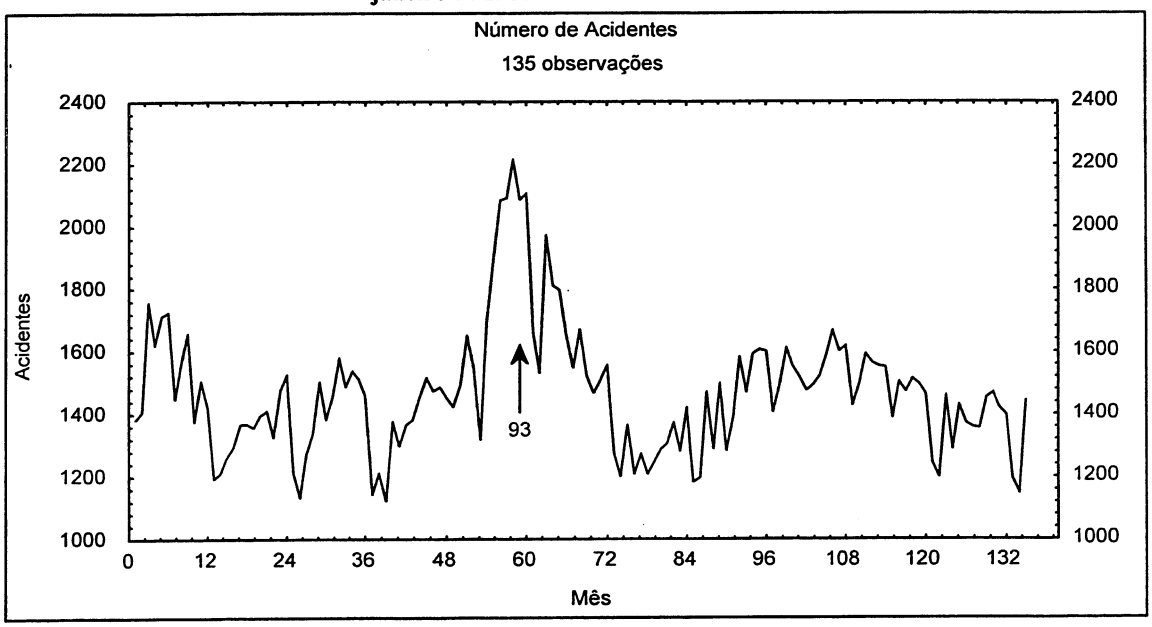
Figura 11 - Série real e estimada no período de janeiro de 1992 a agosto de 1999



Aplicação 3: número de acidentes com os ônibus

Esta série descreve o número de acidentes ocorridos com os ônibus urbanos de São Paulo, no período de janeiro de 1988 a setembro de 1999, retirando-se as observações de abril a setembro de 1999 da série para serem comparadas com os valores de previsão.

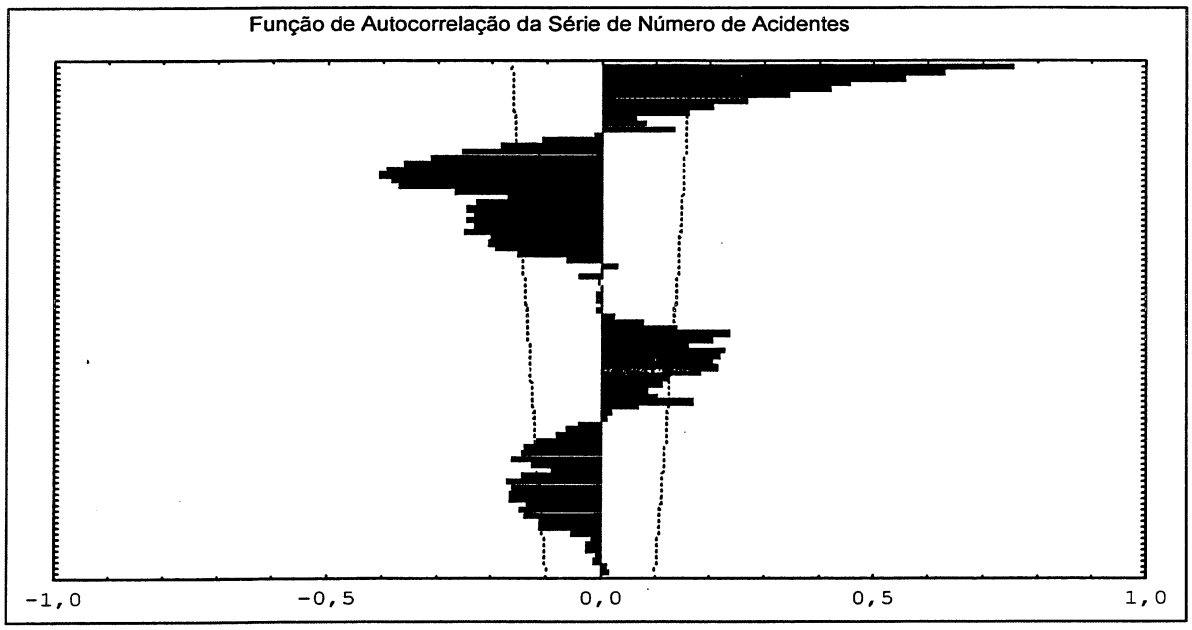
Figura 12 - Série do número de acidentes mensais com os ônibus urbanos no período de janeiro de 1988 a setembro de 1999



Através da **Figura 12**, observa-se que em torno da 58ª observação, outubro de 1992, ocorreu um acréscimo no número de acidentes. O comportamento da série apresenta-se em torno de 1400 acidentes mensais, sendo que no início de 1993 este índice aumenta em média para mais de 2000 acidentes mensais.

Considera-se um período de defasagem para um possível efeito de intervenção ocorrido na 62ª observação, referente a março de 1993, sendo representada por $x_{1,t}$. Este fato ocorre no período da privatização da CMTC, ou seja, nesta época o número de acidentes com os ônibus aumentaram consideravelmente.

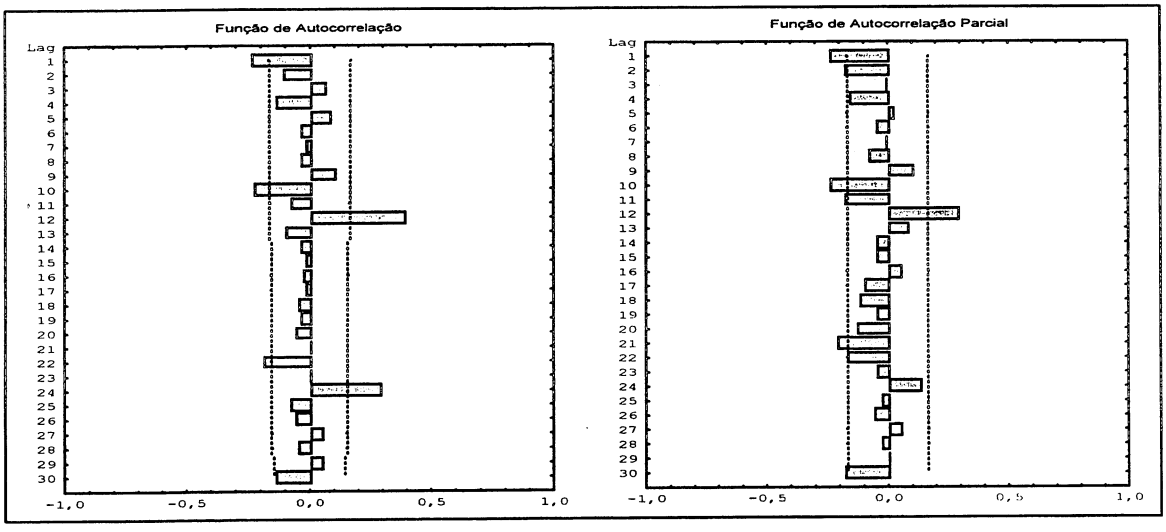
Figura 13 - Função de autocorrelação da série de número de acidentes com os ônibus urbanos



Através da análise gráfica, **Figura 13**, verificou-se efeito de tendência na série e como a componente sazonal da série tende para zero não foi necessário tirar o efeito desta componente para ajustar os modelos de Box e Jenkins.

Analisando a FAC e a FACP da **Figura 14** sem a presença de tendência, observa-se que a série apresenta correlação nos instantes múltiplos de 12, sendo necessário acrescentar uma componente sazonal, nos modelos que serão propostos. Além das componentes sazonais, existe correlação no primeiro instante das FAC e FACP, apresentando estruturas auto-regressivas e de médias-móveis de ordem 1. Observa-se, ainda, que existe correlação no instante 10 da FAC e FACP, mas esta correlação pode estar sendo influenciada pela correlação existente no instante 12.

Figura 14 - Função de autocorrelação e função de autocorrelação parcial para a série diferenciada do número de acidentes mensais com os ônibus urbanos



É proposto dois modelos sem intervenção, com referência na FAC e FACP.

O primeiro modelo tem estrutura de médias-móveis, com correlação nos instantes 1, 12 e 24, portanto indica-se o modelo SARIMA(0,1,1)x(0,0,2)$_{12}$

$$(1-B)\eta_t = (1-\theta_1 B)(1-\Theta_1 B^{12} - \Theta_2 B^{24})a_t \tag{10}$$

Também é interessante propor o modelo SARIMA(0,1,1)x(1,0,0)$_{12}$

$$(1-\Phi_1 B^{12})(1-B)\eta_t = (1-\theta_1 B)a_t \tag{11}$$

Para a intervenção optou-se inicialmente pelas funções de transferências dadas pelas Figuras 1(b) e 1(c), o que significaria no segundo caso que a intervenção poderia ter ocorrido num intervalo de tempo contendo T=62. Após a análise verificou-se que o parâmetro δ não foi significativo. Assim o efeito de intervenção, que ocorreu em março de 1993, representado pela variável binária $x_{1,t}$, no período de privatização da CMTC, foi considerado de efeito abrupto e duração temporária

$$x_{1,t} = \begin{cases} 1, & T = 62 \\ 0, & T \neq 62 \end{cases}$$

e a estrutura da função de transferência tem a forma da Figura 1.b), $v(B) = w_0$.

O primeiro modelo com intervenção tem o resíduo representado através do modelo (11)

$$z_t = w_0 x_{1,t} + \frac{(1-\theta_1 B)}{(1-B)(1-\Phi_1 B^{12})} a_t \tag{12}$$

Para o modelo com intervenção (13), ajustou-se o modelo SARIMA(1,1,0)x(0,0,1), que não havia sido proposto através dos modelos sem intervenção.

$$z_t = w_0 x_{1,t} + \frac{(1-\Theta_1 B^{12})}{(1-B)(1-\phi_1 B)} a_t \qquad (13)$$

Através do modelo com intervenção (13), fica evidente que o resíduo deste modelo não precisa ser necessariamente igual ao modelo sem intervenção, pois com a ocorrência da intervenção o valor residual do modelo pode ser alterado, propondo uma nova estrutura dos modelos de Box e Jenkins que ajuste melhor aos dados.

Através dos critérios de Akaike e do quadrado médio residual, propõe-se o modelo sem intervenção (11) e o modelo com intervenção (12), para fazer previsão. As estimativas destes modelos estão nas **Tabelas 9** e **10**, respectivamente.

Tabela 9 - Estimativas dos parâmetros do modelo sem intervenção

Modelo: SARIMA(0,1,1)x(1,0,0)$_{12}$				
Parâmetro	**Estimativa**	**Erro-padrão**	**Teste t**	**p-value**
θ_1	0,2850	0,0837	3,41	0,001
Φ_1	0,4372	0,0796	5,49	<0,001
Quadrado Médio Residual			15370,22	
AIC			1676,68	
Teste de Box e Pierce			Q=27,15 < χ_{23}=35,18 (ruído branco)	

Tabela 10 - Estimativas dos parâmetros do modelo com intervenção

Modelo: SARIMA(0,1,1)x(1,0,0)$_{12}$ w_0= Março de 1993				
Parâmetro	**Estimativa**	**Erro-padrão**	**Teste t**	**p-value**
θ_1	0,2317	0,0873	2,65	0,008
Φ_1	0,4383	0,0800	5,48	<0,001
w_0	259,3272	110,6137	2,34	0,015
Quadrado Médio Residual			14885,48	
AIC			1673,35	
Teste de Box e Pierce			Q=22,48 < χ_{21}=32,68 (ruído branco)	

Considerando as **Tabelas 9** e **10,** observa-se que as estimativas dos parâmetros dos modelos são significativas, no nível de 5% de significância, e aplicando o teste de Box e Pierce comprova-se que o resíduo dos modelos são um ruído branco, se ajustando bem aos valores reais da série. De acordo com a **Tabela 10**, existe um acréscimo de 259 acidentes no período de março de 1993.

Utilizando os modelos (11) e (12), realiza-se as previsões no período de abril a setembro de 1999. A **Tabela 11** fornece os valores preditos para o modelo com intervenção e a **Tabela 12**, para o modelo sem intervenção.

Observa-se que o modelo com intervenção apresenta erro quadrático médio menor que o modelo sem intervenção, sendo proposto para o ajuste da série do número de acidentes com os ônibus.

Tabela 11- Previsão para o modelo SARIMA$(0,1,1)x(1,0,0)_{12}$ w_0= março de 1993

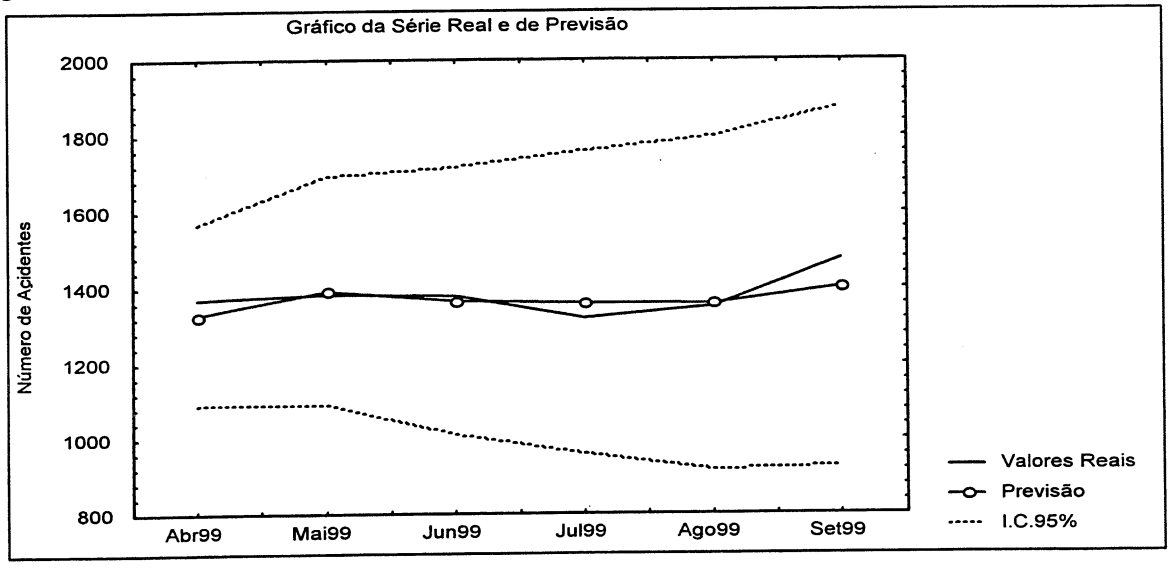
Data	Valor Real	Previsão	LI:95%	LS:95%
Abril 1999	1370	1329	1090	1568
Maio 1999	1384	1392	1090	1694
Junho 1999	1379	1366	1013	1719
Julho 1999	1320	1360	962	1758
Agosto 1999	1350	1358	920	1797
Setembro 1999	1477	1402	926	1877
Erro Quadrático Médio		9203		

Tabela 12 - Previsão para o modelo SARIMA$(0,1,1)x(1,0,0)_{12}$

Data	Valor Real	Previsão	LI:95%	LS:95%
Abril 1999	1370	1321	1079	1565
Maio 1999	1384	1384	1086	1683
Junho 1999	1379	1358	1012	1704
Julho 1999	1320	1352	966	1739
Agosto 1999	1350	1350	927	1775
Setembro 1999	1477	1394	936	1852
Erro Quadrático Médio		10755		

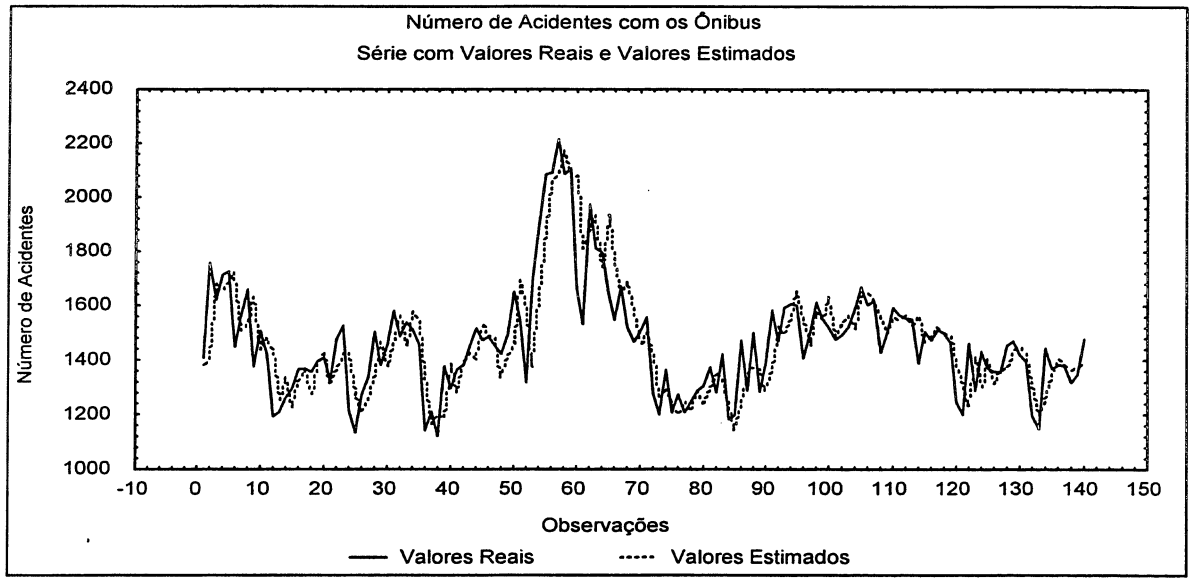
A **Figura 15** mostra que os valores preditos estão seguindo o comportamento dos valores reais da série, no período de abril de 1999 a setembro de 1999, apresentando-se dentro do intervalo de confiança de 95%.

Figura 15 - Valores reais e de previsão com os respectivos IC 95%, no período de abril de 1999 a setembro de 1999



Através da **Figura 16**, compara-se os valores estimados do modelo com intervenção com os valores originais da série, no período de janeiro de 1988 a setembro de 1999, observando que o ajuste do modelo seguiu o comportamento da série real.

Figura 16 - Série real e estimada no período de janeiro de 1988 a setembro de 1999



4. Conclusão

As séries de acidentes e do número médio de passageiros apresentaram correlações sazonais, sendo que o modelo SARIMA se ajustou melhor aos dados. Para a série de assaltos o melhor modelo foi um ARIMA.

Nas três séries estudadas foram propostos quatro modelos, sendo dois sem intervenção e dois com intervenção. Através dos critérios de Akaike, quadrado médio residual e do erro quadrático médio, conclui-se que os modelos com intervenção apresentaram-se mais eficientes do que os modelos sem intervenção, para as três séries.

Na série referente ao número médio de passageiros por ônibus, apesar de o modelo sem intervenção ter gerado previsão melhor do que o modelo com intervenção, a série apresentou um decréscimo significativo em janeiro de 1985 de 136 passageiros por ônibus, decorrente do segundo semestre de 1984.

Na série referente ao número de assaltos, houve efeito de intervenção significativo em dezembro de 1997, aumentando em média 42 assaltos. Em novembro de 1998 houve um decréscimo médio de 289 assaltos, não sendo possível identificar os fatores que influenciaram esta intervenção.

A série do número de acidentes teve efeito de intervenção significativo, no período da privatização da CMTC, referente a março de 1993, aumentando em média 259 acidentes neste período.

Referências bibliográficas

BHATTACHARYYA, M.N.; LAYTON, A.P. Effectiveness of seat belt legislation on the Queensland road toll – An Australian case study in intervention analysis. *Journal of the American Satatistical Association*, Alexandria, v.74, n.367, p.596-603, Sept. 1979.

BOX, G.E.P.; JENKINS, G.M. *Time series analysis, forecasting and control*. San Francisco: Holden-Day, 1970. 575p.

BOX, G.E.P.; JENKINS, G.M.; REINSEL, G.C. *Time series analysis: forecasting and control*. 3 ed. New Jersey: Prentice Hall, 1994. 598p.

BOX, G.E.P.; PIERCE, D.A. Distribution of residual auto-correlations in autorregressive-integrated moving average time series models. *Journal of the American Statistical Association*, Alexandria, v.65, n.332, p.1509-1529, Dec.1970.

BOX, G.E.P.; TIAO, G.C. A change in level of a non-stationary time series. *Biometrika*, London, v.52, n.1/2, p.181-192, June 1965.

BOWERMAN, B.L.; O'CONNELL, R.T. *Forecasting and Time Series: an Applied Approach*. 3 ed., Belmont: Duxbury Press, 1993.

GLASS, G.V. Estimating the effects of intervention into a non-stationary time series. *American Educational Research Journal*, Washington, v.9, n.3, p.463-477, 1972.

GLASS, G.V.; WILLSON, V.L.; GOTTMAN, J.M. *Design and analysis of time series experiments*. Boulder: Colorado Associated University Press, 1975. 241p.

LITTELL, R.C.; MILLEKEN, G.A.; STROUP, W.W.; WOLFEINGER, R.D. *SAS® System for mixed models*. Cary: SAS Institute, 1996. 633p.

PANKRATZ, A. *Forecasting with Dynamic regression Models*. New York: Wiley, 1991. 386p.

PINO, F.A. *Análise de Intervenção em Séries Temporais: Aplicações em Economia Agrícola*. São Paulo: USP, 1980. 253p. (Dissertação – Mestrado em Estatística).

PRIESTLEY, M.B. *Spectral analysis and time series – (Probability and Mathematical Statistical)*. 6 ed., New York: Academic Press, 1989. 890p.

SAS INSTITUTE. SAS/STAT®: user's guide. North Carolina, 1999.

STATISTICA for Windows. Realese 5. Copyright Stat Soft, Inc. 1984-1995.

WOLD, H.O. *A study in the analysis of stationary time series*. Sweden: Uppsala, 1938. 214p.

Agradecimentos

Os autores agradecem aos consultores as sugestões feitas na versão original deste trabalho.

ABSTRACT

This work analyze the effect of urban transportation in city of São Paulo using time series, being include in effect models of intervention with the objective of generate a accuracy forecast. The series used going the mean daily number of passengers, the number of holdups in urban buses and the number of traffic accidents with buses. The identification of models from Box & Jenkins was accomplished by

function of autocorrelation and function autocorrelation partial, increasing parameters for the independent phenomenon of the series, called intervention. The verification of models appropriate was accomplished through of test Box & Pierce, Akaike criterion, residual mean square and errors mean square of forecasted values.

Key-words: Intervention Analysis, SARIMA Models, Urban Transportation

Política editorial

A Revista Brasileira de Estatística - RBEs - objetiva promover a Estatística relevante para aplicação em questões sociais, interpretadas amplamente para incluir questões educacionais, de saúde, demográficas, econômicas, legais, de políticas públicas e de estatísticas oficiais, entre outras. A revista apresenta artigos num formato que permita fácil assimilação pelos membros da comunidade científica em geral. Os artigos devem incluir aplicações práticas como assunto central. Essas aplicações devem ter conteúdo estatístico substancial. As análises devem ser exaustivas e bem apresentadas, mas o emprego de métodos estatísticos inovadores não é essencial para publicação.

Artigos contendo exposição de métodos são aceitáveis, desde que estes sejam relevantes para as áreas cobertas pela revista, auxiliem na compreensão do problema e contenham interpretação clara das expressões matemáticas apresentadas. A apresentação de aplicações ilustrativas envolvendo dados adequados é requerida. Tratamentos algébricos extensos devem ser evitados.

A RBEs tem periodicidade semestral e publicará também artigos escritos a convite e resenhas de livros, bem como artigos abordando os diversos aspectos de metodologias relevantes para órgãos produtores de estatísticas, incluindo:

a) planejamento de pesquisas;
b) avaliação e mensuração de erros em pesquisas;
c) uso e combinação de fontes alternativas de informação; integração de dados;
d) novos desenvolvimentos em metodologia de pesquisa;
e) crítica e imputação de dados;
f) amostragem e estimação;
g) disseminação e confiabilidade de dados;
h) análise de dados;
i) análise de séries temporais;
j) modelos e métodos demográficos; e
k) modelos e métodos econométricos.

Todos os artigos submetidos serão avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs. Os artigos submetidos deverão ser inéditos e não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional. O processo de avaliação é do tipo duplo cego, isto é, os artigos são avaliados sem identificação da autoria, e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos submetidos para publicação deverão ser remetidos em 3 vias (que não serão devolvidas) para:

Pedro Luis do Nascimento Silva
Editor Responsável
Revista Brasileira de Estatística - RBEs
Av. República do Chile 500, 10º andar
Rio de Janeiro – RJ – 20031-170
Tel.: +55 - 21 - 2514 4548
Fax: +55 - 21 - 2514 0039
E-mail: pedrosilva@ibge.gov.br

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

Para cada artigo publicado, serão fornecidas gratuitamente 20 separatas.

Instruções para preparo de originais:

1. 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 filiação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos devem figurar também nesta página;

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

3. 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;

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

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

6. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo, sempre que possível. Quando isto não ocorrer, devem ser traçados em papel branco, como nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho, quer nas legendas ou títulos; e

7. Serão preferidos originais processados pelo editor de texto Word for Windows.

Se o assunto é **Brasil**,
procure o **IBGE**

www.ibge.gov.br
www.ibge.net

atendimento
0800 21 81 81