Presidente da República
Fernando Henrique Cardoso

Ministro de Estado do Planejamento e Orçamento
Paulo de Tarso Almeida Paiva

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Simon Schwartzman

Diretor de Planejamento e Coordenação
Nuno Duarte da Costa Bittencourt

ÓRGÃOS TÉCNICOS SETORIAIS

Diretoria de Pesquisas
Lenildo Fernandes Silva

Diretoria de Geociências
Trento Natali Filho

Diretoria de Informática
Fernando Elyas Nóbrega Nasser

Centro de Documentação e Disseminação de Informações
David Wu Tai

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 58 número 209 janeiro/junho 1997

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 58, n. 209, p. 1-148, jan./jun. 1997

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 Responssável - 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.

Impressão
Centro de Documentação e Disseminação de Informações-CDDI/ IBGE, em meio digital, em 1998.

Capa
Renato J. Aguiar - Divisão de Criação - DIVIC/CDDI

Ilustração da Capa
Marcos Balster - Divisão de Criação - DIVIC/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.

1. Estatística - Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

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

Impresso no Brasil/Printed in Brazil

SUMÁRIO

NOTA DO EDITOR 5

ARTIGOS

ESTATÍSTICAS DA POBREZA

7

Simon Schwartzman

USO DO SISTEMA DIA PARA A DETECÇÃO E CORREÇÃO AUTOMÁTICA DE ERROS NOS DADOS DO QUESTIONÁRIO BÁSICO DO CENSO DEMOGRÁFICO DE 1991

19

Luis Carlos de Souza Oliveira

Rita Luzia Aguiar Lima

Laura Baridó Indá

ALGUNS MODELOS COMPLEXOS DE REGRESSÃO PARA ANÁLISE DE DADOS UNIVARIADOS

53

Gaus M. Cordeiro

EXPERIMENTOS COM INTERCÂMBIO DE DOIS TRATAMENTOS E DOIS PERÍODOS: ESTRATÉGIAS PARA ANÁLISE E ASPECTOS COMPUTACIONAIS

81

Denise A Botter

Julio M. Singer

MÉTODOS AUTOMÁTICOS DE PREVISÃO PARA SÉRIES TEMPORAIS MULTIVARIADAS

105

Enivaldo Carvalho da Rocha

Basilio de Bragança Pereira

POLÍTICA EDITORIAL 147

NOTA DO EDITOR

Prezado(a) Leitor(a),

Este é o primeiro número da Revista Brasileira de Estatística (RBEs) publicado conjuntamente com a Associação Brasileira de Estatística (ABE). A contribuição da ABE na sua produção foi decisiva, com a indicação dos membros do conselho editorial, que vem atuando decisivamente no encorajamento à submissão de trabalhos e no processo de revisão e avaliação dos artigos submetidos. Não foi ainda possível regularizar os prazos de produção da revista, devido a dificuldades com a editoração da mesma no IBGE. Tal problema já foi solucionado e esperamos reduzir bastante o prazo de preparação do próximo número. Isto será possível porque já temos material suficiente e com a editoração adiantada. Para tal foi fundamental a contribuição dos autores, que têm submetido os artigos já editados e em arquivos digitais.

O conselho editorial se reuniu em agosto de 1997 e discutiu idéias para a reformulação da política editorial. Entre as idéias centrais estão a do reforço da revista como um veículo para divulgação de trabalhos aplicados de Estatística, com ênfase nos resultados e não somente na metodologia, e o reforço da prática de captação de trabalhos nas reuniões promovidas pela ABE, tais como o SINAPE (Simpósio Nacional de Probabilidade e Estatística), as Escolas de Séries Temporais e Econometria (ESTE) e de Modelos de Regressão (EMR), e também as Reuniões Regionais. Para tanto, é preciso contar com a colaboração dos organizadores dessas reuniões para incluírem em suas chamadas de trabalhos o incentivo à submissão de artigos para publicação na RBEs.

A revista deverá passar por uma reformulação de sua programação visual. Estaremos trabalhando no novo formato editorial e gráfico nos próximos meses e portanto abertos a sugestões da comunidade de leitores.

Agradecemos o apoio recebido de todos e damos boas vindas aos colegas da ABE que se juntaram ao IBGE na tarefa de produção da RBEs.

Pedro Luis do Nascimento Silva
Editor Responsável pela RBEs

Estatísticas da Pobreza

Simon Schwartzman*

O IBGE e a Comissão Econômica das Nações Unidas para a América Latina -CEPAL- realizaram uma reunião em Santiago do Chile no início de maio de 1997 para examinar o estado da arte na produção de estatísticas sobre pobreza em diversas partes do mundo, e dar início a um "Grupo de Especialistas" (*Expert Group*[1]) que deverá dar continuidade a este trabalho, preparando recomendações que possam ser de utilidade para os institutos de estatística e demais interessados na produção, análise e uso de informações estatísticas a este respeito.

O tema da pobreza tem sido objeto de atenção cada vez mais intensa por parte dos governos, organizações internacionais e, conseqüentemente, institutos de estatística, que retomam, assim, uma tradição de estudos que foi importante na Inglaterra do Século XIX, mas que foi em grande parte substituída, nas décadas seguintes, pelas estatísticas sobre emprego e

[1] O *expert group* é uma modalidade de trabalho da Comissão de Estatística das Nações Unidas, que consiste em um conjunto de representantes de diferentes países que se associam para aprofundar um determinado tema. O *expert group* sobre estatísticas da pobreza é presidido pelo IBGE/Brasil.

* IBGE - Av. Franklin Roosevelt, 166 - Centro-RJ.

desemprego[2]. O fenômeno da pobreza, naturalmente, sempre existiu, mas sua interpretação tem variado muito ao longo do tempo[3]. Tradicionalmente, a condição de pobreza era entendida como algo natural, inevitável e inerente a uma parte grande, se não a maior, da humanidade, mas só se tornava objeto de preocupação de governantes e estudiosos dos fenômenos da economia e das populações quando os pobres, de alguma forma, saíam ou eram arrancados de sua situação de conformismo tradicional, e se transformavam em uma ameaça à ordem constituída. A obsessão inglesa com o tema a partir da Revolução Industrial, manifestada pela complexa legislação das *poor laws* e o grande debate a ela associado, tem como origem os efeitos combinados da grande expansão demográfica e o processo de esvaziamento dos campos, que jogaram milhares de pessoas nas cidades, em condições extremas de privação e pauperismo. A grande discussão, até o Século XIX, era se a pobreza era uma questão moral, conseqüência da falta de ética de trabalho e sentido de responsabilidade dos pobres, ou o efeito inevitável do desenvolvimento da economia industrial e de mercado. Malthus, como é sabido, explicava a pobreza pelo crescimento geométrico das populações, que não teria como ser acompanhado pelo crescimento da produção de alimentos, e jogava a responsabilidade da situação para os próprios pobres, que continuavam a procriar sem pensar nas conseqüências. Outros autores explicavam a pobreza pela preguiça, falta de caráter e excesso de bebida dos pobres. O termo *lumpenproletariat* foi utilizado pelo próprio Marx para descrever o que ele considerava a escória da Revolução Industrial, que não merecia os mesmos cuidados que os proletários, estes sim merecedores de toda a atenção, como portadores do futuro da humanidade.

---

[2] Sobre as estatísticas da pobreza na Inglaterra, veja E. P. Hennock, "The measurement of urban poverty: from the metropolis to the nation, 1880-1920", *Economic History Review,* 2nd ser., XL, 2 (1987), pp. 208-227. Veja também "Les Pauvres: comment les décrire, qu'en faire?", in Alain Desrosières, *La Politique des Grands Nombres*, Paris, Éditions de la Découverte, 1993, pp.271-276.

[3] Veja, a respeito Robert Castel, *Le métamorphoses de la question sociale - une chronique du salariat*, Paris, Fayard, 1995.

A suposição que se firmou nos países capitalistas mais desenvolvidos era que todas as pessoas que quisessem poderiam encontrar trabalho. Mas Marx já falara, no Século XIX, sobre o "exército de reserva" que, ainda que não se confundisse com o *lumpenproletariat*, seria uma característica permanente da economia capitalista. Os ciclos econômicos destruíam empregos de tempos em tempos, e a crise mundial de 1929 colocou milhões de pessoas em situação de pobreza, independentemente de seus valores morais e ética do trabalho. O tema do desemprego começou a ganhar atenção cada vez maior, primeiro como política social - os sindicatos reivindicavam, e os governos acabavam concordando em criar mecanismos de compensação - e mais tarde como coisa a ser medida e avaliada. Era necessário saber como andava o desemprego, não somente para atender aos necessitados, mas também como um indicador importante do próprio nível da atividade econômica. A diferença principal entre os estudos de pobreza do Século XIX e as estatísticas de desemprego do Século XX é que a pobreza era vista como uma característica das pessoas, mesmo que elas pudessem eventualmente mudar, enquanto que o desemprego era visto como um fenômeno estrutural temporário, ainda que em muitos casos esta situação de curto prazo acabasse sendo, na prática, permanente. É assim que a metodologia de mensuração do desemprego, desenvolvida pela Organização Internacional do Trabalho e aplicada na maioria dos países do mundo, e inclusive pelo IBGE, define o desempregado como alguém que faz parte da população economicamente ativa, mas que está, temporariamente, sem trabalho e buscando ativamente uma alternativa, excluindo, desta forma, as pessoas que estão fora do mercado de trabalho de forma permanente.[4]

[4] Veja, sobre a introdução das estatísticas sobre desemprego nos Estados Unidos, "Chômage et inegalités: comment bâtir des objets nouveaux", in A. Desrosières, op. cit., pp. 244-245 .

Fora dos países industrializados a pobreza continuou existindo em grande escala e em muitos casos se agravando, mas não foi nestes países que os modernos sistemas de estatísticas públicas se desenvolveram. Na América Latina, nos anos 50 e 60, o tema da pobreza ressurgiu sob o rótulo de estudos sobre "marginalidade", sobretudo em organizações acadêmicas ou voltadas para a mobilização popular[5], em três vertentes principais. Uma, de inspiração marxista, tratava de interpretar os fenômenos de pobreza em termos do conceito de "exército industrial de reserva". Os pobres da América Latina, que se deslocavam em grandes números dos campos para as cidades, repetindo de alguma forma, séculos depois, a transição demográfica da Revolução Industrial européia, seriam uma criação do próprio capitalismo, que dependeria de sua existência para manter seus altos níveis de lucro e exploração. A premissa não estava de todo errada, já que, de fato, a explosão demográfica, a introdução de técnicas modernas de produção agrícola e a geração de empregos nas cidades de fato explicavam a grande expansão da pobreza urbana, que tornava mais

[5] Veja, entre outros, Acedo Mendoza, Carlos, *América Latina, Marginalidad y subdesarrollo*, Caracas: Fondo Editorial Común, 1974; Germani, Gino, *El concepto de marginalidad: significado, raíces históricas y cuestiones teóricas, con particular referencia a la marginalidad urbana*, Buenos Aires: Ediciones Nueva Visión, c1973; DESAL, *Marginalidad en América Latina; un ensayo de diagnóstico.* Santiago, DESAL, 1969 [c1967]; Margulis, Mario, *Migración y marginalidad en la sociedad argentina,* Buenos Aires, Paidós, 1968; Mattelart, Armand. [y] Manuel A. Garretón, Integración nacional y marginalidad: un ensayo de regionalización social de *Chile, Santiago de Chile, Editorial* del Pacifico, 1965; Núñ, José, Miguel Murmis [y] Juan Carlos Marín, La marginalidad en América Latina; informe preliminar, Buenos Aires: Instituto Torcuato di Tella, Centro de Investigaciones Sociales, 1968; Quijano, *Aníbal, Imperialismo y "marginalidad" en América Latina*, Lima: Mosca Azul Editores, 1977; United Nations. Economic Commission for Latin América, *Bibliografía sobre marginalidad social,* Santiago de Chile, La Biblioteca, 1973; Vekemans, Roger, Ismael Silva [y] Jorge Giusti, *La marginalidad en América Latina: un ensayo de conceptualización*, Santiago de Chile, Centro para el Desarrollo Económico y Social de América latina (DESAL), 1970.

visível, e potencialmente mais explosiva, a tradicional pobreza rural.[6] (a dúvida era se a implantação de uma ordem socialista conseguiria reverter este processo.) A outra vertente dos estudos de marginalidade era a vertente católica, que se confundia em parte com a marxista, mas tinha um tom muito mais claramente ético e moral. A pobreza era vista como produto da exploração, não de um sistema econômico impessoal, mas de classes dominantes gananciosas e desprovidas dos dons da caridade e da solidariedade. A mensuração da pobreza equivaleria à mensuração dos níveis de iniqüidade e injustiça existentes em uma sociedade, a serem reduzidos pelo arrependimento dos ricos e a mobilização dos pobres. Uma terceira vertente vinha do norte, sobretudo dos Estados Unidos, e interpretava o que ocorria em termos culturais. A pobreza era, nesta perspectiva, sobretudo uma questão de atraso cultural ou psicológico, que fazia com que as pessoas não tivessem iniciativa, não fizessem uso de seus recursos, e não buscassem melhorar de vida. O processo de modernização que se espalhava do Norte para o Sul, e do Ocidente para o Oriente, era visto sobretudo como um processo de difusão de valores e atitudes, a serem transportados pelos meios de comunicação de massas e consolidados pelos sistemas educacionais.[7]

Este estoque de teorias sobre a pobreza e suas possíveis soluções não se alterou substancialmente desde então[8], mas o tema da pobreza readquiriu importância, primeiro, porque nenhum dos encaminhamentos propostos nos anos anteriores funcionou, e, segundo, talvez o mais importante, porque os problemas da pobreza começaram a se manifestar

[6] A pobreza urbana, no entanto, sempre existiu nas grandes cidades latino-americanas, que se formaram sobretudo como centros administrativos dos impérios coloniais espanhol e português, que sempre atraíram e mantiveram multidões de pessoas vivendo das sobras do poder político - as "classes dangereuses" estudadas por muitos historiadores no Brasil e em outros países.

[7] Os principais autores associados a estas teorias sobre modernização eram Lucian Pye, David Lerner e Alex Inkeles.

com intensidade nos países industrializados mais avançados, onde ele parecia ter deixado de existir. Nos países do chamado "Terceiro Mundo" o que se presenciou foi que, em quase toda parte, mesmo quando a economia se desenvolvia, como ocorreu no Brasil, a pobreza continuava existindo. Muitos países, sobretudo os da África mas também os da América Latina, viram suas economias estagnarem ou entrarem em processo de involução, ao mesmo tempo em que suas formas mais tradicionais de organização social e econômica eram destruídas, aumentando os níveis de pobreza absoluta, violência urbana e situações intermináveis de conflito armado nas áreas rurais. Nos países ricos, a pobreza aparece, sobretudo, com as novas ondas de migração internas - como a dos negros nos Estados Unidos para as grandes cidades do norte e nordeste - e externas, das antigas colônias para as metrópoles, na França e Inglaterra, ou dos países da Europa Central, do Mediterrâneo e da Península Ibérica para a Alemanha e outros países da Europa Ocidental. Se, em um primeiro momento, estes imigrantes representavam uma mão-de-obra barata e disposta a realizar tarefas não-qualificadas que as populações locais rechaçavam, eles passaram a competir, depois, pelos benefícios dos sistemas de previdência social e pelos empregos, ambos em processo de encolhimento. Além dos imigrantes, a crise do estado de bem-estar social e as transformações tecnológicas na economia criaram novos grupos em situação de pobreza nos países desenvolvidos, sobretudo entre idosos, jovens e antigos empregados em atividades econômicas tradicionais, que têm dificuldades em se reempregar.

As estatísticas de pobreza que se desenvolveram nos últimos anos podem ser classificadas em dois tipos principais: aquelas que buscam medir a pobreza absoluta, ou seja, identificar as pessoas que estão abaixo de um padrão de vida considerado minimamente aceitável, e as que medem a pobreza relativa, ou seja, que buscam identificar as pessoas que

[8] Ao longo dos anos 60 e 70, as teorias da marginalidade foram substituídas, na América Latina, pelas da "dependência", que procuravam buscar explicações e eventuais soluções para a pobreza no plano das relações internacionais.

tenham um nível de vida baixo em relação à sociedade em que vivem[9]. Tanto em um como em outro caso, a renda monetária é utilizada normalmente como indicador. No caso da pobreza relativa, trata-se de identificar as pessoas que se situam abaixo de um ponto qualquer na distribuição de renda, definido arbitrariamente. No caso da pobreza absoluta, trata-se de identificar as pessoas cujos rendimentos são inferiores ao necessário para adquirir um conjunto mínimo de bens e serviços considerados indispensáveis. Uma variante em relação à pobreza absoluta é a chamada "metodologia das necessidades básicas não satisfeitas" - neste caso, trata-se de identificar as pessoas que de fato não conseguem satisfazer necessidades essenciais como habitação, nutrição, educação, saúde, etc., independentemente da renda disponível.

A simplicidade aparente destes conceitos desaparece rapidamente quando eles são levados à prática. Primeiro, como medir a renda. A fonte usual para estas informações são as pesquisas domiciliares anuais, como a PNAD - Pesquisa Nacional por Amostra de Domicílios - mas, se se pretende descer ao nível de municípios ou distritos, a única fonte de informação disponível são os censos decenais, de periodicidade longa, e limitados a informações sucintas. Populações mais pobres muitas vezes possuem rendas não monetárias, produzem para o autoconsumo, ou têm acesso a transferências e doações familiares que não aparecem nas estatísticas usuais. Famílias de composição diferente têm gastos distintos. O custo de vida varia de uma região a outra no mesmo país. E, a rigor, há que decidir se a renda deve incluir ou não benefícios não monetários na área social como educação, saúde, habitação, transportes subsidiados, e outros. Depois, o conceito de "necessidade básica" ou "conjunto mínimo de

[9] Veja, para um exame detalhado destas diferentes metodologias, "Poverty measurement: present status of concepts and methods", documento preparado por Luis Beccaria para a CEPAL para o "Seminar on Poverty Statistics", Santiago, maio de 1997. Veja também Juan Carlos Feres, "Notas sobre la Medición de la pobreza segundo el método del ingreso", *Revista de la CEPAL,* 61, Abril, 1997, 119-134; e Sônia Rocha, "On statistical mapping of poverty: social reality,

bens" também é problemático, e sujeito a grandes variações culturais. Um critério utilizado tem sido a definição de um volume mínimo de calorias ingeridas pelas pessoas, considerado indispensável. Uma vez estabelecido este mínimo, deve-se ver o que as pessoas (ou as famílias) estão ingerindo, e converter em calorias por alguma tabela. Na impossibilidade de medir diretamente a ingestão de alimentos de cada família (isto só foi feito uma única vez no Brasil, nos anos 70, com a pesquisa ENDEF - Estudo Nacional da Despesa Familiar -, de onde derivam as tabelas de conversão utilizadas até hoje), procura-se medir o custo de uma cesta básica de alimentos suficientes para este total de calorias, fazendo uso das pesquisas de índice de preços, e depois comparando os valores encontrados com as informações disponíveis sobre renda monetária, definindo, desta maneira, uma "linha de pobreza" para determinada região e momento.

Este é só um resumo dos procedimentos necessários para a mensuração da pobreza absoluta, mas é suficiente para mostrar o grande número de suposições e mesmo decisões arbitrárias que precisam ser adotadas a cada passo. Estas suposições e decisões não invalidam, necessariamente, os números obtidos ao final do processo, que precisam ser avaliados sobretudo em termos de sua consistência com outras informações relevantes, e por outros procedimentos estatísticos conhecidos. Mas elas levantam três tipos de questões, que merecem ser examinadas em mais profundidade.

A primeira é que o resultado final de uma mensuração tão complexa pode não ser muito diferente do que seria obtido por um método muito mais simples e direto. O Banco Mundial, por exemplo, em alguns de seus estudos, define como pobres as pessoas que ganham menos do que um dólar por dia. É um número arbitrário, mas não necessariamente pior do

concepts and measurement" , documento de trabalho preparado para a reunião do Expert Group on Poverty Statistics, Santiago, maio de 1997.

que medidas muito mais complexas[10]. A segunda é que dados sobre pobreza obtidos em um país dificilmente podem ser comparados com os de outros, produzidos por metodologias distintas, a partir de suposições e decisões operacionais também distintas e independentes. A terceira é que estes dados constituem uma base extremamente precária sobre a qual os países possam definir "linhas de pobreza" oficiais, como referência para suas políticas.

Ainda que estas dificuldades sejam conhecidas, vários países têm adotado linhas de pobreza oficiais, que cumprem inúmeras finalidades. Quando aplicadas a indivíduos ou famílias, elas servem de critério para distribuição de auxílios e benefícios sociais de vários tipos; quando aplicadas a localidades geográficas ou regiões, elas podem servir de base para a definição de prioridades em políticas de investimentos públicos; quando aplicadas a populações específicas, elas podem ser utilizadas para políticas compensatórias; e podem servir de *benchmarks* para o acompanhamento de políticas nacionais de redução da pobreza. Existem, no entanto, vários inconvenientes, que fazem com que outros países prefiram não possuir uma linha de pobreza oficial. O primeiro é o caráter necessariamente arbitrário de qualquer linha de pobreza: diferentes suposições e decisões metodológicas podem conduzir a valores distintos, sem que existam critérios objetivos para optar entre eles. Segundo, uma vez definida uma linha de pobreza oficial, e utilizada para políticas distributivas, ela fica associada a um grande número de interesses, que passam a se opor ao aperfeiçoamento ou modificação dos critérios utilizados inicialmente, pela perda de benefícios ou aumento de gastos públicos que uma modificação destes números pode significar. Austrália e Estados Unidos são exemplos de países que adotam linhas de pobreza

[10] É uma abordagem semelhante à utilizada pela revista *The Economist,* de utilizar o preço do Big Mac em dólares para comparar os valores relativos das moedas dos diferentes países. Este critério é também utilizado no "Relatório de Desenvolvimento Humano" publicado periodicamente pelo PNUD.

desde a década de 60, e têm encontrado dificuldades em reformulá-las, apesar de reconhecerem suas limitações.

No Brasil, diversos pesquisadores têm utilizado os dados do IBGE para estimativas de linhas de pobreza, que podem ser adotadas pela administração pública em suas políticas, mas não existe uma linha de pobreza oficial.[11] A CEPAL, com base em processamento próprio das pesquisas de domicílios dos diversos países, afirma que a pobreza na América Latina teria baixado de 41 a 39% da população entre 1990 e 1994, enquanto que o número de indigentes teria baixado de 18 a 17%.[12]

A grande heterogeneidade dos problemas sugere que os exercícios de mensuração global da pobreza, e a eventual opção por uma linha de pobreza qualquer, devem estar associados à identificação dos diferentes tipos de pobreza existentes em um país, que requerem políticas sociais diferenciadas. De alguma forma, o estudo mais aprofundado das diferentes condições de pobreza implicam uma volta aos antigos dilemas sobre as causas individuais ou estruturais da pobreza. É necessário poder distinguir aquelas situações que resultam do contexto maior de que os grupos mais desfavorecidos participam, daquelas situações em que os problemas da pobreza devem ser tratados no nível dos próprios grupos afetados. Em um extremo, existem situações em que o mercado de trabalho não paga baixos salários, ou não abre possibilidades de emprego, por exemplo, por problemas de competitividade; em outro extremo, os salários baixos estão associados a baixos níveis educacionais nos trabalhadores, o que requer uma ação específica sobre o sistema educacional. Estes dois extremos não são excludentes, já que uma mudança na oferta global de pessoas bem

[11] Estimativas sobre o número de "pobres" no Brasil tem variado, conforme as diferentes metodologias, de 24 a 42 milhões de pessoas. O número de "indigentes" tem também variado em escala semelhante.

[12] CEPAL, *Panorama Social de América Latina, 1996*. Santiago de Chile, CEPAL, 1997.

treinadas deve afetar tanto a competitividade da economia quanto a remuneração recebida pelos diferentes grupos sociais.

Em muitos casos as situações de pobreza estão associadas a um conjunto complexo que os antropólogos costumam denominar de "cultura", e que os médicos denominam de "síndrome". A idéia, em ambos os casos, é que não se trata de problemas de causação simples e tratamento também simples, através da manipulação de uma ou duas variáveis, mas de situações muito mais complicadas. Existe toda uma tradição de estudos antropológicos sobre a pobreza que trabalham com o conceito de cultura, seja em populações marginalizadas na América Latina (como os famosos estudos de Oscar Lewis sobre a família Sánchez no México, dos anos 60), seja em relação à população dos *ghettos* urbanos nos Estados Unidos, ou imigrantes na Europa. Uma contribuição importante destes estudos é que eles permitem entender as estratégias de sobrevivência das populações pobres; outra contribuição é o entendimento sobre como os recursos públicos e privados, orientados para a solução dos problemas de pobreza chegam efetivamente aos setores interessados, e sobre as eventuais dificuldades de adoção de políticas que poderiam mudar as condições de vida destas populações – pela educação, por exemplo –, mas que muitas vezes não conseguem obter os resultados esperados, ou têm resultados negativos, mesmo quando existem recursos disponíveis.

Um problema comum a estes estudos mais qualitativos é o risco de que os problemas da pobreza terminem sendo vistos como insolúveis, ou até mesmo como preferências "culturais" de determinados grupos, que deveriam ser deixados à sua própria sorte. Daí a importância crescente dos estudos que tratam de entender as síndromes de pobreza e os elementos culturais a elas associados não como uma característica intrínseca dos grupos afetados, mas como o resultado de um processo mais complexo de interação entre estes grupos e a sociedade mais ampla, através do qual as identidades, percepções e preconceitos são construídos e reforçados.

A conclusão é simples, mas nem por isto menos importante. Os estudos globais sobre situações de pobreza, realizados a partir de estatísticas de grande representatividade e cobertura, precisam estar acompanhados de estudos em profundidade sobre grupos e situações específicas, sem os quais políticas adequadas de redução da pobreza se tornam muito difíceis de ser implementadas e avaliadas. As formas destes estudos variam, e incluem desde *surveys* detalhados, como a Pesquisa de Padrão de Vida, realizada em 1996-7 pelo IBGE, como trabalhos mais qualitativos, realizados por pesquisadores acadêmicos ou associados a instituições públicas e privadas que atuam na área da redução dos problemas da pobreza. Neste espectro amplo de trabalho, cabe aos órgãos nacionais de estatística, como o IBGE, proporcionar os parâmetros quantitativos mais gerais do conjunto, e ajudar a viabilizar estudos em profundidade dentro de um leque bastante amplo de metodologias.

Uso do Sistema Dia para a Detecção e Correção Automática de Erros nos Dados do Questionário Básico do Censo Demográfico de 1991

Luís Carlos de Souza Oliveira, Laura Baridó Indá,
Rita Luzia Aguiar Lima e Zélia Magalhães Bianchini[1]

1. INTRODUÇÃO

O Censo Demográfico é uma das mais importantes pesquisas que o IBGE realiza. Seus resultados servem aos mais variados tipos de usuários, para os mais diversos propósitos analíticos, de planejamento e tomada de decisão. Além disso, os dados do censo servem de base para todo um conjunto de outras pesquisas desenvolvidas durante a década. Em muitos casos, a dependência dos dados provenientes do Censo Demográfico é crucial para essas pesquisas. Veja-se, por exemplo, o caso das pesquisas

[1] IBGE - Diretoria de Pesquisas.
Av. Chile, 500 - 10º andar - Rio de Janeiro.

domiciliares por amostragem do IBGE, que baseiam a seleção e estimação de suas amostras nas informações provenientes do censo.

Por essa razão, o conhecimento preciso dos dados e informações resultantes do Censo Demográfico de 1991 é indispensável não só para aqueles que dão uso imediato a essas informações, mas também para aqueles que, ao longo dos anos vindouros, terão o censo como base para o planejamento de suas pesquisas. Para que esse conhecimento seja não só preciso, mas completo, é fundamental ter registrado e devidamente documentado como os dados coletados foram tratados nas diversas etapas de processamento, antes de serem armazenados em bases de dados, a partir das quais são recuperados para a produção das tabelas e dos demais produtos de divulgação (arquivos de pronta entrega, sistemas de recuperação *on-line*, etc.), bem como para as futuras análises a que servirão de base.

Na coleta das informações do Censo Demográfico de 1991, foram usados dois modelos de questionário: um questionário básico (denominado CD 1.01) aplicado nas unidades não selecionadas para a amostra contendo perguntas referentes às características que foram investigadas para 100% das unidades domiciliares, e um questionário de amostra (denominado CD 1.02) contendo, além das perguntas que constam do questionário básico, outras perguntas mais detalhadas sobre características do domicílio e das pessoas.

Foram pesquisadas no questionário básico as seguintes variáveis: para domicílios, a espécie (V201), localização (V202), abastecimento de água (V203), escoadouro (V204), uso do escoadouro (V205), condição de ocupação (V206), cômodos (V207), cômodos servindo de dormitório (V208), banheiros (V209), destino do lixo (V210); para pessoas, sexo (V301), parentesco ou relação com o chefe do domicílio (V302), mês e ano de nascimento, idade presumida (caso não soubesse o mês e ano de nascimento) (V303/V304), alfabetização (sabe ler e escrever ou não sabe) (V305); e, adicionalmente para os chefes; última série concluída (V306), grau (V307) e rendimento mensal bruto (V308). Além dessas variáveis foram criadas, também, algumas outras para serem utilizadas na correção

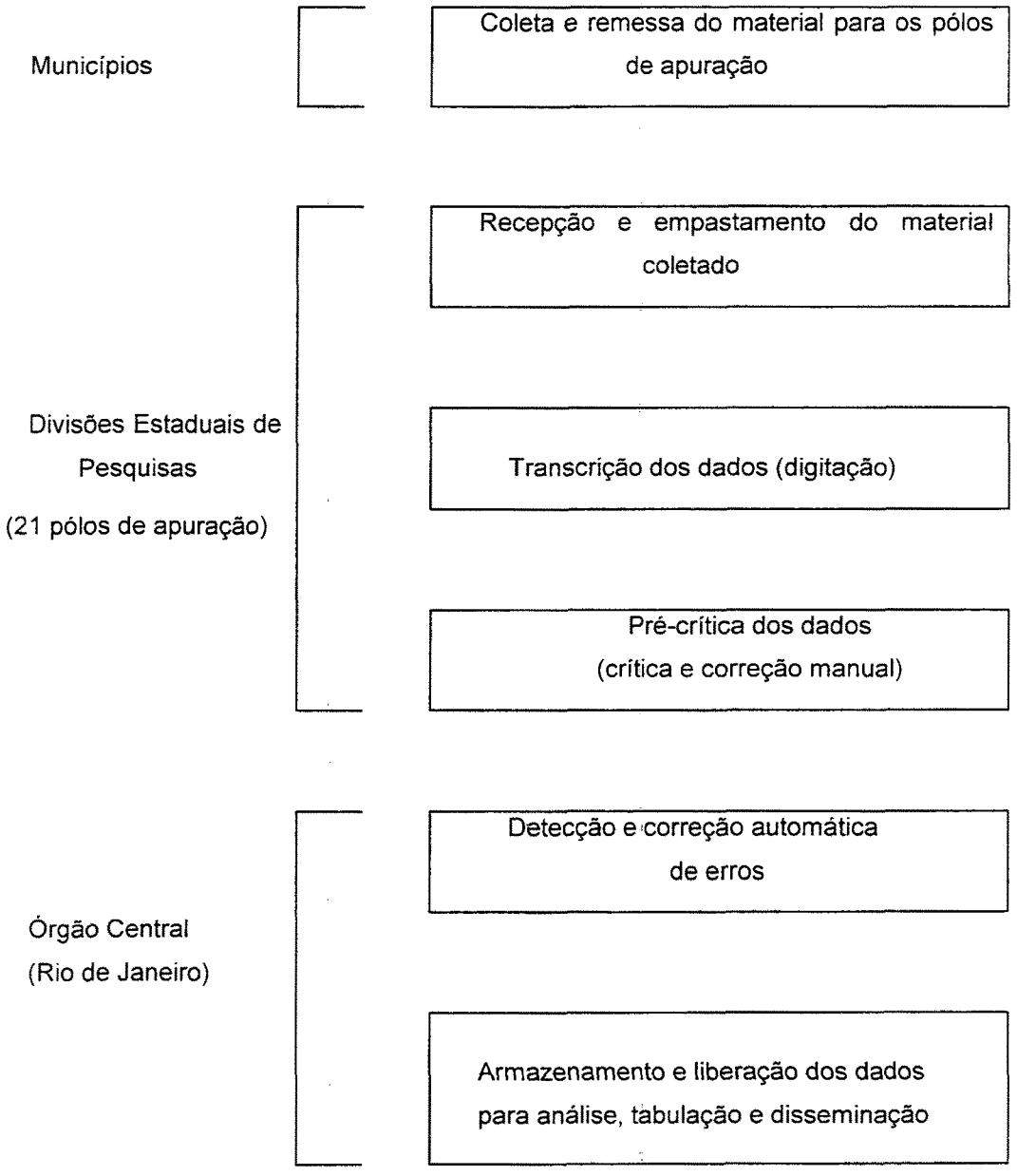
automática, a saber: V2203/V2204 (indicadora de rede geral de água e esgoto, respectivamente), V1062 (tipo do setor), V7100 (total de pessoas), V3043 (idade calculada em anos ou meses completos) e V3080 (indicadora do tipo de rendimento).

Durante o planejamento do Censo Demográfico de 1991, visando a agilizar a divulgação dos dados referentes às características investigadas para 100% das unidades domiciliares, foi definido que para cada questionário da amostra - CD 1.02 seria "criado" um CD 1.01 com as informações correspondentes, transcritas do CD 1.02. Desse modo, as atividades subseqüentes de apuração passaram a ser planejadas de forma isolada para cada um dos questionários. Isto implica que o Censo Demográfico de 1991 pode ser considerado como composto por duas pesquisas independentes[2]: uma pesquisa censitária que abrange as variáveis investigadas no questionário básico CD 1.01 e uma pesquisa por amostra que abrange as variáveis investigadas no CD 1.02.

As grandes etapas de coleta e apuração do CD 1.01 podem ser visualizadas na Figura 1, a saber: coleta e remessa do material para os pólos de apuração, recepção e empastamento do material, transcrição dos dados para formato digital (digitação), pré-crítica (crítica e correção manual de impossibilidades, quantidade e parte das de incompatibilidade), detecção e correção automática dos erros e armazenamento e liberação dos dados para análise, tabulação e disseminação. Este artigo enfoca aspectos relativos à fase de detecção e correção automática dos erros[3].

[2] Ver Albieri e Bianchini (1993).
[3] Este artigo tem como base o documento de Oliveira e outros (1994) que descreve de forma detalhada as especificações adotadas para o tratamento dos dados do questionário básico - CD 1.01.

Figura 1 - Visão geral do processo de apuração do CD 1.01



De modo geral, o processamento de pesquisas envolve a definição de regras de crítica que, quando não satisfeitas, apontam erros nos dados. "Correções" ou imputações são necessárias de modo a tornar consistentes

ou aceitáveis os dados de cada questionário. A operação de crítica e correção de erros detectados nos dados investigados no CD 1.01 envolveu duas fases de trabalho: pré-crítica e correção automática. A pré-crítica correspondeu à primeira fase do processamento dos dados do CD 1.01, durante a qual os erros detectados por programa eram corrigidos manualmente.

A correção automática foi a segunda fase do processamento dos dados, na qual erros detectados eram corrigidos através de imputação por programa. A utilização de métodos para detecção e correção automática de erros tem os seguintes objetivos:

- economia de tempo no processamento dos dados, eliminando a necessidade de pessoas especialmente treinadas para efetuar correções nos questionários com respostas rejeitadas; e
- obtenção de maior homogeneidade nos critérios de correção das inconsistências detectadas, não permitindo que as alterações nos valores das variáveis suspeitas[4] de erro sejam decididas de formas distintas por pessoas diferentes.

Um ponto importante a destacar diz respeito à decisão de quais variáveis deveriam ser corrigidas de forma automática ou manual. Um princípio que norteou a elaboração dos planos da pré-crítica e da correção automática foi o de deixar o mínimo possível de críticas de incompatibilidade ou de inconsistência para serem aplicadas na etapa manual descentralizada.

As críticas tratadas na fase de pré-crítica compreenderam as de impossibilidade, que verificavam se os valores digitados correspondiam aos códigos válidos para cada quesito investigado, incluindo o preenchimento de quesitos obrigatórios; as críticas de quantidade, que faziam a conferência da seqüência da identificação, do número dos questionários dentro de cada pasta, do número da pessoa dentro de cada questionário e dos totais de pessoas; algumas críticas de incompatibilidade que verificavam inconsistências entre registros de pessoas e domicílio, e outras

cuja correção dependia do acesso aos instrumentos de coleta (questionários, folhas de coleta, etc.).

2. A NOVA METODOLOGIA PARA DETECÇÃO E CORREÇÃO AUTOMÁTICA DOS ERROS

O uso de correção automática de erros de consistência foi iniciado no IBGE no Censo Demográfico de 1980, tendo sido objeto de experimentação no Censo Demográfico de 1970 (somente nos questionários da não-amostra). A correção automática do Censo Demográfico de 1980 utilizou dois métodos: a correção determinística e o método da Matriz Dinâmica com variáveis de restrição (*Hot-Deck*)[5].

Uma preocupação fundamental num processo de correção automática é a não geração de novos erros após a imputação. Infelizmente, a metodologia empregada em 1980 não permitia assegurar, *a priori*, que as correções efetuadas num dado registro deixariam o mesmo livre de novas inconsistências. Como conseqüência, os lotes de apuração tiveram que passar pelo sistema de crítica e correção automática mais de uma vez, criando assim um ciclo de processamento indesejável numa pesquisa do porte do censo. Outro problema decorrente da metodologia adotada no Censo de 80 foi a elevada complexidade do trabalho de desenvolvimento do sistema computacional implicado pelo plano de crítica e correção, ante a inexistência naquela época, de ferramentas de automatização do desenvolvimento de sistemas no IBGE, e a natureza exploratória do processo de teste dos programas desenvolvidos.

Desde a sua implantação em 1987, a Divisão de Metodologia (atual Departamento de Metodologia) direcionou esforços para estudar métodos e sistemas de apuração de pesquisas, motivada pelos dilatados prazos de desenvolvimento e execução de sistemas de crítica, bem como os altos custos e insegurança quanto à qualidade dos dados obtidos. Foram

[4] Variável suspeita é a que possui valor inválido ou participa em alguma inconsistência detectada.
[5] Ver Metodologia do Censo Demográfico de 1980.

examinadas alternativas de metodologia e sistemas *software* para apuração de pesquisas, particularmente para correção ou imputação automática de dados, que pudessem ser aplicadas no processamento dos dados do Censo Demográfico de 1991. Além disso, uma avaliação da experiência do censo anterior (1980) revelou que o uso de métodos de correção automática foi considerado, à época, como um sucesso e representou um avanço. Entretanto, nos dias de hoje, os métodos adotados em 1980 seriam de difícil aplicação, por serem extremamente custosos em termos de recursos para a especificação, o desenvolvimento, o teste e a validação dos sistemas requeridos para sua implementação.

Esse estudo conduziu a um exame mais detalhado das experiências vivenciadas pelos países mais desenvolvidos no tratamento dessa questão, e a um aprofundamento no exame das experiências do Canadá, baseada no trabalho pioneiro de Fellegi & Holt (1976), e da Espanha, a partir do trabalho de Rubio e Criado (1988). Estes últimos desenvolveram um sistema generalizado para detecção e correção automática de erros em dados qualitativos, baseado na metodologia proposta por Fellegi & Holt (1976), denominado DIA - *Detección e Imputación Automática de errores para datos cualitativos*. Tal sistema foi cedido ao IBGE em maio de 1989 para estudo e testes, e foi imediatamente aplicado para a realização de um experimento de processamento dos dados do CD 1.01 coletados durante o Censo Experimental de Limeira de 1988. Esse experimento está descrito com detalhes em Silva e outros (1990). Após realizados todos os testes e estudos necessários, foi decidida a utilização do sistema DIA para o processamento do CD 1.01 do Censo Demográfico de 1991.

Em seu artigo, Fellegi & Holt (1976) deram uma nova abordagem, baseada num tratamento matemático rigoroso, à questão da análise de um conjunto de regras de crítica definidas para uma pesquisa. Essa nova abordagem possibilitou integrar perfeitamente os processos de detecção e correção de erros em apenas um único ciclo de processamento. Os princípios básicos no qual se baseia essa metodologia são os seguintes:

- os dados de cada registro devem satisfazer todas as críticas, alterando o menor número possível de variáveis, procurando assim manter a maior quantidade possível de informação original;

- não se faz necessário especificar regras de imputação (correção), ao contrário do processo *Hot-Deck*, as quais são deduzidas *automaticamente das regras de crítica (detecção)*; isto permite assegurar que os registros não estarão em condição de erro após a imputação;

- o processo de imputação procura manter, na medida do possível, as distribuições de freqüências marginais e conjuntas das variáveis, baseando-se nas distribuições dos dados "bons" (aqueles que passaram pela crítica sem serem rejeitados).

Para a implementação da metodologia de Fellegi & Holt, é necessário escrever as regras de crítica especificadas para a pesquisa na **forma normal**[6]. As regras escritas nessa forma são denominadas críticas explícitas. A partir dessas regras de crítica pode-se deduzir logicamente outras críticas, as quais são denominadas críticas implicadas. Dentre essas regras de crítica só são de interesse as regras de crítica implicadas essencialmente novas que, em conjunto com as regras explícitas, constituem o **conjunto completo das regras de crítica**. Fellegi & Holt (1976) demonstraram que sempre se pode obter um conjunto completo de regras de crítica por meio de um algoritmo.

Dispondo do conjunto completo das regras de crítica, para corrigir um registro detectado como inconsistente ou errôneo, basta imputar ao menos um campo ativo[7] em cada crítica do conjunto completo que tenha rejeitado o registro. A seleção faz-se empregando o princípio da menor modificação possível dos dados originais, escolhendo o menor número possível de variáveis a modificar em cada registro inconsistente. Para um registro inconsistente, o menor conjunto de variáveis que contempla as variáveis

---

[6] Uma regra de crítica é escrita na forma normal se for da seguinte forma:
Se ocorrer Vx (subconjunto de códigos da Vx) e ... e Vy (subconjunto de códigos da Vy) $\Rightarrow$ incompatibilidade.

[7] Campo ativo ou explícito é aquele que é incluído na especificação de uma regra de crítica.

ativas nas críticas do conjunto completo que rejeitaram esse registro é chamado de conjunto mínimo de campos a imputar.

Uma vez selecionadas as variáveis a serem imputadas em um certo registro inconsistente, é feita a seleção do código a imputar em cada uma dessas variáveis, a partir do dicionário de valores válidos. Fellegi & Holt (1976) demonstraram que é sempre possível encontrar ao menos um código válido para cada variável selecionada, de forma que imputando tal código, o registro fique "corrigido" e não seja mais rejeitado por nenhuma das regras de crítica do conjunto completo.

3. O SISTEMA DIA

O sistema DIA - *Detección e Imputación Automática de errores para datos cualitativos* - é um pacote computacional desenvolvido no *Instituto Nacional de Estadística (INE)* da Espanha. O DIA é baseado na metodologia de Fellegi & Holt (1976) para a detecção e correção automática de erros, em sua versão para o tratamento de dados categóricos ou qualitativos[8]. Esse sistema permite não só a utilização de correção probabilística sugerida na metodologia de Fellegi & Holt, como também a de correção determinística para o tratamento de erros sistemáticos, garantindo, porém, a consistência entre as críticas de incompatibilidade e as regras para imputação determinística (RIDs)[9].

O sistema DIA opera em ambiente centralizado tipo *mainframe* IBM e requer dados armazenados em arquivos seqüenciais zonados, com um único tipo de registro de tamanho fixo com variáveis categóricas ou qualitativas[10]. O sistema é constituído de dois subsistemas: o de especificação e o de tratamento, cada um formado por vários módulos.

---

[8] Uma descrição mais detalhada pode ser encontrada em Rubio e Criado (1988) e Silva e outros (1990).

[9] Uma regra para imputação determinística é escrita da seguinte forma:
Se Vx (subconjunto de códigos de Vx) e...e Vy (subconjunto de códigos de Vy) $\Rightarrow$ Vx = subconjunto de códigos determinados de Vx.

[10] Ver Hanono (1993).

3.1. Subsistema de Especificação

As funções desenvolvidas pelos módulos que compõem o subsistema de especificação podem ser resumidas em:

- análise sintática das especificações;
- tradução das especificações para o formato interno do sistema;
- geração do conjunto completo de regras de crítica;
- geração e compilação dos programas a serem usados pelo subsistema de tratamento dos dados;
- criação dos arquivos de regras de críticas e auxiliares, no formato a ser usado no subsistema de tratamento dos dados; e
- criação da estratégia de depuração.

Tendo como entrada o plano de depuração especificado para a pesquisa, o sistema DIA procede a uma análise e passa então a gerar as estruturas de dados e os programas necessários para a depuração dos dados, cujo conjunto é denominado "Aplicação DIA".

O subsistema de especificações é formado basicamente pelo analisador de regras, cuja função consiste em comprovar a consistência interna dos conjuntos de regras de incompatibilidades e de regras de imputação determinística, eliminando redundâncias existentes. Esse módulo também verifica a consistência do conjunto de regras de imputação determinística em relação ao conjunto de regras de crítica explícitas.

O plano de depuração divide-se em duas partes:

- especificação do usuário contendo
  - lista das variáveis a depurar e respectivos conjuntos de códigos possíveis;
  - regras de crítica escritas na forma normal; e
  - regras de imputação determinística.
- estratégias de depuração.

Algumas das estratégias de depuração possíveis são definidas a seguir:

- fixação de variáveis - estratégia bastante utilizada, na qual algumas variáveis já depuradas em fases anteriores são

impedidas de receber alterações durante o processo de correção automática ou, ainda, para variáveis auxiliares que participam do conjunto de regras de crítica apenas no intuito de auxiliar a depuração de outras variáveis;

- ponderação de variáveis - esta estratégia permite representar o grau de desconfiança para cada variável a imputar, com pesos variando de 1 a 10; quanto menor o peso, maior o grau de confiabilidade da variável, permitindo definir que determinadas variáveis tenham menos chance de serem escolhidas para imputação que outras.
- critérios de imputação de códigos - é necessário especificar o tipo e a composição da distribuição de freqüências, de onde selecionar códigos para imputação, bem como o método de seleção do código; a imputação em uma variável é feita através da seleção de um código; caso haja vários candidatos, essa seleção é probabilística a partir de uma determinada distribuição de freqüências de registros não suspeitos que se procura manter; essa distribuição pode ser marginal ou conjunta, de acordo com a especificação definida para cada variável. Para o cálculo da distribuição de freqüências do sistema DIA, são contabilizados todos os registros bons (sem erros) e aqueles registros errôneos em que as variáveis envolvidas são não suspeitas.

O sistema DIA possui opções que permitem escolher entre a FNS (freqüência de registros não suspeitos ou a FNS* (freqüência de registros não suspeitos após a atuação das regras de imputação determinística para selecionar códigos a imputar).

Além da especificação da distribuição de freqüência marginal ou conjunta e do tipo de freqüência FNS ou FNS* é ainda necessária uma outra especificação que diz respeito ao método de seleção do código a imputar. O método recomendado é o de seleção com probabilidade proporcional à distribuição de freqüência escolhida. Entretanto o sistema oferece uma alternativa em que a seleção do código pode ser efetuada pela

máxima diferença[11] na distribuição de freqüência de variáveis com imputação marginal.

3.2. Subsistema de Tratamento dos Dados

O subsistema de tratamento dos dados é composto pelos módulos de detecção e de imputação. Para a detecção dos registros com erro o sistema utiliza o dicionário, que contém a descrição das variáveis com os respectivos códigos válidos, o conjunto das regras de incompatibilidade e as críticas derivadas de RIDs.

Um registro é considerado com erro quando:

- alguma de suas variáveis tem um código inválido; e/ou
- falha em alguma das regras de crítica ou alguma RID.

O módulo de detecção separa e gera arquivos dos registros sem erro (bons) e dos registros com erro (maus), identificando nesses últimos os campos a corrigir e as regras que falharam. Caso haja regras de imputação determinística (RIDs), é necessário aplicar o módulo de imputação determinística aos registros "maus". Para cada registro com erro, o sistema verifica a ocorrência de falha em alguma RID e, conforme o caso, efetua as imputações especificadas pelas RIDs que falharam.

A imputação é dita Imputação Determinística Única (IDU) se existir um único código a ser atribuído ao registro que falhou. Por outro lado, se existirem vários códigos possíveis, a variável é passível de imputação mediante a seleção probabilística de um de seus códigos, e neste caso se diz que a variável é pendente de Imputação Determinística Flexível (IDF).

Para os registros que falharam em alguma regra de crítica, aplica-se o módulo de imputação probabilística que executa duas tarefas:

- seleciona o conjunto de variáveis a modificar para cada registro com erro; e
- seleciona o código a ser imputado para cada variável a modificar.

[11] Este era o método adotado na versão 1 do sistema DIA, e foi conservado na versão 2 para manter a compatibilidade entre as duas versões.

Para cada registro com erro são verificadas as críticas do conjunto completo que falharam. Dessas críticas o sistema verifica o conjunto de variáveis ativas e seleciona o conjunto de variáveis para serem imputadas. A seleção das variáveis é feita levando em conta a de maior índice de suspeita, cujo cálculo é baseado na atividade da variável, no peso da variável e, em menor medida e em proporção inversa, no grau de atividade do código ante as regras do conjunto completo. A seguir é apresentada a expressão para o cálculo do Índice de Suspeita de uma dada variável i, denotado por $S(i)$:

$$S(i) = ATIV(i) \,.\, \left[PESO(i) + \left(1 - PESO(i)\right) \,.\, P_j \right] \qquad (1)$$

onde:

$ATIV(i)$ é a atividade da variável i, e definida como o número de regras de crítica do conjunto completo em que a variável está ativa; é o fator que mais contribui para o cálculo do $S(i)$ e varia entre zero e o número de regras de crítica do conjunto completo.

$PESO(i)$ é o peso associado à variável i;

$\left[(1 - \text{PESO(i)}) \,.\, \text{P}_\text{j}\right]$ pode ser interpretado como um incremento à suspeita subjetiva e serve para resolver empates hipotéticos;

P_j é construído a partir do grau de atividade do código j (GACj) usando a expressão (2) abaixo; quanto maior o grau de atividade de um código menor será o valor associado a P_j; a divisão por 10 visa a diminuir o impacto deste fator, na composição do índice de suspeita (1);

$$\text{P}_\text{j} = \frac{1 - \dfrac{\text{GAC}_\text{j}}{\text{MAX}_{\text{GAC}} + 1}}{10} \qquad (2)$$

GACj é o grau de atividade de um código j, definido como a quantidade de regras de crítica do conjunto completo em que o código está ativo;

MAX_{GAC} é definido como o MAX $\{GAC_j \mid j \in$ conjunto de códigos válidos de todas as variáveis$\}$.

No caso de ocorrência de empate, isto é, mais de uma variável com o valor máximo do Índice de Suspeita, é feita uma seleção (ao acaso) entre elas para decidir qual será imputada. Após a seleção das variáveis a imputar de um registro com erro, são então determinados os códigos a serem imputados para cada uma delas. Se só existir um código possível que seja diferente do código atribuído ao registro rejeitado, este é eleito e diz-se que ocorreu Imputação Probabilística Única (IPU) da variável. Se existirem vários códigos possíveis, é feita a seleção probabilística de um deles e diz-se que ocorreu Imputação Probabilística Flexível (IPF) da variável. A seleção do código é feita de acordo com a estratégia de tratamento definida no subsistema de especificação, levando em conta a distribuição de freqüência marginal ou conjunta, o tipo de freqüência (FNS) ou (FNS*) e o método de seleção do código (proporcional ou pela máxima diferença na distribuição de freqüências).

O sistema DIA fornece um conjunto amplo de tabelas com informações para subsidiar a análise do processo de detecção e correção automática dos erros, dentre as quais destacam-se: número de registros processados sem erro (bons) e de registros com algum erro (maus); número de registros errôneos por tipo de erro e o número de críticas que falharam; número de registros que falharam em cada regra de crítica; número de registros por número de variáveis imputadas; número de registros imputados por tipo de imputação; e distribuição marginal de cada variável, em que considera o arquivo de entrada (antes da correção), o arquivo dos registros bons e o arquivo corrigido.

O sistema DIA garante que a correção é realizada com uma única passagem, ou seja, uma vez tratados os registros com erro, nenhum registro será rejeitado em uma nova rodada do módulo de detecção considerando o mesmo conjunto de regras de crítica. Além da vantagem de evitar os ciclos de processamento, um outro aspecto altamente positivo

é o fato de que não há necessidade de desenvolvimento de programas de imputação. Com isso já é possível perceber o ganho significativo desse método em relação ao adotado no Censo de 80. Isto ocorre porque é possível concentrar os esforços na especificação do plano de depuração e na análise dos resultados da imputação, ao invés de no desenvolvimento e teste de programas.

4. A APLICAÇÃO DO SISTEMA DIA AOS DADOS DO CD 1.01

Assim que os dados do CD 1.01 foram transmitidos para a sede do IBGE no Rio de Janeiro, o primeiro procedimento adotado na etapa centralizada foi a consolidação dos dados transmitidos, para em seguida formar lotes de questionários com vistas à execução da detecção e correção automática dos erros. A formação de lotes foi baseada nos seguintes critérios[12]:

- obtenção do menor número possível de lotes, para minimizar o número de relatórios a serem analisados, após a aplicação do sistema DIA;
- obtenção de lotes com tamanho mínimo acima de um limite pré-especificado a fim de viabilizar a utilização de distribuições de registros bons (sem erros) "estáveis" como base para a imputação;
- geração dos lotes levando em conta a situação do domicílio (urbana e rural) bem como a divisão geográfica do País, contemplando a divisão de cada Unidade da Federação através da ordenação dos respectivos setores, segundo a microrregião, município, distrito e subdistrito de modo a garantir, na medida do possível, maior homogeneidade quanto às características de cada região durante o processo de correção.

Considerando esses critérios, os dados investigados no CD 1.01 foram separados em 528 lotes de questionários, sendo 403 lotes provenientes de setores urbanos e 125 lotes de setores rurais, de aproximadamente 70 000 domicílios cada.

[12] Ver Silva, Oliveira e Oliveira (1992).

Considerando a estreita ligação existente entre as especificações da pré-crítica e da correção automática, as variáveis corrigidas na pré-crítica tornaram-se fixas durante a fase de correção automática, visando, assim, ao não comprometimento daquilo que já havia sido corrigido e, conseqüentemente, a manutenção da coerência interna dos questionários sem recorrer a ciclos de acertos.

4.1. Estratégia para o Tratamento dos Dados

Foi planejado executar o tratamento dos dados em três aplicações DIA para cada lote de apuração: uma para dados de domicílios e duas referentes às características de pessoas. Quanto às aplicações de pessoas, uma foi desenvolvida para tratar as características dos chefes dos domicílios e individuais em domicílios coletivos, e a outra para as características dos não-chefes de domicílios, considerando a idade do chefe já corrigida para tratar a idade dos demais componentes do domicílio. A principal razão para adotar duas aplicações do sistema DIA para a detecção e correção dos erros em características de pessoas foi possibilitar a construção da variável auxiliar "faixa de idade do chefe do domicílio", *obtida a partir dos dados "limpos" da variável idade dos chefes*. Dessa forma, foi possível tratar a variável idade dos não-chefes usando a distribuição conjunta da variável auxiliar "faixa de idade do chefe" com a variável "parentesco ou relação com o chefe do domicílio".

As variáveis do CD 1.01 que foram submetidas ao processo de detecção e correção automática dos erros, por aplicação, são as seguintes:

- características do domicílio: localização, abastecimento de água, uso de instalação sanitária, condição de ocupação, cômodos, cômodos servindo de dormitório, banheiros e destino do lixo;
- características de chefes ou individuais em domicílios coletivos: idade calculada com base no mês e ano de nascimento/idade presumida, alfabetização, última série concluída, grau da última série concluída e tipo de rendimento; e

- características de não-chefes: idade calculada com base no mês e ano de nascimento/idade presumida, alfabetização.

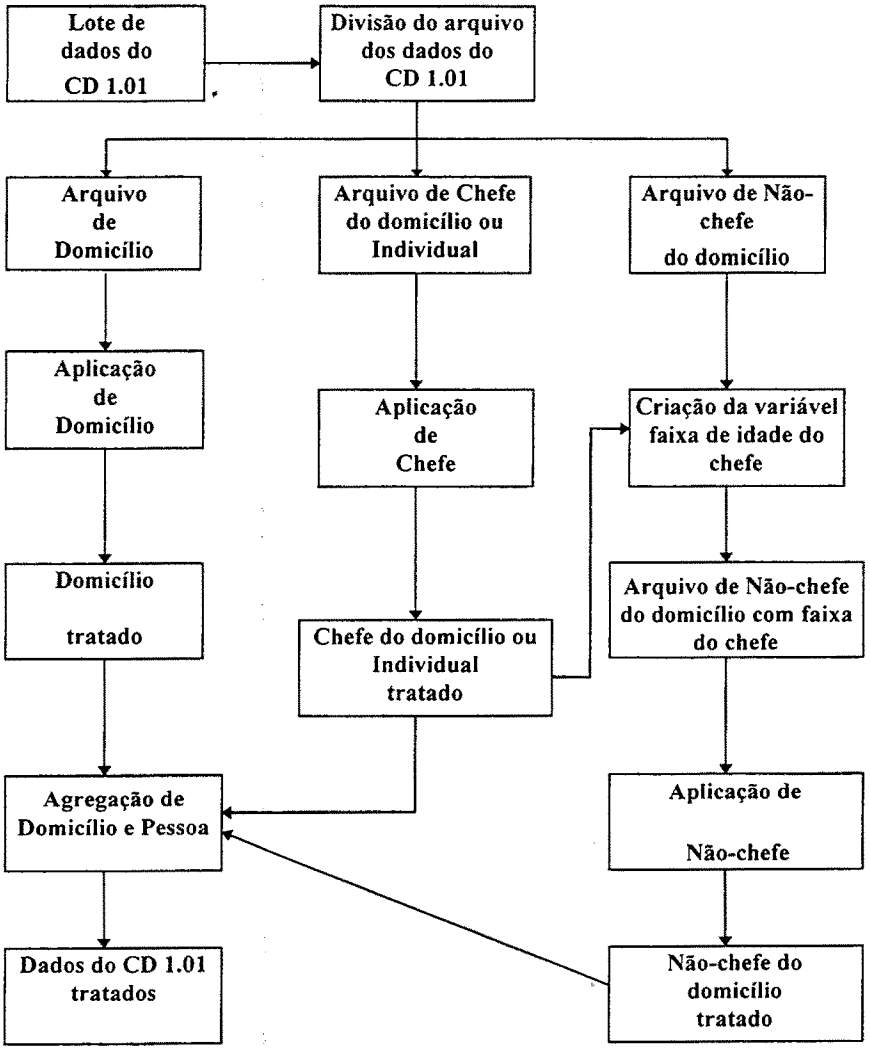
No Anexo 1 são apresentadas as regras de crítica especificadas por aplicação DIA. São apresentados no Quadro 1 os quantitativos de regras de crítica especificadas e do conjunto completo, segundo as aplicações definidas para a depuração dos dados e na Figura 2 o fluxo da operacionalização da execução da correção automática.

Quadro 1 - Número de regras de crítica, segundo as aplicações DIA

Aplicação DIA	Número de regras de crítica		
	Incompatibilidade	Imputação determinística	Conjunto completo
TOTAL	143	18	241
Domicílios	87	2	168
Chefes e Individuais	51	14	67
Não-chefes	5	2	6

Figura 2

Fluxo da operacionalização da execução da correção automática



4.2. Execução da Correção Automática

O sistema de produção foi desenvolvido de modo a submeter cada lote de questionários previamente formado às aplicações de domicílio, de chefe e de não-chefe, nessa ordem, já que a aplicação de não-chefe dependia da realização prévia da correção dos dados dos chefes. Durante a execução da correção foram tratadas as inconsistências verificadas entre as variáveis de um mesmo registro assim como os valores inválidos, dentre eles os valores ignorados de algumas variáveis. Estes foram, em alguns casos, considerados "inválidos" durante o processo de imputação, afim de que fossem imputados em seu lugar valores válidos coerentes com a distribuição dos registros bons (sem erros). Isto se justifica porque os usuários das informações, de um modo geral, tendem a "imputar implicitamente" esses valores distribuindo proporcionalmente pelas categorias das variáveis.

Antes da depuração definitiva das informações de domicílio, chefe e não-chefe, foram feitos testes com lotes de trabalho de algumas Unidades da Federação, já liberados da fase de pré-crítica, com o intuito de verificar todo o funcionamento da operação, detectar possíveis problemas que poderiam aparecer posteriormente e, com base nos relatórios de saída da correção, revisar o plano de análise da correção automática. Foi em função desses testes que ficou definida a estratégia de aplicar um peso diferenciado para a variável idade, de modo a controlar a sua imputação para ter menos chance de ser imputada.

Em se tratando de uma metodologia utilizada pela primeira vez em uma pesquisa do IBGE, é natural a ocorrência de alguns ajustes ou calibrações. Destaque deve ser dado para a facilidade com que foram feitas as alterações no conjunto de regras de crítica ou na estratégia de tratamento dos dados com o uso do sistema DIA. O Censo Demográfico de 1991 teve a grande oportunidade de utilizar uma ferramenta poderosíssima e bastante eficiente, obtendo assim, um grande avanço em termos de apuração censitária.

4.3. Imputação de Questionários Faltosos

Por ocasião do empastamento dos questionários foram constatadas faltas de alguns questionários. Durante a pré-crítica, foram criados questionários faltosos e recuperadas das folhas de coleta as informações referentes às variáveis: total de homens, total de mulheres e espécie do domicílio para permitir que esses questionários fossem submetidos ao processo de imputação dos dados.

Foi feita uma avaliação de que no caso de um questionário faltoso, praticamente todo em branco, a imputação pelo sistema DIA, como é feita variável a variável, ficaria muito próxima a uma fabricação pura e simples de um questionário fictício. Isto porque o processo é ideal para tratar os casos de não resposta parcial e erros entre campos de um questionário. Para evitar a criação de questionários fictícios, decidiu-se adotar o método *Hot-Deck* seqüencial para a imputação dos campos sem declaração dos questionários gravados nos arquivos dos faltosos, de acordo com o seguinte procedimento: procurava-se no lote um questionário para servir de "doador" que tivesse os mesmos valores que o questionário faltoso para as variáveis disponíveis. Caso fosse encontrado, todas as variáveis de domicílios e de pessoas eram imputadas segundo o questionário doador, com exceção da identificação.

Foram feitas imputações em 25 441 questionários faltosos[13] correspondendo a 97 352 pessoas com imputações em faltosos. Foram eliminados 4 questionários faltosos por falta de doadores no respectivo lote. Os questionários faltosos representam 0,07% do total de questionários CD 1.01.

[13] Incluindo 42 setores do Maranhão cujos questionários haviam sido perdidos que correspondem a 5 921 domicílios e 24 972 pessoas.

5. ANÁLISE DO PROCESSO DE CORREÇÃO AUTOMÁTICA

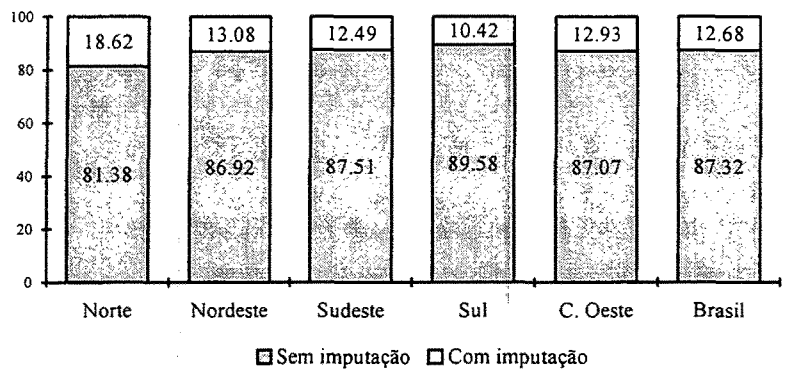
A análise da correção automática para a liberação dos dados para armazenamento e tabulação foi realizada em três níveis geográficos: lote, município e subdistrito, sendo este último nível aplicado apenas para os municípios de algumas capitais que compreendem: Rio de Janeiro, São Paulo, Brasília, Salvador, Recife, Belo Horizonte e Fortaleza[14]. Os resultados da correção automática foram armazenados em arquivos de lotes, tendo sido efetuada uma consolidação das tabelas especiais a nível de Unidades da Federação, Regiões e Brasil. Algumas estatísticas foram selecionadas, afim de revelar como funcionou a aplicação da metodologia de Fellegi & Holt e do sistema DIA para correção automática dos erros detectados nos dados referentes ao CD 1.01 do Censo Demográfico de 1991.

A nível Brasil, 87,32% dos questionários ficaram isentos de qualquer correção automática, sendo que o processo de imputação foi executado em apenas 12,68% dos questionários, como pode ser constatado através do Gráfico 1.

Um questionário foi considerado com alguma imputação quando pelo menos um campo (no registro do domicílio ou de uma pessoa do domicílio - chefe ou não-chefe) teve código imputado pelo sistema DIA ou pelo procedimento adotado para tratar de questionário faltoso. No Gráfico 1 é apresentada a distribuição de questionários por ocorrência de imputação para cada uma das Regiões e para o Brasil. O maior percentual de questionários "com imputação" na Região Norte é decorrente, em grande parte, do percentual observado para o Amazonas, cujo valor foi de 29,69%.

[14] Uma descrição do plano de análise da correção automática do CD 1.01 encontra-se em IBGE (1993).

Gráfico 1
Distribuição percentual do número de questionários por ocorrência de imputação



Uma outra avaliação efetuada diz respeito ao percentual de registros com alguma imputação por tipo de aplicação do sistema DIA. A Tabela 1 apresenta esses percentuais por Região e para o Brasil, para as aplicações de domicílio, chefe e não-chefe, respectivamente.

Tabela 1 - Percentual de registros com alguma imputação, por aplicação DIA, para Brasil e Regiões

Brasil e Regiões	Aplicação DIA		
	Domicílio	Chefe	Não-chefe
Brasil	4,91	4,85	1,36
Norte	8,59	5,86	1,74
Nordeste	5,48	3,68	1,45
Sudeste	4,38	5,64	1,38
Sul	3,77	4,40	1,06
Centro-Oeste	5,94	4,08	1,13

Observa-se que a aplicação de não-chefe apresenta percentuais de registros com alguma imputação bem inferiores aos percentuais obtidos para as aplicações de chefe e de domicílio, o que pode ser explicado pelo menor número de variáveis sujeitas à imputação para essa aplicação. Por outro lado, nas Regiões Sul e Sudeste o percentual de registros com imputação na aplicação de chefe é superior ao da aplicação de domicílio, num sentido inverso ao das demais Regiões, o que levou a que para o total do País esses percentuais ficassem praticamente iguais, em torno de 5%.

Devido à arquitetura das aplicações que originou a separação dos registros de domicílios, chefes e não-chefes de um mesmo questionário em arquivos distintos para o processamento pelo sistema DIA, é importante analisar os resultados da correção automática a nível de questionário afim de verificar a sua integridade.

A Tabela 2 apresenta, dentre os questionários com alguma imputação, o percentual daqueles com imputação em apenas uma aplicação ou mais de uma, por tipo de aplicação, para Brasil e Regiões. Cabe registrar que, a nível Brasil, dentre os questionários que tiveram alguma imputação através do sistema DIA, 93,12% foram "corrigidos" em apenas uma das aplicações (domicílio, chefe ou não-chefe), 6,33% em duas das aplicações e apenas 0,55% nas três aplicações simultaneamente.

Tabela 2 - Distribuição percentual de questionários com alguma imputação, por tipo de aplicação DIA, para Brasil e Regiões

Brasil e Regiões	Uma aplicação				2 ou 3 aplicações
	Só domicílio	Só chefe	Só não-chefe	Total	
Brasil	34,13	33,17	25,82	93,12	6,88
Norte	39,09	25,26	26,37	90,72	9,28
Nordeste	37,56	24,43	32,00	93,99	6,01
Sudeste	29,94	39,21	23,32	92,47	7,52
Sul	32,74	38,01	23,66	94,41	5,59
Centro-Oeste	45,00	27,37	21,96	94,33	5,57

Um outro aspecto do processo de correção automática dos dados do CD 1.01 diz respeito ao número de variáveis imputadas em cada aplicação. As Tabelas 3, 4 e 5 apresentam as distribuições de registros com alguma imputação na aplicação de domicílio, chefe e não-chefe, respectivamente, por número de variáveis imputadas para as Regiões e Brasil. Aproximadamente 90% dos registros com alguma imputação tiveram imputação em apenas uma variável em cada aplicação, sendo 89,09% em domicílio, 88,53% em chefe e 94,14% em não-chefe. Considerando-se que, pela Tabela 2, 93,12% dos questionários tiveram imputação em apenas uma das aplicações, pode-se inferir que mais de 80% dos questionários que tiveram imputação foram corrigidos com a imputação de apenas **uma** variável. Tais resultados estão de acordo com o princípio na qual se baseia a metodologia de Fellegi & Holt "os dados de cada registro devem satisfazer todas as críticas alterando o menor número possível de variáveis, procurando assim manter a maior quantidade possível de informação original".

Na Tabela 3 observa-se a ocorrência de registros com imputações em 9 variáveis, na aplicação de domicílio. Isto se justifica porque a variável espécie do domicílio foi tratada na pré-crítica e permaneceu fixa durante todo o processo automático de correção; em conseqüência, quaisquer inconsistências detectadas entre a variável espécie do domicílio e as demais variáveis correspondentes às características de domicílios só puderam ser corrigidas através da imputação das outras 9 variáveis envolvidas nessa aplicação.

Tabela 3 - Distribuição percentual de registros com alguma imputação na aplicação de Domicílio, por número de variáveis imputadas, para Brasil e Regiões

Brasil e Regiões	Número de variáveis imputadas			
	1	2	3 a 8	9
Brasil	89,09	4,50	3,34	3,07
Norte	89,62	5,20	2,39	2,80
Nordeste	92,43	3,97	2,81	1,80
Sudeste	85,45	4,88	5,26	4,41
Sul	91,00	4,32	2,18	2,50
Centro-Oeste	92,09	3,81	2,03	2,06

Tabela 4 - Distribuição percentual de registros com alguma imputação na aplicação de Chefe, por número de variáveis imputadas, para Brasil e Regiões

Brasil e Regiões	Número de variáveis imputadas		
	1	2	3 a 5
Brasil	88,53	9,27	1,92
Norte	89,94	9,30	0,74
Nordeste	91,17	7,88	0,90
Sudeste	86,31	10,23	2,97
Sul	90,81	8,57	0,56
Centro - Oeste	92,77	6,75	0,46

Tabela 5 - Distribuição percentual de registros com alguma imputação na aplicação de Não-chefe, por número de variáveis imputadas, para Brasil e Regiões

Brasil e Regiões	Número de variáveis imputadas	
	1	2
Brasil	94,14	5,86
Norte	96,56	3,44
Nordeste	97,47	2,53
Sudeste	89,71	10,29
Sul	97,64	2,36
Centro - Oeste	97,44	2,56

As análises aqui apresentadas não esgotam outras análises mais aprofundadas que podem ser efetuadas, pois encontram-se disponíveis os registros de entrada e de saída da correção automática, bem como a identificação daqueles que tiveram alguma imputação. O sistema fornece todos os recursos necessários para uma análise e monitoramento adequados do processo, como se recomenda. Das análises efetuadas pode-se concluir que o efeito da imputação sobre os resultados foi pequeno e estava dentro do que se esperava, não introduzindo distorções indesejáveis nos dados.

6. Considerações Finais

A aplicação do sistema DIA para a detecção e correção automática dos erros do Censo Demográfico de 1991 representou um grande avanço no uso de sistemas generalizados no IBGE. Entre as principais vantagens, destacam-se:

- integração lógica entre a detecção dos erros e a imputação em apenas um ciclo de processamento;
- preservação tanto quanto possível da distribuição dos dados bons, seja ela marginal ou conjunta;
- utilização de estratégia de imputação que altera o menor número possível de variáveis em cada registro;
- controle e acompanhamento da imputação favorecidos, já que o sistema fornece recursos adequados para a análise e monitoramento do processo;
- redução do custo de desenvolvimento de sistemas; e
- ganhos de velocidade, com redução do tempo de execução da imputação.

O sucesso obtido na apuração do Questionário Básico - CD 1.01 do Censo Demográfico de 1991 com a utilização do sistema DIA e a avaliação dos ganhos significativos de qualidade do processo levou à decisão de utilização desse sistema de imputação também para a depuração dos dados do Questionário da Amostra - CD 1.02[15].

Desse modo, seria oportuno conduzir estudos para avaliar a viabilidade da utilização da metodologia de Fellegi & Holt, através do sistema DIA, em outras pesquisas domiciliares do IBGE, em especial, na Pesquisa Nacional por Amostra de Domicílios - PNAD.

[15] Ver Oliveira e outros (1996).

Anexo 1 - Planos de crítica para a detecção e correção automática de erros, pelo sistema DIA

APLICAÇÃO DE DOMICÍLIOS

Regras de imputação determinística

1) Se V201 (1) e V1062 (1) e V202 (1-2) ⇒ V202 = 3

2) Se V201 (1) e V1062 (1) e V202 (4-5) ⇒ V202 = 6

Regras de incompatibilidade

1) Se V201 (1) e V202 ( )

2) Se V201 (2-3) e V202 (1-7)

3) Se V201 (1) e V203 ( )

4) Se V201 (2-3) e V203 (1-6)

5) Se V201 (1) e V204 ( )

6) Se V201 (2-3) e V204 (0-7)

7) Se V201 (1) e V205 ( )

8) Se V201 (2-3) e V205 (0-2)

9) Se V201 (1) e V206 ( )

10) Se V201 (2-3) e V206 (1-6)

11) Se V201 (1) e V207 ( )

12) Se V201 (2-3) e V207 (1-30)

13) Se V201 (1) e V208 ( )

14) Se V201 (2-3) e V208 (0-8)

15) Se V201 (1) e V209 ( )

16) Se V201 (2-3) e V209 (0-5)

17) Se V201 (1) e V210 ( )

18) Se V201 (2-3) e V210 (1-7)

19) Se V2203 (2) e V203 (1,4)

20) Se V205 (1-2) e V204 (0)

21) Se V209 (1-5) e V204 (0)

22) Se V205 (0) e V204 (1-7)

23) Se V204 (2) e V204 (1)
24) Se V205 (0) e V209 (1-5)
25) Se V208 (0,2-8) e V207 (1)
26) Se V209 (1-5) e V207 (1)
27) Se V208 (0,3-8) e V207 (2)
28) Se V209 (1-5) e V208 (2) e V207 (2)
29) Se V209 (2-5) e V207 (2)
30) Se V208 (0,4-8) e V207 (3)
31) Se V209 (2-5) e V208 (2) e V207 (3)
32) Se V209 (1-5) e V208 (3) e V207 (3)
33) Se V209 (3-5) e V207(3)
34) Se V208 (0,5-8) e V207 (4)
35) Se V209 (3-5) e V208 (2) e V207 (4)
36) Se V209 (2-5) e V208 (3) e V207 (4)
37) Se V209 (1-5) e V208 (4) e V207 (4)
38) Se V209 (4-5) e V207 (4)
39) Se V208 (0,6-8) e V207 (5)
40) Se V209 (4-5) e V208 (2) e V207 (5)
41) Se V209 (3-5) e V208 (3) e V207 (5)
42) Se V209 (2-5) e V208 (4) e V207 (5)
43) Se V209 (1-5) e V208 (5) e V207 (5)
44) Se V209 (5) e V207 (5)
45) Se V208 (0,7-8) e V207 (6)
46) Se V209 (5) e V208 (2) e V207 (6)
47) Se V209 (4-5) e V208 (3) e V207 (6)
48) Se V209 (3-5) e V208 (4) e V207 (6)
49) Se V209 (2-5) e V208 (5) e V207 (6)
50) Se V209 (1-5) e V208 (6) e V207 (6)
51) Se V208 (0,8) e V207 (7)
52) Se V209 (5) e V208 (3) e V207(7)
53) Se V209 (4-5) e V208 (4) e V207 (7)
54) Se V209 (3-5) e V208 (5) e V207 (7)

55) Se V209 (2-5) e V208 (6) e V207 (7)
56) Se V209 (1-5) e V208 (7) e V207 (7)
57) Se V208 (0) e V207 (8)
58) Se V209 (5) e V208 (4) e V207 (8)
59) Se V209 (4-5) e V208 (5) e V207 (8)
60) Se V209 (3-5) e V208 (6) e V207 (8)
61) Se V209 (2-5) e V208 (7) e V207 (8)
62) Se V209 (1-5) e V208 (8) e V207 (8)
63) Se V209 (1-5) e V208 (0) e V207 (9)
64) Se V209 (5) e V208 (5) e V207 (9)
65) Se V209 (4-5) e V208 (6) e V207 (9)
66) Se V209 (3-5) e V208 (7) e V207 (9)
67) Se V209 (2-5) e V208 (8) e V207 (9)
68) Se V209 (2-5) e V208 (0) e V207 (10)
69) Se V209 (5) e V208 (6) e V207(10)
70) Se V209 (4-5) e V208 (7) e V207 (10)
71) Se V209 (3-5) e V208 (8) e V207 (10)
72) Se V209 (3-5) e V208 (0) e V207 (11)
73) Se V209 (5) e V208 (7) e V207 (11)
74) Se V209 (4-5) e V208 (8) e V207 (11)
75) Se V209 (4-5) e V208 (0) e V207 (12)
76) Se V209 (5) e V208 (8) e V207 (12)
77) Se V209 (5) e V208 (0) e V207 (13)
78) Se V208 (0) e V7100 (1-8)
79) Se V208 (0, 2-8) e V7100 (1)
80) Se V208 (0,3-8) e V7100 (1-2)
81) Se V208 (0,4-8) e V7100 (1-3)
82) Se V208 (0,5-8) e V7100 (1-4)
83) Se V208 (0,6-8) e V7100 (1-5)
84) Se V208 (0,7-8) e V7100 (1-6)
85) Se V208 (0,8) e V7100 (1-7)
86) Se V205 (1) e V209 (0) e V204 (1-3)

87) Se V201 (1) e V1062 (1) e V202 (1-2,4-5)

APLICAÇÃO DE CHEFES

Regras de imputação determinística

1) Se V305 (1-2) e V302 (20) e V3043 (1-4,900-911) ⇒ V305 = 3
2) Se V305 (3) e V302 (20) e V3043 (5) ⇒ V305 = 1-2
3) Se V306 (1) e V307 (2) e V3043 (6-24) ⇒ V307 = 4
4) Se V306 (2) e V307 (2) e V3043 (7-25) ⇒ V307 = 4
5) Se V306 (3) e V307 (2) e V3043 (8-26) ⇒ V307 = 4
6) Se V306 (4) e V307 (2) e V3043 (9-27) ⇒ V307 = 4
7) Se V306 (1) e V307 (3) e V3043 (10-28) ⇒ V306 = 5 e V307 = 4
8) Se V306 (2) e V307 (3) e V3043 (11-29) ⇒ V306 = 6 e V307 = 4
9) Se V306 (3) e V307 (3) e V3043 (12-30) ⇒ V306 = 7 e V307 = 4
10) SeV306 (4) e V307(3) e V3043 (13-31) ⇒ V306 = 8 e V307 = 4
11) SeV306 (1) e V307 (6) e V3043 (14-32) ⇒ V307 =5
12) SeV306 (2) e V307 (6) e V3043 (15-33) ⇒ V307 =5
13) SeV306 (3) e V307 (6) e V3043 (16-34) ⇒ V307 =5
14) SeV306 (4) e V307 (6) e V3043 (17-35) ⇒ V307 =5

Regras de incompatibilidade

1) Se V305 (1-2) e V306 ( )
2) Se V305 (3) e V306 (0-8)
3) Se V305 (1-2) e V307 ( )
4) Se V305 (3) e V307 (0-8)
5) Se V305 (1-2) e V3080 ( )
6) Se V305 (3) e V3080 (0-1,9)
7) Se V302 (1) e V3043 (1-9, 900-911)
8) Se V305 (1-2) e V3043 (1-4, 900-911)

9) Se V305 (3) e V3043 (5-130)
10) Se V306 (1) e V307 (2) e V3043 (1-24, 900-911)
11) Se V306 (2) e V307 (2) e V3043 (1-25,900-911)
12) Se V306 (3) e V307 (2) e V3043 (1-26, 900-911)
13) Se V306 (4) e V307 (2) e V3043 (1-27, 900-911)
14) Se V306 (5) e V307 (2) e V3043 (1-28, 900-911)
15) Se V306 (6) e V307 (2) e V3043 (1-29, 900-911)
16) Se V306 (1) e V307 (3) e V3043 (1-28, 900-911)
17) Se V306 (2) e V307 (3) e V3043 (1-29,900-911)
18) Se V306 (3) e V307 (3) e V3043 (1-30, 900-911)
19) Se V306 (4) e V307 (3) e V3043 (1-31, 900-911)
20) Se V306 (5) e V307 (3) e V3043 (1-65, 900-911)
21) Se V306 (1) e V307 (4) e V3043 (1-5, 900-911)
22) Se V306 (2) e V307 (4) e V3043 (1-6, 900-911)
23) Se V306 (3) e V307 (4) e V3043 (1-7, 900-911)
24) Se V306 (4) e V307 (4) e V3043 (1-8, 900-911)
25) Se V306 (5) e V307 (4) e V3043 (1-9, 900-911)
26) Se V306 (6) e V307 (4) e V3043 (1-10, 900-911)
27) Se V306 (7) e V307 (4) e V3043 (1-11, 900-911)
28) Se V306 (8) e V307 (4) e V3043 (1-12, 900-911)
29) Se V306 (1) e V307 (5) e V3043 (1-13, 900-911)
30) Se V306 (2) e V307 (5) e V3043 (1-14, 900-911)
31) Se V306 (3) e V307 (5) e V3043 (1-15, 900-911)
32) Se V306 (4) e V307 (5) e V3043 (1-16, 900-911)
33) Se V306 (1) e V307 (6) e V3043 (1-32, 900-911)
34) Se V306 (2) e V307 (6) e V3043 (1-33, 900-911)
35) Se V306 (3) e V307 (6) e V3043 (1-34, 900-911)
36) Se V306 (4) e V307 (6) e V3043 (1-35, 900-911)
37) Se V306 (1) e V307 (7) e V3043 (1-16, 900-911)
38) Se V306 (2) e V307 (7) e V3043 (1-17, 900-911)
39) Se V306 (3) e V307 (7) e V3043 (1-18, 900-911)
40) Se V306 (4) e V307 (7) e V3043 (1-19,900-911)

41) Se V306 (5) e V307 (7) e V3043 (1-20, 900-911)

42) Se V306 (6) e V307 (7) e V3043 (1-21, 900-911)

43) Se V307 (1) e V3043 (1-13, 900-911)

44) Se V307 (8) e V3043 (1-20, 900-911)

45) Se V305 (2) e V306 (3-8)

46) Se V305 (2) e V307 (3, 5-8)

47) Se V306 (1-8) e V307 (0-1, 8)

48) Se V306 (0, 6-8) e V307 (3)

49) Se V306 (0) e V307 (4)

50) Se V306 (0, 5-8) e V307 (5-6)

51) Se V306 (0, 7-8) e V307 (2,7)

APLICAÇÃO DE NÃO-CHEFES

Regras de imputação determinística

1) Se V305 (1-2) e V302 (3-4, 8, 10-14, 16) e V3043 (1-4, 900-911) $\Rightarrow$V305 = 3

2) Se V305 (3) e V3043 (5) $\Rightarrow$ V305 = 1-2

Regras de incompatibilidade

1) Se V302 (2, 9, 15) e V3043 (1-9, 900-911)

2) Se V302 (5-6) e V3043 (1-19, 900-911)

3) Se V302 (7) e V3043 (1-29, 900-911)

4) Se V305 (1-2) e V3043 (1-4, 900-911)

5) Se V305 (3) e V3043 (5-130)

REFERÊNCIAS

ALBIERI, S. e BIANCHINI, Z.M. Censo Demográfico de 1991. Sobre a independência da apuração do CD 1.01 e CD 1.02. Rio de Janeiro: IBGE, Divisão de Metodologia, 1993. 20p.

ANÁLISE da correção automática no Censo Demográfico de 1991. Rio de Janeiro: IBGE, Coordenação Técnica do Censo Demográfico, [1993].

CRIADO, I.V. e CABRIA, M.S.B. Procedimiento de depuración de datos estadísticos. Vitoria-Gasteiz: EUSTAT - Instituto Vasco de Estadistica, 1990. 169p.

FELLEGI, I.P. e HOLT, D. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, v.71, p.17-35. 1976.

HANONO, R.M. *DIA* - Integração à arquitetura de informática do IBGE. Rio de Janeiro: IBGE, 1993. (INFOTEC v.2 n. 9).

LAS FRECUENCIAS para la imputacion DIA. Madrid: Instituto Nacional de Estadística (INE), [1993]. (mimeo). 15p.

LA SELECCION de variables DIA. Madrid: Instituto Nacional de Estadística (INE), [1993]. (mimeo). 23p.

LA SELECCION de codigos DIA. Madrid: Instituto Nacional de Estadística (INE), [1993]. (mimeo). 9p.

METODOLOGIA do Censo Demográfico de 1980. Rio de Janeiro: IBGE, 1983. 477p. (Série Relatórios Metodológicos, n. 3).

OLIVEIRA, L.C.S., INDÁ, L.B., LIMA, R.L.A. e BIANCHINI, Z.M. Apuração dos dados investigados no questionário básico (CD 1.01) do Censo Demográfico de 1991. Rio de Janeiro: IBGE, 1994. (Textos para Discussão, n. 71).

OLIVEIRA, L.C.S., INDÁ, L.B., MENDONÇA, M.S. e LIMA, R.L.A. Apuração dos dados investigados no questionário da amostra (CD 1.02) do Censo Demográfico de 1991. Rio de Janeiro: IBGE, 1996. (Textos para Discussão n. 86).

RUBIO, E.G. e CRIADO, I.V. Sistema DIA - sistema de Detección e Imputación Automática de errores para datos cualitativos. Volumen I. DIA: descripción del sistema. Madrid: Instituto Nacional de Estadística (INE), 1988.

SILVA, P.L.N., OLIVEIRA, E.M. e OLIVEIRA, L.C.S. Sobre o critério de formação de lotes de apuração do questionário básico CD 1.01 (versão 2). Rio de Janeiro: IBGE, 1992. 10p.

SILVA, P.L.N, OLIVEIRA, E.M., OLIVEIRA, L.C.S. e LIMA, R.L.A. Uma nova metodologia para correção automática no Censo Demográfico brasileiro: experimentação e primeiros resultados. Rio de Janeiro: IBGE, 1990. 102p. (Textos para Discussão, n. 28).

Resumo

Este artigo aborda aspectos relativos ao uso do Sistema DIA para a detecção e correção automática de erros nos dados investigados no questionário básico do Censo Demográfico de 1991. O Sistema DIA foi desenvolvido pelo Instituto Nacional de Estadística (INE) da Espanha e permite executar o processo de detecção e correção de erros em dados categóricos em apenas um ciclo de processamento. Além disso, fornece um amplo conjunto de informações para subsidiar a análise da correção automática. As análises efetuadas apontam a qualidade dos dados investigados e a eficiência do sistema adotado, que assegura consistência e preserva a distribuição básica dos dados.

Abstract

This paper deals with aspects of the application of the generalized data editing and imputation software named DIA to the 1991 Population Census Basic Questionnaire. This software, developed by the Spanish National Statistical Institute, handles editing and imputation of categorical data in one processing cycle and provides comprehensive information to control and assess the automatic correction process. The analysis reveals the data quality and efficiency of the software adopted, which ensures data consistency while preserving basic distribution properties.

Alguns Modelos Complexos de Regressão para Análise de Dados Univariados

Gauss M. Cordeiro*

1. INTRODUÇÃO

Neste artigo de revisão apresentamos alguns modelos complexos de regressão que são bastante úteis na análise de dados univariados e procuramos destacar as suas potencialidades. Estes modelos são adequados para serem ajustados a dados independentes, cuja abordagem usual via os modelos clássicos de regressão não é satisfatória.

Na Seção 2 introduzimos, de forma bastante resumida, os modelos lineares generalizados e citamos as principais referências para o estudo de algumas de suas extensões. A Seção 3 faz o mesmo para os modelos aditivos generalizados. Na Seção 4 apresentamos uma nova classe de modelos via uma dupla transformação de parâmetros que inclui os modelos lineares generalizados e os modelos de Box e Cox (1964), como casos especiais. Damos ainda um exemplo de aplicação desta nova classe de modelos.

* Departamento de Estatística - CCEN/UFPE.
Cidade Universitária, 50740-540, Recife.

Na Seção 5 tratamos dos modelos não-exponenciais e apresentamos alguns exemplos. Descrevemos ainda o algoritmo de ajustamento destes modelos e os testes de adequação. A Seção 6 aborda os modelos de quase-verossimilhança, destacando aqueles com função de variância paramétrica e com parâmetro de dispersão não-constante. Nas Seções 7 e 8 introduzimos os modelos semiparamétricos e os modelos de regressão com estrutura de autocorrelação interna. Finalmente, a Seção 9 trata de outros modelos especiais.

2. MODELOS LINEARES GENERALIZADOS

Os modelos lineares generalizados (MLGs), também denominados modelos exponenciais lineares, têm desempenhado um papel importante na Estatística atual devido ao grande número de técnicas que englobam e às facilidades de ajustamento. Nestes modelos os dados $y = (y_1,\ldots,y_n)^T$ são supostos independentes com distribuição de probabilidade pertencente à família exponencial de distribuições. Esta família é indexada por dois parâmetros θ e ϕ e denotada por $F(\theta,\phi)$. A função de probabilidade ou densidade de $F(\theta,\phi)$ é expressa como

$$f(y;\theta,\phi) = \exp\left[\phi\left\{y\theta - b(\theta)\right\} + c(y,\phi)\right], \qquad (1)$$

onde a média $E(y) = \mu = db(\theta)/d\theta$ é uma função unívoca $\mu = g^{-1}(\eta)$ da estrutura usual $\eta = X\beta$ do modelo clássico de regressão. O parâmetro θ é denominado canônico e ϕ parâmetro de precisão. O inverso de ϕ é chamado parâmetro de dispersão.

Admite-se, ainda, que a matriz modelo X seja de posto completo p e que o vetor de parâmetros lineares desconhecidos seja $\beta = (\beta_1,\cdots,\beta_p)^T$. O vetor η é conhecido na literatura especializada como preditor linear. Várias distribuições (discretas e contínuas) importantes como a normal, gama, normal inversa, binomial e Poisson são membros de (1) e os seguintes modelos são casos especiais dos MLGs: modelo log-linear para a análise de tabelas de contingência, modelos logístico e *probit* para o estudo de proporções, modelo de regressão linear com erro normal, modelos estruturais com erro gama e outros modelos familiares. Diversas aplicações dos modelos acima estão apresentadas em Cordeiro (1986) e McCullagh e Nelder (1989) e nas referências que constam desses livros. Entretanto, os MLGs não englobam dados correlacionados, distribuições fora da família exponencial (1) e estruturas não-lineares (Ratkowsky, 1983) para o preditor η. Alguns modelos especiais de regressão que não são MLGs genuínos podem, entretanto, ser ajustados via os algoritmos iterativos dos MLGs, mediante pequenas alterações (vide Cordeiro e Paula,

1992). Entre estes modelos, que são bastante úteis nas aplicações práticas, mencionamos:

i) os modelos normais definidos pela componente sistemática

$$\mu = 1/(\beta_0 + \beta_1 x + \cdots + \beta_p x^p), \tag{2}$$

que são ajustados no *software* GLIM (*Generalized Linear Interactive Modelling*; *Payne*, 1986) com função de ligação inversa. Estes modelos são comumente denominados de polinômios inversos;

ii) os modelos normais parcialmente não-lineares em covariáveis

$$\mu = Z^T \alpha + g(X; \beta), \tag{3}$$

onde a função g(.;.) é não-linear em pelo menos uma componente de β, e a componente sistemática contém uma parte linear $Z^T\alpha$. O ajuste deste modelo pode ser feito através da linearização de (3) e reajuste iterativo do MLG resultante;

iii) os modelos de Box e Cox (1964), onde a variável resposta depende de parâmetros de transformação não-lineares.

Demonstra-se que o procedimento para estimar β nos MLGs equivale ao cálculo repetido de regressões lineares ponderadas (Jorgensen, 1983; Green, 1984). Várias extensões dos MLGs podem ser vistas ainda em Cordeiro (1986), McCullagh e Nelder (1989), e Cordeiro e Paula (1992). Nas seções seguintes apresentamos alternativamente alguns modelos adicionais para a análise de dados complexos que poderão ser ajustados no programa GLIM. Métodos de inferência nos MLGs, incluindo técnicas de diagnóstico e análise de resíduos, podem ser encontrados nas quatro últimas referências.

3. MODELOS ADITIVOS GENERALIZADOS

Os modelos aditivos são modelos não-paramétricos de regressão cujos efeitos dos preditores são aditivos. Um modelo aditivo é definido pela equação $y = \alpha + \sum_i f_i(x_i) + \varepsilon$, onde os erros em ε são supostos independentes dos x_i's com $E(\varepsilon) = 0$ e $Var(\varepsilon) = \sigma^2$, os f_i's são funções univariadas, arbitrárias, uma para cada preditor. Cada f_i representa uma função suavizadora (vide Seção 7) de natureza não-paramétrica definida de forma conveniente, como, por exemplo, por uma certa média ponderada de pontos y's correspondentes a alguma

vizinhança de x_i. Os f_i's são, individualmente, estimados através de regularizadores gráficos.

O modelo aditivo generalizado (MAG) difere do MLG, definido na Seção 2, pelo fato de um preditor aditivo substituir o preditor linear $n = X\beta$. Assim, o MAG é definido por uma distribuição em (1) cuja média $\mu = E(y|x_1,\cdots,x_p)$ se relaciona com o preditor linear por

$$\eta = g(\mu) = \alpha + \sum_{i=1}^{p} f_i(x_i). \tag{4}$$

A estimação de α e $f_1,\cdots,f_p$ é feita por um algoritmo iterativo do programa GAIM (*Generalized Additive Interactive Modelling*), descrito em Hastie e Tibshirani (1990; Seção 6.3). Vários exemplos de aplicação do algoritmo de ajustamento, descrevendo o uso do GAIM na análise de dados reais, são apresentados por estes autores. Alternativamente, o ajustamento dos MAGs pode ser feito dentro do programa GLIM por macros especiais.

A inferência nos MAGs segue procedimentos similares daqueles dos MLGs. Considere um MAG M_p definido por (1) e (4). Sejam $\hat{l}_p$ o máximo da log-verossimilhança segundo M_p e $\hat{l}_n$ a log-verossimilhança maximizada segundo o modelo saturado, isto é, sobre todas as médias μ_i's e sem a componente sistemática (4). Para o teste de adequação do modelo podemos comparar o desvio $D_p = 2(\hat{l}_n - \hat{l}_p)$, que desempenha o mesmo papel da soma dos quadrados dos resíduos nos modelos normais-lineares, com a distribuição χ^2_{n-p}. O desvio pode ainda ser usado para comparar MAGs encaixados da mesma maneira que nos MLGs, sendo, portanto, equivalente ao teste da razão de verossimilhança. Por exemplo, suponha que η_1 e η_2 são dois modelos lineares seguindo (4), sendo η_1 encaixado em η_2. Segundo condições gerais de regularidade e supondo que η_1 é verdadeiro, a diferença $D_{p1} - D_{p2}$ tem distribuição qui-quadrado assintoticamente, onde D_{pi} é o desvio do modelo definido por η_i.

Pode-se demonstrar que o desvio de um MAG pode ser escrito em forma quadrática, isto é, por

$$D_p = (y - \hat{\mu})^T M(y - \hat{\mu}), \tag{5}$$

onde M é a matriz de pesos usada no algoritmo de ajustamento. A expressão (5) mostra a evidência do papel de D_p nos MAGs análogo ao da

soma dos quadrados dos resíduos no modelo clássico de regressão. Maiores detalhes das técnicas de inferência nos MAGs podem ser vistos em Hastie e Tibshirani (1990).

4. UMA GENERALIZAÇÃO DOS MODELOS DE BOX E COX

Define-se aqui, seguindo Cordeiro e Cribari-Neto (1994), uma classe de modelos indexada por duas transformações do tipo potência que de alguma forma generaliza os modelos de Box e Cox (1964). Esta classe expressa por

$$y_i^* = (y_i^{\lambda_1} - 1)/\lambda_1 \sim FE(\theta_i, \phi),\ (\mu_i^{\lambda_2} - 1)/\lambda_1 = \eta_i, \qquad (6)$$

para $i = 1, \cdots, n$, onde $E(y^*) = \mu$ e $\eta = X\beta$. Assim, os dados brutos $y_1, \cdots, y_n$ são transformados por um parâmetro λ_1, produzindo dados modificados $y_1^*, \cdots, y_n^*$ que devem seguir alguma distribuição na família exponencial de distribuições (1), com as médias dos $y_i^{*}s$ sendo ainda transformadas por um parâmetro λ_2 de tal maneira a produzir o preditor linear $\eta = X\beta$, isto é, $E(y_i^*) = \mu_i$ e $(\mu_i^{\lambda_2} - 1)/\lambda_2 = \eta_i$.

Esta classe de modelos engloba como casos especiais vários modelos importantes. Os próprios MLGs são definidos por $\lambda_1 = 1$ e os modelos de Box e Cox por $\lambda_2 = 1$ supondo que $F(\theta, \phi)$ é a distribuição normal. Quando $\lambda_1 = 0$ e $\lambda_2 = -1$ admite-se erros da família exponencial na escala logaritmica e linearidade na escala inversa. A grande desvantagem prática dos modelos de Box e Cox em relação à classe (6) é exigir um único λ produzindo dois efeitos: normalidade do erro e linearidade da componente sistemática.

Uma subclasse dos modelos (6) de interesse é deduzida supondo que os dados transformados são normais. Quando $\lambda_1 = \lambda_2 = 1$ obtém-se o modelo clássico de regressão. Outro caso especial é definido por $\lambda_1 = 0$ e $\lambda_2 = -1$ e, portanto, supõe erros normais na escala logaritmica e linearidade na escala inversa, sendo equivalente ao modelo $N(\mu, \sigma^2)$ com ligação $\exp(-\mu) = 1 - \eta$.

A classe de modelos definida em (6) pode ser representada no contexto dos MLGs pela família $F(\theta, \phi)$ com função de ligação aproximada

$$\eta = \{(1+\lambda_1\mu)^{\lambda_2/\lambda_1} - 1\} / \lambda_1, \qquad (7)$$

supondo que λ_1 e λ_2 são não-nulos. A equação (7) possibilita ajustar os modelos (6) via o GLIM.

Sejam $\hat{\mu}_1, \cdots, \hat{\mu}_n$ as estimativas de máxima verossimilhança das médias dos y^*'s baseadas nos dados originais. Estas estimativas são obtidas maximizando-se a log-verossimilhança tratada como função de β e de λ_1 e λ_2, cuja expressão $l = l(\lambda_1, \lambda_2)$ iguala

$$l = \sum_{i=1}^{n} \exp[\phi\{y_i\theta_i - b(\theta_i) + c(y_i, \phi_i)\}] + (\lambda_1 - 1)\sum_{i=1}^{n} \log y_i.$$

A maximização da função acima é feita utilizando técnicas iterativas. Em geral, fixa-se λ_1 e λ_2 e maximiza-se em relação a β. A escolha destes parâmetros de transformação visa a obter a maior log-verossimilhança maximizada. Para um certo conjunto de dados y pode-se, então, ajustar vários modelos definidos em (6) para valores discretos adequados dos parâmetros λ_1 e λ_2, escolhendo aqueles $\hat{\lambda}_1$ e $\hat{\lambda}_2$ que produzem a maior log-verossimilhança maximizada $\hat{l} = l(\hat{\lambda}_1, \hat{\lambda}_2)$ em relação aos parâmetros lineares.

No caso da distribuição em (1) ser normal, as log-verossimilhanças maximais são obtidas da expressão

$$\hat{l} = -\frac{n}{2}\log\hat{\sigma}^2(\lambda_1, \lambda_2) + (\lambda_1 - 1)\sum_{i=1}^{n} \log y_i, \qquad (8)$$

onde $\hat{\sigma}^2$, como função de λ_1 e λ_2, iguala à soma dos quadrados dos resíduos (isto é, o desvio do modelo) dividido por n. A adequação global do modelo, correspondente aos valores escolhidos $\hat{\lambda}_1$ e $\hat{\lambda}_2$, pode ser verificada traçando-se os gráficos de y^* versus $\hat{\mu}$ e de y versus $\hat{E}(y)$, onde $E(y)$ é deduzido de expansão em série de Taylor, que resulta em

$$E(y) = (\lambda_1\mu + 1)^{1/\lambda_1} + \frac{1}{2}\sigma^2(1-\lambda_1)(\lambda_1\mu + 1)^{(1-2\lambda_1)/\lambda_1}.$$

Para ilustrar uma aplicação da classe de modelos introduzida nesta seção, consideramos o exemplo tratado em Cordeiro e Cribari-Neto (1994),

referente à análise de 25 dados de mobilidade social apresentados na Tabela 1, onde y_{ij} representa a freqüência observada de famílias com o pai tendo a profissão "i" e o filho a profissão "j". A estrutura linear do modelo é definida por $\eta_{ij} = \beta + pai(i) + filho(j), \quad i,j = 1,\cdots,5$, onde os parâmetros lineares pai(i) e filho(j) são os efeitos das profissões i e j do pai e filho, respectivamente, e β uma média geral. Um modelo log-linear de independência de linhas e colunas seria fortemente rejeitado, isto é, y_{ij} com distribuição de Poisson $P(\mu_{ij})$ e $\log(\mu_{ij}) = \eta_{ij}$. Este fato conduziria ao ajustamento do modelo saturado com 25 parâmetros correspondentes às médias desconhecidas que apenas reproduziria os próprios dados. Propõe-se, então, um modelo normal definido em (6) com a escolha de λ_1 e λ_2 visando a obter a maior log-verossimilhança maximizada (8) em relação aos parâmetros lineares. A vantagem maior deste modelo é que ele tem bem menos parâmetros (9 parâmetros lineares e dois de tranformação) sem conduzir a grandes dificuldades na sua interpretação.

TABELA 1: MOBILIDADE SOCIAL NA GRÃ-BRETANHA
Freqüências de profissões pai/filho, com os 5 níveis seguintes: 1-executivo (mais alto), 2-subordinado ao executivo, 3-administrativo, 4-profissional habilitado e 5-sem habilitação (mais baixo)

		Filho				
		1	2	3	4	5
Pai	1	50	45	8	18	8
	2	28	174	84	154	55
	3	11	78	110	223	96
	4	14	150	185	714	447
	5	3	42	72	320	411

A Tabela 2 apresenta valores para esses parâmetros de transformação com as correspondentes log-verossimilhanças maximais. Desta tabela concluímos que o modelo com maior log-verossimilhança maximizada corresponde à transformação da raiz quadrada $(\lambda_1 = 0.5)$ dos dados para produzir normalidade com ligação inversa $(\lambda_2 = -1)$. As estimativas na estrutura linear (os erros padrão estão entre parênteses), considerando uma parametrização para os parâmetros lineares com pai(1) = filho(1) = 0, são: β = 0.29 (0.10), pai(2) = -0.09 (0.06), pai(3) = -0.10 (0.06), pai(4) = -0.12 (0.06), pai(5) = -0.11 (0.06), filho(2) = -0.12 (0.08), filho(3) = -0.12 (0.08), filho(4) = -0.14 (0.08) e filho(5) = -0.14 (0.08). A significância dos parâmetros não pode aqui ser medida pelo Teste de Nulidade de Wald, isto é, dividindo-se as estimativas pelos erros padrão correspondentes, pois este teste é uma aproximação de primeira ordem. O

procedimento correto é fazer o teste pela redução no desvio provocada pelo acréscimo do parâmetro de interesse. Alguns dos parâmetros foram significativos com este teste e outros não. Entretanto, o teste global do modelo implica na sua aceitação. Com o intuito de se ter uma melhor interpretação do modelo todos os parâmetros foram mantidos. Observa-se que, quando λ_1 é negativo, a log-verossimilhança maximizada praticamente independe de λ_2.

TABELA 2- VEROSSIMILHANÇAS MAXIMIZADAS PARA VÁRIOS MODELOS NORMAIS DEFINIDOS EM (6)

	λ_1=	-1.5	-1.0	-0.5	0	0.5	1.0	1.5
	-1.0	-164.8	-136.4	-112.8	-96.11	-89.85	-95.13	-135.8
	-0.5	-164.7	-136.4	-112.9	-96.89	-90.97	-95.95	-109.8
λ_2=	0	-164.6	-136.3	-113.0	-97.91	-93.05	-98.01	-111.7
	0.5	-164.6	-136.2	-113.3	-99.24	-93.48	-101.6	-117.5
	1.0	-164.5	-136.1	-113.4	-100.9	-102.6	-116.8	-135.8

Torna-se importante enfatizar que um modelo mais adequado poderia ser encontrado através de um estudo detalhado dos resíduos, implicando em procurar distribuições alternativas dentro da família $F(\theta,\phi)$. Dos valores ajustados ($\hat{\mu}_{ij}$) aos dados originais pode-se calcular as probabilidades de transição do tipo

$$\text{Prob[filho(j)|pai(i)]}= \hat{\mu}_{ij} / \hat{\mu}_{i+} ,$$

para $i,j = 1,\cdots,5$, onde $\hat{\mu}_{i+} = \mu_{i1}+\cdots+\hat{\mu}_{i5}$ e, portanto, determinar a mobilidade social. Por exemplo, no caso, se o pai pertence a classe social 3 existem chances de 22% e 20% do seu filho retroagir às classes 1 e 2, respectivamente, chance de 20% do filho continuar na classe 3 e chances de 19% de passar para cada umas das classes 4 ou 5.

5. MODELOS NÃO-EXPONENCIAIS

Os **modelos não-exponenciais** (Cordeiro e Paula, 1989a; 1992) de regressão, embora não tão amplamente usados como os modelos exponenciais, têm despertado grande interesse nos últimos anos, principalmente, pelo desenvolvimento de diversos *softwares* estatísticos de ajustamento. Estes modelos também são denominados de **modelos de**

dispersão (Jorgensen, 1987a,b), porque quando $a(\phi)$ (vide equação (9)) é constante, este parâmetro desempenha um papel de dispersão análogo ao da variância do modelo normal-linear. Aqui, apresentamos vários modelos não-exponenciais, supondo que os dados $y_1,\cdots,y_n$ são independentes e seguem a família de distribuições

$$f(y;\theta_i,\phi_i) = \exp\{t(y,\theta_i)/a(\phi_i) + c(y,\phi_i)\}, \qquad (9)$$

onde $a(\cdot)$, $t(\cdot,\cdot)$ e $c(\cdot,\cdot)$ são funções bem comportadas, conhecidas, com $a(\phi_i) > 0$, suposto conhecido para $i = 1,\cdots,n$. Os modelos dados por (9) são chamados modelos não-exponenciais desde que $t(y,\theta)$ não tenha a forma $y\theta - b(\theta)$, pois a igualdade conduz aos modelos exponenciais definidos em (1). Jorgensen (1987a,b) denomina também estes últimos modelos de **modelos exponenciais de dispersão**.

A componente sistemática do modelo definido por (9) relaciona o parâmetro θ com um vetor β de dimensão p

$$h(\theta) = \eta = g(X;\beta), \qquad (10)$$

onde X é uma matriz de constantes conhecidas e h(.) é um tipo de função de ligação que expressa o parâmetro θ em termos do preditor η. Diversas aplicações dos modelos definidos pelas equações (9) e (10) a dados reais estão em Cordeiro (1986), Cordeiro e Paula (1989a,b,c,1992) e Jorgensen (1983,1984,1986,1987a,b).

Fica claro denominarmos os modelos definidos por (9) e (10) de modelos não-exponenciais não-lineares para corresponder com a não-exponencialidade da distribuição do erro e a não-linearidade do vetor η. Propriedades dos modelos não-exponenciais podem ser vistas em Jorgensen (1983, 1986, 1987a,b), com bastante ênfase na teoria assintótica. Cordeiro e Paula (1989a, 1992) apresentam um algoritmo de ajustamento destes modelos via o GLIM, usando as potencialidades da diretiva *offset*.

Vários modelos não-lineares importantes desenvolvidos recentemente são casos especiais de (9), como, por exemplo: modelos da forma $Y = X\beta + \sigma\varepsilon$, onde os ε 's embora independentes podem ter qualquer distribuição de probabilidade (Stirling, 1984), modelos de regressão para dados ordinais (McCullagh, 1980), modelos com função de ligação composta com $E(Y) = \mu = C\gamma$, sendo γ uma função não-linear de β e, ainda, modelos definidos pela estimação condicional e marginal da verossimilhança (Jorgensen, 1984).

Os exemplos, a seguir, são casos especiais de modelos definidos por (9) e (10):

a) modelo log-gama com densidade igual a

$$f(y,\theta,\phi) = c(\phi)\exp\left[\phi\{y-\theta-\exp(y-\theta)\}\right],\quad y>0,$$

onde $c(\phi)$ é uma função normalizadora;

b) modelo hiperbólico com densidade

$$f(y;\theta) = y^{-1}\exp\left[-\{y^{-1}\exp(\theta)+y\exp(-\theta)\}/2\right]/2K_0(1),$$

onde $K_v(\cdot)$ é a função de Bessel com índice v;

c) modelo log-normal inverso generalizado com

$$t(y,\theta) = \alpha(y-\theta)-\beta\cosh(y-\theta),$$

supondo α e β conhecidos;

d) família de modelos da forma

$$f(y,\theta,\phi) = \delta\phi^{1/\delta}\exp\left\{-\phi|y-\theta|^{\delta}\right\}\Big/\Gamma(\delta^{-1}),$$

sendo $\delta \neq 2$, onde $\Gamma(\delta) = \int_0^{\infty} x^{\delta-1}e^{-\delta x}dx$. Para $\delta = 1$ tem-se o modelo de Laplace;

e) modelo logarítmico com densidade definida por

$$t(y,\theta) = y\log\theta - \log\left\{-\log(1-\theta)\right\},\ y = 1,2,\cdots \text{ e } 0<\theta<1;$$

f) modelo em série de potência com densidade

$$f(y,\theta) = \exp\left\{\log a_y + y\log\theta - \log b(\theta)\right\},\quad y = 0,1,2,\cdots,\quad \theta>0,$$

$a_y \geq 0$ e $b(\theta) = \sum_{y=0}^{\infty} a_y\theta^y$;

g) modelo beta-BETA $\left(\phi\theta,\phi(1-\theta)\right)$ com média θ e ϕ como um parâmetro de dispersão;

h) modelo de Von Mises definido por $t(y,\theta) = \cos(y-\theta)$.

Supondo agora que $t(y,\theta)$ envolve um parâmetro c conhecido para cada observação, sendo $t(y,\theta) = t(y,\theta,c)$, $p(\phi) = \phi = 1$ e $q(y,\phi) = q(y,c)$, pode-se ainda definir os seguintes modelos especiais no contexto dos modelos especificados por (9) e (10):

i) normal - $N\left(\theta,c^2\theta^2\right)$, log-normal - $LN(\theta,c^2\theta^2)$, normal inverso - $N^{-}(\theta,c^2\theta^2)$ com média θ e coeficiente de variação c;

j) gama - $G(\theta,c)$ com média θ e parâmetro de escala c;

l) Weibull - $W(\theta,c)$ com média θ e parâmetro de forma c; e

m) todos os modelos que dependem de um único parâmetro θ desde que $t(y,\theta)$ tenha qualquer forma diferente de $y\theta - b(\theta)$.

O leitor deve notar que os modelos descritos em (i) e (j) não pertencem à classe dos MLGs, pois, em cada caso, usa-se uma parametrização diferente. Todos os modelos não-exponenciais (a) a (m) podem ter componente sistemática linear $\eta = X\beta$ ou não-linear $\eta = g(X;\beta)$ para os parâmetros β's e, portanto, são chamados de modelos não-exponenciais lineares e modelos não-exponenciais não-lineares, respectivamente. Resultados assintóticos para os modelos exponenciais lineares e não-lineares podem ser encontrados em Cordeiro (1983, 1985, 1987) e Cordeiro e Paula (1989b, 1992), respectivamente. Vide ainda Cordeiro e Paula (1989b, 1989c), Cordeiro e Ferrari (1991), Cordeiro e McCullagh (1991), Cordeiro, Ferrari e Paula (1993) e Cordeiro, Botter e Paula (1994).

Para os modelos não-exponenciais, supondo que todas as observações, têm o mesmo parâmetro ϕ, a log-verossimilhança tem a forma

$$l(\beta,\phi) = t(y;\theta) / a(\phi) + c(y;\phi),$$

onde $t(y;\theta)=\sum_{i=1}^{n} t(y_i,\theta_i)$ e $c(y;\phi)=\sum_{i=1}^{n} c(y_i,\phi)$, sendo $\theta=\theta(\beta)$ uma função de β através da estrutura de ligação. Aqui β e ϕ são parâmetros ortogonais (Cox e Reid, 1987), pois $E\{-\partial^2 l(\beta,\phi)/\partial\beta\,\partial\phi\}=0$. Neste caso, a matriz de informação particionada em β e ϕ é bloco-diagonal. Quando o parâmetro de dispersão não é constante, a log-verossimilhança para os parâmetros β 's fica expressa por

$$l(\beta)=\sum_{i=1}^{n} t(y_i,\theta_i)/a(\phi_i)+\sum_{i=1}^{n} c(y_i,\phi_i) \qquad (11)$$

sendo θ uma função de β definida pela componente sistemática (10). Como as observações são independentes $l(\beta)$ é aditiva e iguala à soma de n contribuições: $l(\beta)=\sum_{i=1}^{n} l\{\theta_i(\beta);y_i\}$.

5.1. Algoritmo de Ajustamento

A função escore $U(\beta)=\partial l(\beta)/\partial\beta$ é dada por $U(\beta)=\tilde{X}^T\Phi WH\tilde{t}$, onde $\Phi=diag\{a(\phi)^{-1}\}$, $W=diag\{-D_2(\theta)(d\theta/d\eta)^2\}$, $H=diag\{-D_2(\theta)^{-1}\,d\eta/d\theta\}$, com $D_2(\theta)=E\{t''(y,\theta)\}$, $\tilde{t}=\left(t'(y_1,\theta_1),\cdots,t'(y_n,\theta_n)\right)^T$ e $\tilde{X}=\partial\eta/\partial\beta$ sendo uma matriz n x p com elementos $\partial\eta_i/\partial\beta_r$. No caso de g(.;.) ser linear, $\tilde{X}$ iguala à matriz de planejamento X. Quando $t(y,\theta)=y\theta-b(\theta)$, W e H reduzem-se às definições usuais dos modelos lineares generalizados e $\tilde{t}=y-\mu$..

Seja $\hat{\beta}$ a estimativa de máxima verossimilhança de β. As equações $U(\hat{\beta})=0$ são, em geral, não-lineares e a solução $\hat{\beta}$ deve ser obtida por procedimentos iterativos, tais como:

$$K(\beta^{(m)})\beta^{(m+1)}=\tilde{X}^{(m)T}\Phi W^{(m)}y^{*(m)}; \qquad (12)$$

com $y^* = X\tilde{\beta} + H\tilde{\iota}$, onde $E\{J(\beta)\} = K(\beta) = \tilde{X}^T W\Phi\tilde{X}$ e

$$J(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta\ \partial\beta^T} = -\sum_{i=1}^{n}\frac{\partial l(\beta)}{\partial\eta_i}\frac{\partial^2\eta_i}{\partial\beta\ \partial\beta^T} - \left(\frac{\partial\eta}{\partial\beta}\right)^T\frac{\partial^2 l(\beta)}{\partial\eta\ \partial\eta^T}\frac{\partial\eta}{\partial\beta}$$

é a matriz das derivadas de segunda ordem da log-verossimilhança com sinal menos. Assim, $K(\beta)$ é a matriz de informação de Fisher. O processo iterativo (12) é deduzido da expansão de $U(\beta)$ em série de Taylor até primeira ordem. Deste procedimento iterativo obtém-se as *n* estimativas $\hat{\theta}_1,\cdots,\hat{\theta}_n$.

A solução das equações de máxima verossimilhança para o modelo não-exponencial não-linear equivale, portanto, a calcular repetidamente uma regressão linear ponderada de uma variável dependente modificada y^* sobre a matriz $\tilde{X}$, com esta matriz e a função de peso $W\Phi$ se modificando no processo iterativo. Nota-se que $Cov(y^*) = W^{-1}\Phi^{-1}$. A inicialização do processo iterativo pode ser feita a partir das estimativas $\hat{\hat{\theta}}_i$, $i = 1,\cdots,n$ correspondentes ao modelo saturado, isto é, sem a componente sistemática (10). Observa-se que $\hat{\hat{\theta}}_i$ é a solução da equação $l'(y_i,\hat{\hat{\theta}}_i) = 0$, para $i = 1,\cdots,n$. Entretanto, para os modelos não-lineares, torna-se necessário uma escolha adicional para β com o objetivo de inicializar $\tilde{X}$. A implementação do algoritmo anterior no GLIM está dada em Cordeiro e Paula (1989a, 1992).

Jorgensen (1984) define uma classe de algoritmos iterativos para calcular as estimativas de máxima verossimilhança de um modelo exponencial de dispersão, supondo que a log-verossimilhança é escrita na forma $l\{\eta(\beta)\}$, por

$$\beta^{(m+1)} = \left(D^{(m)T}A^{(m)}D^{(m)}\right)^{-1}D^{(m)T}A^{(m)}v^{(m)} \qquad (13)$$

onde os sobre escritos indicam as etapas do processo iterativo, $D = \partial\eta/\partial\beta$ é a matriz de planejamento local, $A = E\{-\partial^2 l(\beta)/\partial\eta\partial\eta^T\}$ é a matriz de informação de Fisher e $v = D\beta + A^{-1}u$ é uma variável dependente modificada local, sendo

$u = \partial l(\beta) / \partial \beta$ o vetor escore $n \times 1$. Uma discussão geral de escolhas alternativas da matriz A visando a robustecer o processo iterativo (13) pode ser encontrada em Jorgensen (1984). No caso do parâmetro ϕ ser constante este não afeta as estimativas dos β's.

5.2. Testes de Adequação

Considere agora um modelo M_p não-exponencial não-linear definido por (9) e (10) com p parâmetros. Sejam $\hat{l}_p$ o máximo da log-verossimilhança segundo M_p, cujas estimativas são $\hat{\theta}_1, \cdots, \hat{\theta}_n$ e $\hat{l}_n$ a log-verossimilhança maximizada segundo o modelo saturado, cujas estimativas são $\hat{\hat{\theta}}_1, \cdots, \hat{\hat{\theta}}_n$. Para o teste de adequação do modelo ajustado usa-se uma das estatísticas:

$$D_p = 2\sum_{i=1}^{n} \left\{ t\left(y_i, \hat{\hat{\theta}}_i\right) - t\left(y_i, \hat{\theta}_i\right) \right\} \Big/ a(\phi_i)$$

ou

$$X_p^2 = \sum_{i=1}^{n} \left\{ y_i - \hat{E}(y_i) \right\}^2 \Big/ \hat{Var}(y_i),$$

onde $\hat{E}(y_i)$ e $\hat{Var}(y_i)$ são as estimativas de máxima verossimilhança, segundo M_p, da média e da variância de y_i. Qualquer uma dessas estatísticas mede a discrepância entre os dados $y_1, \cdots, y_n$ e os seus valores ajustados, sendo D_p uma óbvia extensão do desvio do MLG, denominado aqui, também, de desvio do modelo.

Para usarmos as estatísticas acima torna-se necessário conhecer alguma estimativa consistente do parâmetro ϕ, como, por exemplo, a sua estimativa de máxima verossimilhança, calculada de

$$a(\hat{\phi})^2 \sum_{i=1}^{n} c'\left(y_i; \hat{\phi}\right) - a'(\hat{\phi}) \sum_{i=1}^{n} t\left(y_i; \hat{\theta}_i\right) = 0.$$

Esta equação, em geral não-linear, pode ser expressa igualando uma certa função de $\hat{\phi}$ e dos dados ao desvio do modelo M_p. Uma solução

explícita para $\hat{\phi}$ é raramente possível, como no modelo (d) de Laplace apresentado nesta seção, que vale $\hat{\phi} = -\delta^{-1}\sum_{i=1}^{n}\left|y_i - \hat{\theta}_i\right|^{\delta}$.

Para o teste do modelo M_p, definem-se os seus graus de liberdade por $\nu = n - p$. O modelo será rejeitado se os valores das estatísticas D_p e X_p^2 forem superiores ao ponto crítico $\chi^2_{n-p}(\alpha)$ da distribuição χ^2_{n-p} correspondente ao nível de significância α. Este teste é aproximado e um aperfeiçoamento através do cálculo do fator de correção de Bartlett foi desenvolvido, recentemente, por Cordeiro, Botter e Paula (1994).

As medidas usuais de diagnóstico nos MLGs (vide Cordeiro, 1986 e McCullagh e Nelder, 1989) podem ser estendidas para os modelos não-exponenciais. Os resíduos são definidos como raízes quadradas das componentes do desvio D_p, com o sinal dado pela diferença $y_i - \hat{\mu}_i$. A análise gráfica desses resíduos permite detectar anomalias locais no modelo. O efeito local (influência) de uma observação j sobre o modelo é medido pela diferença $\hat{\beta} - \hat{\beta}_{(j)}$, onde $\hat{\beta}_{(j)}$ representa a estimativa de β sem a observação j. Em geral, teremos de ajustar r+1 modelos caso hajam r pontos a pesquisar as suas influências.

6. MODELOS DE QUASE-VEROSSIMILHANÇA

Nos modelos de quase-verossimilhança as variáveis são consideradas independentes, sem ser necessário especificar qualquer distribuição para erro estocástico, e a componente sistemática é dada por

$$E(y_i) = \mu_i(\beta), \quad Var(y_i) = \phi V_i(\mu_i). \tag{14}$$

Nas relações (14) os μ_i's são funções conhecidas dos regressores, os V_i's são funções conhecidas das médias desconhecidas (em geral $V_i(.) = V(.)$ ou $V_i(.) = a_i V(.)$ para valores conhecidos dos a_i's) onde de agora por diante ϕ representará um parâmetro de dispersão (e não mais de precisão), possivelmente desconhecido, podendo ainda ser uma função de regressores adicionais. Usualmente, $\mu_i(\beta)$ equivale à componente sistemática do MLG.

Define-se a *log-quase-verossimilhança* para uma única observação, supondo apenas que a média e a variância da distribuição existem, por

$$Q = Q(y;\mu) = \phi^{-1} \int (y-\mu) V(\mu)^{-1} \, d\mu . \qquad (15)$$

O método de quase-verossimilhança generaliza o método de mínimos quadrados, pois este supõe $V(\mu_i)$ constante. Os modelos de quase-verossimilhança são equivalentes aos MLGs com as seguintes funções de variância:

$V_i(\mu_i)$	Distribuição
constante	normal
μ_i	Poisson
$\mu_i(1-\mu)$	binomial
μ_i^2	gama
$\mu_i + k\mu_i^2 (k>0)$	binomial negativa
μ_i^3	normal inversa

Wedderburn (1974) demonstrou que a quase-verossimilhança tem propriedades semelhantes à verossimilhança como, por exemplo,

$$E\{\partial Q / \partial \mu\} = 0, \; E\left\{[\partial Q / \partial \mu]^2\right\} = -E\{\partial^2 Q / \partial \mu^2\} = 1 / \{\phi V(\mu)\}. \qquad (16)$$

Uma terceira propriedade importante entre os logaritmos da verossimilhança l e da quase-verossimilhança Q, supondo para ambos uma mesma função de variância, é dada por

$$-E\{\partial^2 Q / \partial \mu^2\} \leq -E\{\partial^2 l / \partial \mu^2\} \qquad (17)$$

Se y seguir a família exponencial (1) de distribuições tem-se $V(\mu) = d\mu / d\theta$ e, portanto, $Q = \phi^{-1} \int (y-\mu) d\theta$. Como $\mu = b'(\theta)$, então, Q tem expressão idêntica à log-verossimilhança da distribuição de y. A igualdade em (17) somente ocorre no caso de l ser a log-verossimilhança da família exponencial. O lado esquerdo de (17) é uma medida da informação quando se conhece apenas a relação entre a variância e a média dos dados, enquanto o lado direito é a informação usual de Fisher obtida pelo conhecimento da distribuição dos dados. A quantidade não-negativa $E\{\partial^2 (Q-1) / \partial \mu^2\}$ é a informação que se ganha quando ao conhecimento da relação variância-média se acrescenta a informação da forma da distribuição dos dados. A suposição dos dados pertencer à família exponencial equivale, portanto, à informação minimal

obtida do simples conhecimento da relação funcional variância-média dos dados. O leitor interessado em aplicações a dados reais dos modelos de quase-verossimilhança poderá consultar Nelder (1985), Nelder e Pregibon (1987), Cordeiro e Demétrio (1989) e McCullagh e Nelder (1989, Capítulos 9 e 10).

A log-quase-verossimilhança para *n* observações iguala à soma de *n* contribuições definidas por (15). As estimativas de máxima quase-verossimilhança $\tilde{\beta}_1,\cdots,\tilde{\beta}_p$ são obtidas maximizando esta soma. Supondo ϕ constante para os *n* dados $y_1,\cdots,y_n$, chega-se ao seguinte sistema de equações para os $\tilde{\beta}$'s

$$\sum_{i=1}^{n}\left(y_i-\mu_i\right)\cdot\left(\partial\mu_i/\partial\beta_r\right)/V_i\left(\mu_i\right)=0 \qquad (18)$$

para $r=1,\cdots,p$, as quais independem de ϕ .

A maximização da quase-verossimilhança generaliza o método de mínimos quadrados, correspondente ao caso de $V(\mu)$ constante. Pode-se demonstrar (McCullagh, 1983) que as equações de máxima quase-verossimilhança produzem as melhores estimativas lineares não-viesadas, o que representa uma generalização do teorema de Gauss-Markov. Os modelos de quase-verossimilhança podem ser ajustados facilmente usando macros especiais de *softwares* como GENSTAT, GLIM ou SAS.

Na análise de dados em forma de contagens trabalha-se com o erro de Poisson supondo que $Var(Y_i)=\phi\,\mu_i$. O parâmetro ϕ é estimado igualando a razão de quase-verossimilhança $2\left[Q(y;y)-Q(y;\tilde{\mu})\right]$ aos graus de liberdade $n-p$ da distribuição χ^2 de referência, ou então, através da fórmula

$$\tilde{\phi}=(n-p)^{-1}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}_i\right)^2/\tilde{\mu}_i \qquad (19)$$

Os dados apresentarão superdispersão se $\tilde{\phi}>1$ e subdispersão no caso contrário. Assim, dados contínuos que representam durações de tempo com superdispersão podem ser modelados por $Var(Y_i)=\phi\,\mu_i^2$ supondo $\phi>1$ e dados na forma de contagens com superdispersão por

$V(\mu) = \mu + \lambda\ \mu^2$ (binomial negativa) ou por $V(\mu) = \mu + \lambda\,\mu + \gamma\ \mu^2$. Para proporções usa-se $V(\mu\ ) = \mu\,(1-\mu\,)$ ou $\mu^2(1-\mu)^2$.

A definição da quase-verossimilhança (15) permite fazer comparações de modelos com preditores lineares diferentes ou com funções de ligação diferentes. Entretanto, não se pode comparar, com os mesmos dados, funções de variância diferentes. Nelder e Pregibon (1987) propuseram uma definição de *log-quase-verossimilhança estendida* Q^+ , a partir da variância e da média dos dados, que permite fazer esta comparação, cuja fórmula é

$$Q^+ = -1/2\sum_i \log\left\{2\pi\phi_i V(y_i)\right\} - 1/2\sum_i D(y_i;\mu_i)/\phi_i \qquad (20)$$

Em (20) o somatório varre todas as observações e a função $D(y;\mu)$ denominada de *quase-desvio*, é uma simples extensão do desvio do MLG, definida para uma observação (Nelder, 1985) por

$$D(y;\mu) = -2\int_y^{\mu}(y-x)V(x)^{-1}dx \qquad (21)$$

isto é, $D(y;\mu) = 2\phi\left[Q(y;y) - Q(y;\mu)\right]$. A função quase-desvio para os dados iguala $\sum_i D(y_i;\tilde{\mu}_i)$. Para as funções de variância dos MLGs, a função quase-desvio reduz-se aos desvios destes modelos.

6.1. Modelo de quase-verossimilhança com função de variância paramétrica

Admite-se, agora, o seguinte modelo de quase-verossimilhança com função de variância paramétrica

$$E(Y_i) = \mu_i(\beta), \quad Var(Y_i) = \phi\ V_\lambda(\mu_i), \qquad (22)$$

onde λ é um parâmetro desconhecido na função de variância. Um exemplo de função de variância paramétrica é a função potência definida por $V_\lambda(\mu) = \mu^\lambda$, $\lambda \geq 0$, que engloba as variâncias das distribuições normal, Poisson, gama e normal inversa. Para um valor fixo de λ pode-se, ainda, utilizar as equações dadas em (18). A escolha da estimativa de λ visa a minimizar a função desvio $\sum_{i=1}^{n} D_\lambda(y_i;\tilde{\mu}_i)$, onde

$$D(y;\mu) = -2\int_{y}^{\mu}(y-\mu)V_{\lambda}(\mu)^{-1}d\mu \qquad (23)$$

Uma situação em que ocorre, naturalmente, a função de variância paramétrica corresponde ao preditor linear $\eta = X\beta$ com uma componente aleatória independente extra ε de variância λ , produzindo o preditor modificado $\eta^{*} = \eta + \varepsilon$. Até primeira ordem, obtém-se a média e a variância modificadas $E(y)^{*} = \mu + \varepsilon\, d\mu / d\eta$ e $Var(y)^{*} = \phi\, V(\mu) + \lambda\,(d\mu / d\eta)^{2}$ e , portanto, a função de variância torna-se parametrizada por λ . Uma outra situação diz respeito a uma variável resposta Y , representando uma soma de variáveis IID, cujo número de variáveis é também uma variável aleatória de média μ e variância $V(\mu)$. É fácil verificar que os parâmetros extras que aparecem na função de variância de Y incluirão os dois primeiros momentos das variáveis IID.

Para um valor fixo de λ pode-se, ainda, utilizar as equações dadas em (18) para obter as estimativas de máxima quase-verossimilhança dos β's. A estimativa de λ corresponderá ao maior valor da quase-verossimilhança estendida maximizada, obtida da expressão (20), tratada como função de λ , ou ainda ao menor valor do *desvio estendido* $-2\,Q^{+}(\lambda)$ dado por

$$\min_{\lambda} -2Q^{+}(\lambda).$$

Seria melhor maximizar conjuntamente Q^{+} em relação a β e λ , embora este processo exija o cálculo da função escore em relação ao parâmetro λ , o que é bastante complicado. Sendo assim, considera-se o ajustamento do modelo com função de variância $V(\mu_i) = \mu_i + \lambda\,\mu_i^{2}$, escolhendo para λ o valor que minimiza o desvio estendido. A função desvio estendida para este modelo é calculada de (20) como

$$-2Q^{+}(\lambda) = \sum_{i}\left\{\log\left[2\pi\left(y_i + \lambda\, y_i^{2}\right)\right] + D(y_i;\tilde{\mu}_i)\right\} \qquad (24)$$

com $D(y_i;\mu_i)$ obtido de (21). Temos, então

$$D(y;\mu) = 2y\log\left\{\frac{(\mu+\alpha)y}{(\mu+\alpha)\mu}\right\} + 2\alpha\log\left\{\frac{\mu+\alpha}{y+\alpha}\right\} \qquad (25)$$

com $\alpha = 1/\lambda$.

Para ilustrar uma aplicação das equações (24) e (25), considera-se os dados da Tabela 3 representando os números de falhas em 32 peças de tecidos de uma fábrica, de comprimentos variáveis (Hinde, 1982). O ajustamento do modelo de regressão de Poisson para o número de falhas, com covariável definida pelo logaritmo do tamanho da peça (log(com)), conduz a uma estimativa $\widetilde{\phi}$, obtida de (19), igual a 2.267, implicando que os dados estão superdispersos. Neste ajustamento, o preditor linear estimado (erros padrão entre parênteses) corresponde a

$$\widetilde{\eta} = \log\widetilde{\mu} = -4.17(1.14) + 1.00(0.18)\log(com)$$

TABELA 3- NÚMEROS DE FALHAS EM 32 PEÇAS DE TECIDOS DE COMPRIMENTOS VARIÁVEIS

Comprimento da peça em metros / número de falhas

551/6	**651/4**	**832/17**	**375/9**	**715/14**	**868/8**	**271/5**	**630/7**	**491/7**
372/7	**645/6**	**441/8**	**895/28**	**458/4**	**642/10**	**492/4**	**543/8**	**842/9**
905/23	**542/9**	**522/6**	**122/1**	**657/9**	**170/4**	**738/9**	**371/14**	**735/17**
749/10	**495/7**	**716/3**	**952/9**	**417/2**				

Sendo assim, ajusta-se o modelo com função de variância $V_\lambda(\mu_i) = \mu_i + \lambda\,\mu_i^2$, escolhendo λ como o valor que minimiza o desvio estendido. A função desvio estendida para este modelo é calculada de (24) e (25). A partir do gráfico de $-2Q^+(\lambda)$ versus λ encontraremos a estimativa $\widetilde{\lambda}$ de λ, aproximadamente, igual a 0.13. Para este valor de λ (melhor modelo em $V_\lambda(\mu)$), o preditor linear estimado correspondente é $\widetilde{\eta} = -3.78(3.59) + 0.94(0.56)\log(com)$. O gráfico dos valores observados versus os valores ajustados, segundo este modelo, mostra-se próximo da primeira bissetriz, implicando na aceitação do modelo. Ainda, o gráfico dos resíduos padronizados $(y-\widetilde{\mu})/V_{0.13}(\widetilde{\mu})^{1/2}$ versus $\widetilde{\mu}$, suporta também a adequação do modelo. Este exemplo mostra a grande utilidade de incorporar um parâmetro desconhecido na função de variância do modelo.

6.2. Modelos de quase-verossimilhança com parâmetro de dispersão não-constante

Para esta classe de modelos supõe-se

$$E(Y_i)=\mu_i(\beta),\quad Var(Y_i)=\phi_i(\gamma)V(\mu_i) \tag{26}$$

onde $\phi_i(\cdot)$ são funções conhecidas dos parâmetros desconhecidos γ's e $V(\mu)$ pode ser da forma $V_\lambda(\mu)$. Para γ fixo utiliza-se (18) para calcular as estimativas de quase-verossimilhança dos β's, podendo γ ser escolhido visando a maximizar a quase-verossimilhança estendida definida por Nelder e Pregibon (1987)

$$Q^+(y,\mu)=-1/2\sum_i \log\{2\pi\,\phi_i(\gamma)V_\lambda(y_i)\}-1/2\sum_i D_\lambda(y_i;\mu_i)/\phi_i(\gamma),$$

sendo o somatório sobre todas as observações e a função desvio determinada em (23) para algum λ fixo. A idéia básica é usar $Q^*(y;\mu)$ como uma log-verossimilhança para fazer inferência no modelo (26) sobre β, λ ou γ no modelo (26).

Define-se agora a seguinte classe especial de modelos de quase-verossimilhança com parâmetro de dispersão não-constante

$$\eta=g(\mu)=X\beta,\ \tau=h(\phi)=Z\gamma, \tag{27}$$

onde $\mu_i=E(y_i)$, $Var(y_i)=\phi_i V(\mu_i)$, X e Z são matrizes $n\times p$ e $n\times q$ de posto completo p e q, β e γ são vetores de parâmetros desconhecidos de dimensões $p\times 1$ e $q\times 1$, respectivamente, com $g(.)$ e $h(.)$ funções de ligação conhecidas. Supondo γ fixo pode-se utilizar (18) para calcular a estimativa de quase-verossimilhança de β, e então γ será escolhido visando a maximizar a quase-verossimilhança estendida maximal $Q^+(\gamma)$ como função de γ. A estimativa de γ será o valor correspondente ao maior valor $Q^+(\gamma)$. A idéia básica é usar Q^+ como o análogo da log-verossimilhança na inferência sobre β ou γ.

As componentes *quase-escore* são dadas por

$$U_\beta^+ = \partial Q^+ / \partial \beta = X^T W H(y-\mu), \quad U_\gamma^+ = \partial Q^+ / \partial \gamma = 0.5 Z^T L(D-\phi)$$

onde $W = diag\{\phi^{-1}V(\mu)^{-1}g'(\mu)^{-2}\}$, $H = diag\{g'(\mu)\}$, $L = diag\{\phi^{-2}h'(\mu)^{-1}\}$, e ainda $D = \left(D(y_1;\mu_1),\cdots,D(y_n;\mu_n)\right)^T$. As estimativas de quase-verossimilhança de β e γ são calculadas através do sistema não-linear resultante da igualdade de U_β^+ e U_γ^+ a zero. Este sistema pode ser resolvido por qualquer *software* estatístico-computacional contendo métodos não-lineares de otimização.

Nelder e Pregibon (1987) sugeriram um algoritmo em duas etapas para estimar β e γ que tem o inconveniente de não fornecer as estruturas de covariância assintótica das estimativas de máxima quase-verossimilhança desses parâmetros. Cordeiro e Demétrio (1989) apresentam um algoritmo para estimar conjuntamente esses parâmetros que pode ser implementado facilmente no GLIM. Na dedução deste algoritmo expandiu-se a função desvio

$$D(y,\mu) = -2\int_y^\mu (y-x)V(x)^{-1}dx,$$

em série de Taylor considerando a família potência com expoente fixado visando a obter aproximações razoáveis para os momentos de $D(y;\mu)$. Eles implementaram o algoritmo no GLIM, fazendo uma aplicação na análise de dados reais através de um modelo logístico duplo. Também desenvolveram testes assintóticos para hipóteses compostas sobre os parâmetros β e γ.

7. MODELOS SEMIPARAMÉTRICOS

Os modelos semiparamétricos foram propostos por Green e Yandell (1985) quando definiram o preditor linear η como a parte usual $X\beta$ dos MLGs mais uma parte *s(t)*, onde *s(.)* é alguma função regular cujo argumento *t* pode representar uma medida de distância, tempo, etc. A função *s(.)* é especificada por uma soma $s(t) = \sum_{i=1}^{q} \gamma_i g_i(t)$ de *q* funções básicas $g_1,\cdots,g_q$ sendo γ's os parâmetros desconhecidos. O problema de maximização consiste em definir uma log-verossimilhança penalizada como função dos parâmetros β e γ e maximizá-la, ou seja,

$$\max_{\beta,\gamma}\left\{l[\eta(\beta,\gamma)]-\lambda J[s(\gamma)]/2\right\},$$

onde J[.] é representativo de uma penalidade sobre a não-suavidade de *s(.)* e γ uma constante que indica o compromisso entre a suavidade de *s(.)* e a maximização de $l[\eta(\beta,\gamma)]$. Em geral, admite-se para J[.] a forma quadrática $\gamma^T K\gamma$, com *K* uma matriz de ordem *q* simétrica não-negativa. Se *t* tem dimensão um, a penalidade da não-suavidade da curva *s(t)* iguala $\int\{s''(t)\}^2 dt$ que é uma expressão comumente usada para suavizar uma curva.

Uma outra alternativa para estimar a função *s(t)* é usar um suavizador linear do tipo $s(t_i)=\gamma_{0i}+\gamma_{1i}t_i$, onde esses γ's representam parâmetros ajustados por mínimos quadrados às $n_i=\left([wn]-1\right)/2$ observações de cada lado de t_i e *w* representa a amplitude do suavizador, escolhido distante dos extremos do intervalo $(1/n,2)$. A notação $[x]$ significa a parte inteira do número *x*.

8. MODELOS DE REGRESSÃO COM ESTRUTURA DE AUTOCORRELAÇÃO INTERNA

Os modelos mais comuns com estrutura de autocorrelação interna são os modelos de séries temporais. Os modelos ARMA de Box e Jenkins poderão ser colocados na forma de ajustamento dos MLGs a partir da decomposição das suas matrizes de informação *K* na forma, L^TL supondo que estas matrizes estão particionadas em relação à variância σ^2 e aos parâmetros lineares (parâmetros da parte auto-regressiva e de médias móveis). Esta decomposição pode ser sempre feita numericamente, por programas especiais, acoplando à mesma o algoritmo de ajustamento dos MLGs.

Outros modelos de séries temporais apresentam equações de máxima verossimilhança idênticas às equações (18). Por exemplo, considere a estimação da densidade espectral $f(w;\beta)$ de uma série temporal estacionária y_t, $t=1,\cdots,n.$. Pode-se usar o método de máxima verossimilhança supondo que as ordenadas do periodograma $I(w_i)$, $w_i=2\pi i/n$, $0<i<[n/2]$ são variáveis aleatórias exponenciais independentes. A log-verossimilhança a ser maximizada iguala

$$l(\beta) = -\sum_{i=1}^{n} \{\log f(w_i;\beta) + I(w_i)/f(w_i;\beta)\},$$

o que implica nas equações de máxima verossimilhança $(r = 1,\cdots,p)$ serem do mesmo tipo de (18)

$$\sum_{i=1}^{n} \{[I(w_i) - f(w_i;\beta)]/[f(w_i;\beta)^2]\} \partial f(w_i;\beta)/\partial\beta_r = 0. \qquad (28)$$

Estas equações podem ser resolvidas sem grandes dificuldades no GLIM, particularmente, se a densidade espectral pertencer à família exponencial (1).

9. OUTROS MODELOS ESPECIAIS

Considera-se agora que os dados y_{it}, $i = 1,\cdots,n$, supostos independentes, de n indivíduos no tempo $t = 1,\cdots,k$, com variáveis explicativas $x_{itr}(r = 1,\cdots,p)$, seguem a seguinte estrutura, bastante apropriada para experimentos com medidas repetidas e para dados longitudinais

$$E(y_{it}) = \mu_{it}, Var(y_{it}) = \phi\, V(\mu_{it}), g(\mu_{it}) = \sum_{i=1}^{p} x_{itr}\,\beta_r. \qquad (29)$$

As equações de quase-verossimilhança dadas em (18) podem ser usadas para estimar os parâmetros $\beta's$. Essas estimativas são consistentes e assintoticamente normais, com matriz de covariância que depende da estrutura $V(.)$ especificada para os y's. A estrutura $V(.)$ deverá ser estimada inicialmente, embora esta estimativa inicial exerça, em geral, pouca influência nas estimativas finais dos $\beta's$. Uma forma conveniente para optar por $V(.)$ é definí-la indexada por um pequeno número de parâmetros. Escolhendo *k(k-1)/2* parâmetros tem-se eficiência assintótica global quando *n* tende para infinito.

REFERÊNCIAS

Box, G. E. P. e Cox, D. R. (1964) An analysis of transformations (with discussion). J. Roy. Statist. Soc. B, 26, 211-252.

Cordeiro, G. M. (1983) Improved likelihood ratio statistics for generalized linear models. J. Roy. Statist. Soc. B, 55, 404-413.

Cordeiro, G. M. (1985) The null expected deviance for an extended class of generalized linear models. Lecture Notes in Statistics, 32, 27-34, Springer-Verlag, New-York.

Cordeiro, G. M. (1986) Modelos Lineares Generalizados. Livro Texto, VII Simpósio Nacional de Probabilidade e Estatística, IMPA, Rio de Janeiro.

Cordeiro, G. M. (1987) On the corrections to the likelihood ratio statistics. Biometrika, 74, 265-274.

Cordeiro, G. M, Botter, D. e Paula, G. A. (1994) Improved likelihood ratio tests for dispersion models. Intern. Statist. Rev., 62, 257-276.

Cordeiro, G. M. e Cribari-Neto, F. (1994) On Bartlett corrections, bias reduction and a new class of transformations. Braz. J. Prob. Statist., 7, 179-200.

Cordeiro, G. M. e Demétrio, C. G. B. (1989) Estimation and tests in a quasi-likelihood model with a non-constant dispersion parameter. Lecture Notes in Statistics, 57, 95-104, Springer-Verlag, New York.

Cordeiro, G. M. e Ferrari, S. L. de P. (1991) A modified score test statistic having chi-squared distribution to order n^{-1}. Biometrika, 78, 573-582.

Cordeiro, G. M., Ferrari, S. L. de P. e Paula, G. A. (1993) Improved score tests for generalized linear models. J. Roy. Statist. Soc. B, 55, 661-674.

Cordeiro, G. M. e McCullagh, P. (1991). Bias correction in generalized linear models. J. Roy. Statist. Soc. B, 53, 629-643.

Cordeiro, G. M. e Paula, G. A. (1989a) Fitting non-exponential family nonlinear models in GLIM by using the offset facility. Lecture Notes in Statistics, 57, 105-114, Springer-Verlag, New York.

Cordeiro, G. M. e Paula, G. A. (1989b) Improved likelihood ratio statistics for exponential family non-linear models, Biometrika, 76, 93-100.

Cordeiro, G. M. e Paula, G. A. (1989c) Modelos de Regressão para Análise de Dados Univariados. Livro Texto, XVII Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro.

Cordeiro, G. M. e Paula, G. A. (1992) Estimation, large-sample parametric tests and diagnostics for non-exponential family nonlinear models. Communications in Statistics, Simulation and Computation, 21, 149-172.

Cox, D. R. e Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). J. Roy. Statist. Soc. B, 49, 1-39.

Green, P. J. (1984) Iteratively reweighted least squares for maximum likelihood regression and some robust and resistant alternatives (with discussion). J. Roy. Statist. Soc. B, 46, 149-192.

Green, P. J. e Yandell, B. S. (1985) Semi-parametric generalized linear models. Lecture Notes in Statistics, 32, 44-55, Springer-Verlag, New York.

Hastie, T. J. e Tibshirani, R. J. (1990) Generalized Additive Models. Chapman and Hall, London.

Hinde, J. (1982) Compound Poisson regression models. Lecture Notes in Statistics, 14, 66-77, Springer-Verlag, New York.

Jorgensen, B. (1983) Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika, 70, 19-28.

Jorgensen, B. (1984) The delta algorithm and GLIM. Int. Statist. Rev. 52, 283-300.

Jorgensen, B. (1986) Some properties of exponential dispersion models. Scand. J. Statist. 13, 187-198.

Jorgensen, B. (1987a) Exponential dispersion models (with discussion). Roy. Statist. Soc. B, 49, 127-162.

Jorgensen, B. (1987b) Small-dispersion asymptotics. Bras. J. Prob. Statist., 1, 59-90.

McCullagh, P. (1980) Regression models for ordinal data (with discussion) J. Roy. Statist. Soc. B, 42, 109-142.

McCullagh, P. (1983) Quasi-likelihood functions. Ann. Statist., 11, 59-67.

McCullagh, P. e Nelder, J. A. (1989) Generalized Linear Models, 2nd ed., Chapman and Hall, London.

Nelder, J. A. (1985) Quasi-likelihood and GLIM. Lecture Notes in Statistics, 32, 120-127, Springer-Verlag, New York.

Nelder, J. A. e Pregibon, D. (1987) An extendeded quasi-likelihood function. Biometrika, 74, 221-232.

Payne, C. D. (1986) The GLIM System Release 3.77 Manual, Numerical Algorithms Group, Oxford.

Ratkowsky, D. A. (1983) Nonlinear Regression Modelling: A unified practical approach. Marcel Dekker, Inc., New York and Basel.

Stirling, W. D. (1984) Iteratively reweighted least squares for models with a linear part. Appl. Statist., 33, 7-17.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton Method. Biometrika, 61, 439-447.

RESUMO

Os modelos lineares generalizados são uma extensão dos modelos lineares clássicos e incorporam uma ampla classe de modelos estatísticos de regressão necessários para muitos problemas práticos importantes. Nos últimos dez anos apareceram muitas extensões do procedimento de ajustamento do pacote estatístico GLIM *Generalized Linear Interactive Modelling* para incluir modelos que estão fora da estrutura dos modelos lineares generalizados. O programa GLIM tem sido um padrão de acessibilidade, flexibilidade e mérito educacional para a análise de dados reais.

Um amplo espectro de desenvolvimentos teóricos dos modelos lineares generalizados é discutido, como por exemplo, os modelos aditivos generalizados, os modelos não-exponenciais, os modelos de quase-verossimilhança e os modelos semi-paramétricos. A maior parte da modelagem estatística descrita neste artigo pode ser desenvolvida via o GLIM. Este uso amplo do GLIM sugere meios em que resultados padrão dos modelos lineares generalizados, tais como inferência em grandes amostras, testes de ajustamento, etc., podem ser estendidos para outros modelos. Alguns exemplos de análise de dados reais são apresentados.

ABSTRACT

Generalized linear models are an extension of classical linear models and already seem to encompass quite a broad range of statistical regression models needed for many important practical problems. Over the last ten years there have been many requests to extend the fitting procedure of the GLIM ("Generalized Linear Interactive Modelling") statistical package to include models well outside of the framework of the generalized linear models. The program GLIM has already set a standard of accessibility, flexibility and educational merit for analysing real data.

A wide range of theoretical developments of the generalized linear models are discussed, for example, generalized additive models, non-exponential models, quasi-likelihood models and semi-parametric models. Much statistical modelling described in this paper may be carried out using GLIM. This wider use of GLIM suggests ways in which standard results for the generalized linear models, such as large- sample inference, goodness-of-fit tests, etc., may be extended for other models. Some examples of analysis of real data are provided.

Experimentos com Intercâmbio de Dois Tratamentos e Dois Períodos: Estratégias para Análise e Aspectos Computacionais

Denise A. Botter *

Julio M. Singer *

1. INTRODUÇÃO

Em muitas investigações nas áreas de Agricultura, Educação, Psicologia, Odontologia e Medicina entre outras, há interesse na comparação de $t \geq 2$ tratamentos (fertilizantes, métodos de ensino, drogas, etc.) relativamente à média de alguma variável resposta (produção de algum cereal, nota obtida em um teste, taxa de hemoglobina no sangue, etc.). No caso mais simples, envolvendo a comparação entre $t = 2$ tratamentos, um planejamento experimental bastante utilizado é o planejamento completamente casualizado, onde, de um total de n unidades experimentais disponíveis, n_1 são escolhidas aleatoriamente para receber o tratamento 1, enquanto as $n_2 (= n - n_1)$ restantes recebem o

* Universidade de São Paulo - USP.

tratamento 2. A análise dos dados obtidos através deste tipo de planejamento é bastante simples e pode ser encontrada na maioria dos textos que tratam de planejamento de experimentos, como Cox (1958), Winer (1971) ou Milliken e Johnson (1984). Quando as unidades experimentais utilizadas são animais ou seres humanos, ao planejamento completamente casualizado podem estar associadas as seguintes desvantagens:

- O número de unidades experimentais requerido para se detectarem diferenças com magnitudes de interesse nem sempre está disponível (por exemplo, em ensaios clínicos quando as unidades experimentais são voluntários).
- As unidades experimentais podem não ser suficientemente homogêneas.

Uma alternativa ao planejamento completamente casualizado é o planejamento com intercâmbio (*crossover* ou *changeover*). Nesse caso, n_1 unidades experimentais farão parte do grupo 1 onde recebem, num primeiro período (período 1), o tratamento 1 e, num segundo período (período 2), o tratamento 2; as n_2 unidades experimentais restantes farão parte do grupo 2 onde recebem os tratamentos numa seqüência invertida, isto é, no período 1 recebem o tratamento 2 e no período 2 recebem o tratamento 1. Um exemplo típico na área de Psicobiologia está descrito em Carlini (1988).

Em princípio, o planejamento com intercâmbio pode apresentar as seguintes vantagens relativamente ao planejamento completamente casualizado:

- Permite um maior controle da variabilidade entre unidades experimentais (cada indivíduo serve como seu próprio controle).
- Possibilita a diminuição do número requerido de unidades experimentais.

Para que tais vantagens tenham valor prático, é necessário que os períodos 1 e 2 sejam suficientemente espaçados a fim de que se possa garantir a inexistência do efeito residual do tratamento 1 sobre o tratamento 2 (efeito residual 1) e do efeito residual do tratamento 2 sobre o tratamento 1 (efeito residual 2). Todavia, na maioria das vezes, o espaçamento entre os períodos 1 e 2 (*washout period*) é determinado subjetivamente, gerando dúvida quanto à inexistência dos efeitos residuais.

Grizzle (1965) propôs um teste para verificar se a magnitude do efeito residual 1 é a mesma que a do efeito residual 2 utilizando dados de experimentos com intercâmbio. Se os efeitos residuais 1 e 2 forem diferentes, somente as observações colhidas no período 1 devem ser utilizadas, isto é, a análise estatística é a mesma do planejamento completamente casualizado, porém com perda de informações, pois as observações colhidas no período 2 devem ser desprezadas. Caso haja igualdade entre os efeitos residuais 1 e 2, a análise equivale àquela realizada para planejamentos do tipo *split-plot* (ver Cox (1958), por exemplo). A igualdade entre os efeitos residuais não implica que tais efeitos sejam nulos, mas que, se existirem, têm a mesma intensidade em ambas as seqüências de tratamentos.

Na Seção 2 apresentamos os modelos utilizados na análise de dados obtidos sob planejamento com intercâmbio envolvendo dois tratamentos e dois períodos. As estratégias de análise, incluindo os aspectos computacionais, estão presentes na Seção 3. Na Seção 4 ilustramos a utilização da técnica proposta com análises de dois conjuntos de dados. Finalmente, algumas recomendações quanto à utilização desse tipo de planejamento são consideradas na Seção 5.

2. MODELOS PARA O PLANEJAMENTO COM INTERCÂMBIO

Com base em Grizzle (1965) e Brown (1980), apresentamos dois modelos de análise de dados obtidos sob planejamento com intercâmbio:

- Modelo I - Permite testar a igualdade entre os efeitos residuais e entre os efeitos dos períodos. Permite, também, testar a igualdade entre os efeitos dos tratamentos utilizando somente as observações colhidas no período 1.
- Modelo II - Supõe a inexistência ou a igualdade entre os efeitos residuais e permite testar a igualdade entre os efeitos dos grupos, dos tratamentos e dos períodos, utilizando as observações colhidas nos dois períodos.

Um modelo (Modelo I) que inclui os efeitos residuais é dado por

$$Y_{ijk} = \mu + \xi_{ij} + \pi_k + \phi_{d(i,k)} + \lambda_{d(i,k-1)} + \varepsilon_{ijk} \text{ , com}$$

$$i = 1,2;\ j = 1,\cdots n_i;\ k = 1,2;\ d(i,0) = 0;\ d(1,1) = d(2,2) = 1;$$
$$d = (1,2) = d(2,1) = 2;$$
$$\pi_1 + \pi_2 = \phi_1 + \phi_2 = \lambda_1 + \lambda_2 = 0;\ \lambda_0 = 0,$$

onde

y_{ijk} representa a resposta da j-ésima unidade experimental do i-ésimo grupo no k-ésimo período,

μ representa a média global,

ξ_{ij} representa o efeito aleatório da j-ésima unidade experimental do i-ésimo grupo,

π_k representa o efeito fixo do k-ésimo período,

ϕ_d representa o efeito fixo do d-ésimo tratamento,

λ_d representa o efeito fixo do d-ésimo efeito residual,

ε_{ijk} representa um erro aleatório.

Supomos que

$\xi_{ij} \sim N(0,\sigma_u^2)$; $\varepsilon_{ijk} \sim N(0,\sigma_e^2)$; $i = 1,2$; $j = 1,\cdots,n_i$; $k = 1,2$ e que essas variáveis aleatórias são todas independentes.

A partir deste modelo é fácil observar que a variância de uma observação qualquer é $\sigma^2 = \sigma_u^2 + \sigma_e^2$ e que a covariância e a correlação entre observações colhidas na mesma unidade experimental são, respectivamente, σ_u^2 e $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$.

As hipóteses de interesse, neste caso, podem ser escritas como

$$H_1 : \lambda_1 = \lambda_2, H_2 : \pi_1 = \pi_2, \text{ e } H_3 : \phi_1 = \phi_2,$$

representando, respectivamente, a igualdade de efeitos residuais, de períodos e de tratamentos.

Na Tabela 2.1 apresentamos os estimadores dos contrastes $\lambda_1 - \lambda_2, \pi_1 - \pi_2$ e $\phi_1 - \phi_2$, suas variâncias e seus valores esperados.

Tabela 2.1- Estimadores dos contrastes sob o Modelo I

Contraste	Estimador	Variância do estimador	Valor esperado do estimador
$\lambda_1 - \lambda_2$	$\bar{y}_{1.1} + \bar{y}_{1.2} - \bar{y}_{2.1} - \bar{y}_{2.2}$	$\frac{2n}{n_1 n_2}\sigma^2(1+\rho)$	$\lambda_1 - \lambda_2$
$\pi_1 - \pi_2$	$\frac{1}{2}(\bar{y}_{1.1} + \bar{y}_{2.1} - \bar{y}_{1.2} - \bar{y}_{2.2})$	$\frac{n}{2n_1 n_2}\sigma^2(1-\rho)$	$\pi_1 - \pi_2$
$\phi_1 - \phi_2$	$\bar{y}_{1.1} - \bar{y}_{2.1}$	$\frac{n}{n_1 n_2}\sigma^2$	$\phi_1 - \phi_2$

Observemos que $\bar{y}_{1.1} - \bar{y}_{2.1}$ estima a diferença entre os efeitos de tratamento utilizando somente as observações colhidas no período 1.

Na Tabela 2.2 apresentamos a análise de variância associada ao Modelo I.

Observemos que:

- A igualdade de efeitos residuais pode ser testada contra o Erro 1 que corresponde à variação **entre** unidades experimentais.
- A igualdade de efeitos de tratamentos **não** pode ser testada contra o Erro 2 que corresponde à variação **intra-unidades** experimentais, mas pode ser testada contra o Erro 3 construído por Grizzle (1965). Notemos que a soma de quadrados SQE3 utiliza somente as observações colhidas no período 1 e equivale à soma de quadrados do resíduo associado a um planejamento completamente casualizado.

Tabela 2.2- Análise de variância associada ao Modelo I

Fonte de variação	Soma de quadrados	GL	Valor esperado do quadrado médio
Efeito residual	$SQER = \frac{n_1 n_2}{2n}\left(\bar{y}_{1.1} + \bar{y}_{1.2} - \bar{y}_{2.1} - \bar{y}_{2.2}\right)^2$	1	$\sigma^2(1+\rho) + \frac{2n_1 n_2}{n}(\lambda_1 - \lambda_2)^2$
Erro 1 (entre)	$\mathrm{SQE1} = \frac{1}{2}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij.}^2 - \sum_{i=1}^{2}\frac{y_{i..}^2}{n_i}\right)$	$n-2$	$\sigma^2(1+\rho)$
Período	$SQP = \frac{n_1 n_2}{2n}\left(\bar{y}_{1.1} + \bar{y}_{2.1} - \bar{y}_{1.2} - \bar{y}_{2.2}\right)^2$	1	$\sigma^2(1-\rho) + \frac{2n_1 n_2}{n}(\pi_1 - \pi_2)^2$
Tratamento	$SQTR = \frac{n_1 n_2}{n}\left(\bar{y}_{1.1} - \bar{y}_{2.1}\right)^2$	1	$\sigma^2 + \frac{n_1 n_2}{n}(\phi_1 - \phi_2)^2$
Erro 2 (intra)	$\mathrm{SQE2} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk}^2 - SQE1 - \sum_{i=1}^{2}\sum_{k=1}^{2}\frac{y_{i.k}^2}{n_i}$	$n-2$	$\sigma^2(1-\rho)$
Erro 3	$\mathrm{SQE3} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij1}^2 - \sum_{i=1}^{2}\frac{y_{i.1}^2}{n_i}$	$n-2$	σ^2
Total	$SQT = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk}^2 \frac{y^2\ldots}{2n}$	$2n-1$	

Um modelo (Modelo II) adequado à situação de inexistência ou igualdade entre os efeitos residuais é dado por

$$y_{ijk} = \mu + \gamma_i + \xi_{ij} + \pi_k + \phi_{d(i,k)} + \varepsilon_{ijk},$$

com

$$i = 1,2;\ j = 1,\cdots,n_i;\ k = 1,2;\ d = (1,1) = d(2,2) = 1;\ d = (1,2) = d(2,1) = 2;$$
$$\gamma_1 + \gamma_2 = \pi_1 + \pi_2 = \phi_1 + \phi_2 = 0,$$

onde $\mu, \xi_{ij}, \pi_k, \phi_{d(i,k)}$ e ε_{ijk} têm as mesmas interpretações e obedecem às mesmas suposições mencionadas acima e γ_i representa o efeito fixo do i-ésimo grupo.

As hipóteses de interesse, neste caso, podem ser escritas como

$$H_1 : \gamma_1 = \gamma_2, H_2 : \pi_1 = \pi_2, \text{ e } H_3 : \phi_1 = \phi_2,$$

onde H_1 representa a igualdade de efeitos de grupos e as demais hipóteses têm as mesmas interpretações estabelecidas anteriormente.

Na Tabela 2.3 apresentamos os estimadores dos contrastes $\gamma_1 - \gamma_2 = \pi_1 - \pi_2$ e $\phi_1 - \phi_2$, suas variâncias e seus valores esperados. A análise de variância associada ao Modelo II é apresentada na Tabela 2.4.

Tabela 2.3- Estimadores dos contrastes sob o Modelo II

Contraste	Estimador	Variância do estimador	Valor esperado do estimador
$\gamma_1-\gamma_2$	$\frac{1}{2}(\bar{y}_{1.1}+\bar{y}_{1.2}-\bar{y}_{2.1}-\bar{y}_{2.2})$	$\frac{n}{2n_1n_2}\sigma^2(1+\rho)$	$\gamma_1-\gamma_2$
$\pi_1-\pi_2$	$\frac{1}{2}(\bar{y}_{1.1}+\bar{y}_{2.1}-\bar{y}_{1.2}-\bar{y}_{2.2})$	$\frac{2n}{n_1n_2}\sigma^2(1-\rho)$	$\pi_1-\pi_2$
$\phi_1-\phi_2$	$\frac{1}{2}(\bar{y}_{1.1}+\bar{y}_{2.2}-\bar{y}_{1.2}-\bar{y}_{2.1})$	$\frac{n}{2n_1n_2}\sigma^2$	$\phi_1-\phi_2$

Observemos que:

• O estimador do contraste $\gamma_1-\gamma_2$ é igual à metade do estimador do contraste $\lambda_1-\lambda_2$ obtido sob o Modelo I; além disto, o quadrado médio devido aos efeitos residuais obtido sob o Modelo I é igual ao quadrado médio devido aos grupos. Este fato mostra que os efeitos residuais estão confundidos com os efeitos dos grupos, isto é, testar a igualdade de efeitos residuais sob o Modelo I equivale a testar a igualdade de efeitos dos grupos sob o Modelo II.

• O quadrado médio devido aos períodos sob o Modelo II é igual ao quadrado médio devido aos períodos sob o Modelo I. Isto mostra que testar a igualdade de efeitos dos períodos sob o Modelo II equivale a testar a igualdade de efeitos dos períodos sob o Modelo I.

• O estimador do contraste $\phi_1-\phi_2$ e a soma de quadrados devida aos tratamentos são construídos através das observações colhidas nos períodos 1 e 2. Este resultado já era esperado, pois, sob o Modelo II, os efeitos residuais ou são nulos, ou têm a mesma intensidade.

Tabela 2.4- Análise de variância associada ao Modelo II

Fonte de variação	Soma de quadrados	GL	Valor esperado do quadrado médio
Grupo	$SQG = \frac{n_1 n_2}{2n}\left(\bar{y}_{1.1} + \bar{y}_{1.2} - \bar{y}_{2.1} - \bar{y}_{2.2}\right)^2$	1	$\sigma^2(1+\rho) + \frac{2n_1 n_2}{n}(\gamma_1 - \gamma_2)^2$
Erro 1 (entre)	$\mathrm{SQE1} = \frac{1}{2}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij.}^2 - \sum_{i=1}^{2}\frac{y_{i..}^2}{n_i}\right)$	$n-2$	$\sigma^2(1+\rho)$
Período	$SQP = \frac{n_1 n_2}{2n}\left(\bar{y}_{1.1} + \bar{y}_{2.1} - \bar{y}_{1.2} - \bar{y}_{2.2}\right)^2$	1	$\sigma^2(1-\rho) + \frac{2n_1 n_2}{n}(\pi_1 - \pi_2)^2$
Tratamento	$SQTR = \frac{n_1 n_2}{2n}\left(\bar{y}_{1.1} + \bar{y}_{2.2} - \bar{y}_{1.2} - \bar{y}_{2.1}\right)^2$	1	$\sigma^2(1-\rho) + \frac{2n_1 n_2}{n}(\phi_1 - \phi_2)^2$
Erro 2 (intra)	$\mathrm{SQE2} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk}^2 - SQE1 - \sum_{i=1}^{2}\sum_{k=1}^{2}\frac{y_{i.k}^2}{n_i}$	$n-2$	$\sigma^2(1-\rho)$
Total	$\mathrm{SQT} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk}^2 \frac{y_{...}^2}{2n}$	$2n-1$	

3. ESTRATÉGIAS DE ANÁLISE E ASPECTOS COMPUTACIONAIS

Uma estratégia para a análise de dados obtidos segundo um planejamento com intercâmbio com dois tratamentos e dois períodos consiste em:

- Testar a igualdade dos efeitos residuais através das estatísticas

$$F_{ER} = \frac{SQER}{SQE1/(n-2)} \text{ ou } F_{ER} = \frac{SQG}{SQE1/(n-2)},$$

obtidas das Tabelas 2.2 ou 2.4, respectivamente.

- Se houver evidência de diferença entre os efeitos residuais, testar a igualdade entre os efeitos dos tratamentos através da estatística

$$F_{TR} = \frac{SQTR}{SQE3/(n-2)},$$

apresentada na Tabela 2.2, utilizando somente as observações colhidas no período 1.

- Se houver razões para supor que os efeitos residuais são iguais ou não existem, testar as igualdades entre os efeitos dos grupos e dos tratamentos através, respectivamente, das estatísticas

$$F_G = \frac{SQG}{SQE1/(n-2)} \text{ e } F_{TR} = \frac{SQTR}{SQE2/(n-2)},$$

obtidas da Tabela 2.4.

- O teste da igualdade entre os efeitos de períodos pode ser realizado através da estatística

$$F_P = \frac{SQP}{SQE2/(n-2)},$$

obtida da Tabela 2.2 ou da Tabela 2.4.

Os cálculos correspondentes às análises de variância para dados obtidos sobre planejamentos com intercâmbio envolvendo dois tratamentos e dois períodos podem ser facilmente obtidos com o auxílio de

calculadoras. No entanto, a possibilidade de utilização de pacotes computacionais para esse fim é sempre de interesse, especialmente quando o número de problemas a serem analisados é grande. Nesta seção indicamos algumas alternativas disponíveis.

- Utilização de sub-rotinas específicas para análise de variância (GLM no SAS ou 2V e 4V no BMDP, por exemplo).
  - Considerar um modelo *split-plot* tendo como fatores SEQÜÊNCIA (ou EFEITO RESIDUAL), UNIDADE EXPERIMENTAL (Erro *entre)*, TRATAMENTO e SEQÜÊNCIA * TRATAMENTO (PERÍODO); se SEQÜÊNCIA for não significativo, avaliar a significância de TRATAMENTO (e PERÍODO, se houver interesse); caso contrário:
  - Avaliar a significância de TRATAMENTO através de um modelo completamente casualizado baseado somente nos dados do primeiro período e tendo TRATAMENTO como único fator. Os testes para avaliação do efeito de TRATAMENTO e/ou PERÍODO num modelo que incorpora a igualdade dos *efeitos residuais* são idênticos àqueles considerados no item acima (ver Wallenstein e Fisher (1977) e Gomez-Marin e McHugh (1983)).
- Utilização de sub-rotinas para a análise do Modelo Linear Geral e/ou Regressão (GLM e REG no SAS, 1R no BMDP, MODLIN e REGRESS no NTIA, por exemplo).

Neste caso uma matriz de especificação (planejamento) deve ser construída contendo os valores das seguintes variáveis explicativas (fatores): SEQÜÊNCIA,UNIDADE EXPERIMENTAL (hierárquico em relação à SEQÜÊNCIA), PERÍODO e TRATAMENTO. A Tabela 3.1 apresenta as linhas típicas dessa matriz. Os testes das hipóteses de interesse podem ser obtidos

Tabela 3.1- Linhas típicas da matriz de especificação associada ao modelo que inclui o fator SEQUÊNCIA

Seq.	Per.	Uni. Exp.	ξ_{1j}	ξ_{2j}	γ_1	π_1	ϕ_1
1	1	y_{1j1}	1	0	n_2	1	1
	2	y_{1j2}	1	0	n_2	−1	−1
1	1	$y_{1n_1 1}$	−1	0	n_2	1	1
	2	$y_{1n_1 2}$	−1	0	n_2	−1	−1
2	1	y_{2j1}	0	1	$-n_1$	1	−1
	2	y_{2j2}	0	1	$-n_1$	−1	1
2	1	$y_{2n_2 1}$	0	−1	$-n_1$	1	−1
	2	$y_{2n_2 2}$	0	−1	$-n_1$	−1	1

Nota: $j = 1,\cdots,n_i - 1$ para a i -ésima seqüência, $i = 1,2$.

através da especificação de contrastes nos parâmetros dos modelos; em alguns casos, as sub-rotinas têm comandos apropriados para esse fim (GLM no SAS e MODLIN e REGRESS no NTIA, por exemplo); em outros, os cálculos podem ser realizados através da diferença entre a soma dos quadrados do resíduo associado ao modelo contendo todos os fatores e a soma dos quadrados do resíduo associado ao modelo contendo todos os fatores **exceto** o fator que se deseja testar.

Caso o teste da igualdade de efeitos dos tratamentos deva ser baseado somente nas observações do primeiro período, a matriz de especificação deve conter apenas os valores do fator TRATAMENTO (veja as linhas típicas dessa matriz na Tabela 3.2).

Tabela 3.2 - Linhas típicas da matriz de especificação associada ao modelo para o planejamento completamente casualizado

Seq.	Per.	Unid. Exp.	ϕ_1
1	1	y_{1j1}	n_2
2	1	y_{2j1}	$-n_1$

Nota: $j = 1,\cdots,n_i$ para a i -ésima seqüência, $i = 1,2$.

4. EXEMPLOS

Os dados hipotéticos apresentados na Tabela 4.1 constituem um exemplo típico dos problemas discutidos aqui. Eles são baseados num estudo cuja análise estatística foi realizada no Centro de Estatística da USP (ver Botter et al. (1993)). O objetivo principal é avaliar, através de um planejamento *com intercâmbio envolvendo dois tratamentos e dois* períodos, respostas comportamentais induzidas por uma dose aguda de álcool (tratamento 1) e por um mergulho simulado em câmara hiperbárica (tratamento 2), em mergulhadores sadios. As variáveis resposta são escores de escalas de sintomas físicos e subjetivos associados à sensação de cabeça vazia (X) e sensação de desconforto (Y).

Tabela 4.1- Escores hipotéticos de uma escala de sintomas físicos e subjetivos

			Unid.	Exp.				Unid.	Exp.
Seq.	Per.	Trat.	X	Y	Seq.	Per.	Trat.	X	Y
1	1	1	4	5	2	1	2	3	4
1	2	2	11	0	2	2	1	27	1
1	1	1	13	1	2	1	2	1	10
1	2	2	33	4	2	2	1	19	6
1	1	1	16	3	2	1	2	20	10
1	2	2	28	6	2	2	1	14	5
1	1	1	30	7	2	1	2	11	4
1	2	2	29	8	2	2	1	5	0
1	1	1	19	1	2	1	2	18	2
1	2	2	22	0	2	2	1	2	1
1	1	1	10	6	2	1	2	26	4
1	2	2	31	4	2	2	1	17	9
1	1	1	13	0	2	1	2	13	2
1	2	2	35	8	2	2	1	11	2
1	1	1	11	1	2	1	2	0	16
1	2	2	30	2	2	2	1	7	7
1	1	1	5	4	2	1	2	9	10
1	2	2	33	10	2	2	1	22	6
1	1	1	22	4	2	1	2	8	6
1	2	2	23	4	2	2	1	15	1

Os comandos das sub-rotinas GLM do SAS, 1R e 2V do BMDP para obtenção das análises de variância associadas às variáveis X e Y, apresentadas respectivamente nas Tabelas 4.2 e 4.3, podem ser encontrados no Apêndice.

Os resultados da Tabela 4.2 sugerem que a hipótese de igualdade de efeitos residuais seja rejeitada e que a hipótese de igualdade de efeitos de tratamentos, cujo teste baseou-se somente nas observações colhidas no período 1, não seja rejeitada.

Tabela 4.2- Análise de variância associada à variável X

FV	GL	SQ	QM	F	p
Grupo	1	722,50	722,50	12,83	0,0021
Erro 1 (entre)	18	1013,60	56,31		
Tratamento	1	57,80	57,80	0,86	0,3666
Erro 3	18	1213,00	67,39		

Já, os resultados da Tabela 4.3 sugerem que as hipóteses de igualdade de efeitos residuais e de efeitos de períodos não sejam rejeitadas e que a hipótese de igualdade de efeitos de tratamentos seja rejeitada. Convém ressaltar que o teste de igualdade de efeitos de tratamentos baseou-se nas observações colhidas nos dois períodos.

Tabela 4.3- Análise de variância associada à variável Y

FV	GL	SQ	QM	F	p
Grupo	1	19,60	19,60	1,16	0,2956
Erro 1 (entre)	18	304,00	16,89		
Período	1	6,40	6,40	0,91	0,3539
Tratamento	1	48,40	48,40	6,85	0,0175
Erro 2 (intra)	18	505,60	7,07		

5. COMENTÁRIOS

Grizzle (1965) mostrou que o planejamento com intercâmbio de dois tratamentos e dois períodos é preferível ao completamente casualizado quando podemos supor a inexistência ou a igualdade dos efeitos residuais e quando a correlação entre as respostas aos dois tratamentos é positiva. Brown (1980) analisou o teste de igualdade dos efeitos residuais proposto por Grizzle (1965) concluindo que seu poder não é satisfatório quando amostras de tamanho pequeno ou moderado são utilizadas. Dessa forma,

esse teste não oferece subsídios suficientes para a escolha entre o planejamento com intercâmbio e o completamente casualizado.

No entanto, como este teste é o único disponível na literatura pesquisada, sugerimos que não seja desprezado, mas que seus resultados sejam utilizados com cautela.

Existem muitos pesquisadores que utilizam o planejamento com intercâmbio para seus experimentos e necessitam de um procedimento para a análise de seus dados. A eles dirigimos este trabalho, recomendando que, quando possível, tentem espaçar ao máximo os períodos de aplicação dos tratamentos, a fim de que se possa reduzir a chance de existência de efeitos residuais significativos. Uma extensão da estratégia considerada aqui para o caso três tratamentos e três períodos é apresentada em Botter e Singer (1991). O leitor interessado em mais detalhes sobre a análise de dados obtidos sobre planejamento com intercâmbio deve consultar Jones e Kenward (1989), por exemplo.

APÊNDICE

I. Modelo de regressão que inclui o fator seqüência

O arquivo de dados deve seguir a estrutura apresentada na Tabela 3.1.

a) Sub-rotina 1R do BMDP

```
\INPUT FILE = 'Nome do arquivo de dados'.
VARIABLES = 22.
FORMAT = FREE.
\VARIABLE NAMES = Escore, I11, I12, I13, I14, I15, I16, I17, I18, I19, I21,
                  I22, I23, I24, I25, I26, I27, I28, I29, Per., Trat., Seq.
\REGRESS DEPENDENT = Escore.
         INDEP = I11, I12, I13, I14, I15, I16, I17, I18, I19, I21, I22, I23,
                 I24, I25, I26, I27, I28, I29, Per., Trat., Seq.
\END
```

b) Sub-rotina GLM do SAS

```
DATA Título
INFILE 'Nome do arquivo de dados';
INPUT Escore I11 I12 I13 I14 I15 I16 I17 I18 I19 I21 I22 I23 I24 I25 I26 I27
                I28 I29 Per., Trat., Seq.;
PROC GLM;
MODEL Escore = I11 I12 I13 I14 I15 I16 I17 I18 I19 I21 I22 I23 I24 I25 I26
                  I27 I28 I29 Per., Trat., Seq./SS1;
CONTRAST 'Per.' Per.1;
CONTRAST 'Trat.' Trat .1;
CONTRAST 'Seq.' Seq .1;
CONTRAST 'Unid.Exp.(Seq)'
I1 1 I2 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0 I25 0
I26 0 I27 0 I28 0 I29 0,
I11 0 I12 1 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 1 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 1 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 1 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 1 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 1 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 1 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 1 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
```

I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 1 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 1 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 1 I24 0
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 1
I25 0 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 1 I26 0 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 1 I27 0 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 1 I28 0 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 1 I29 0,
I11 0 I12 0 I13 0 I14 0 I15 0 I16 0 I17 0 I18 0 I19 0 I21 0 I22 0 I23 0 I24 0
I25 0 I26 0 I27 0 I28 0 I29 1;
RUN;

Embora não esteja explícito, o modelo de regressão considerado pelas sub-rotinas acima tem intercepto não nulo.

II. Modelo de análise de variância

a) Modelo que inclui o fator seqüência (Split-plot)

a.1) Sub-rotina 2V do BMDP
\INPUT TITLE IS *'Título'*.
VARIABLES = 3.
FORMAT = FREE.
FILE = *'Nome do arquivo de dados'*.
\VARIABLE NAMES = Seq.,Trat.1,Trat.2.

```
\GROUP CODES(Seq.) = 1,2.
        NAMES(Seq.) = Seq.1,Seq.2.
\DESIGN DEPENDENT = Trat.1 to Trat.2.
        LEVEL = 2.
        NAME = Trat.
        GROUPING = Seq.
\END
```

a.2) Sub-rotina GLM do SAS

```
DATA Título;
INFILE 'Nome do arquivo de dados';
INPUT Escore Unid.Exp. Per Trat. Seq.;
PROC GLM;
CLASS Unid.Exp. Per Trat. Seq.;
MODEL Escore = Seq. Unid.Exp.(Seq.) Per. Trat.;
TEST H=Seq. E=Unid.Exp.(Seq.);
RUN;
```

b) Planejamento completamente casualizado

b.1) Sub-rotina 2V do BMDP

```
\INPUT TITLE IS 'Título'.
        VARIABLES = 2.
        FORMAT = FREE.
        FILE = 'Nome do arquivo de dados'.
\VARIABLE NAMES = Escore,Trat.
\GROUP CODES(Trat.) = 1,2.
        NAMES(Trat.) = Trat.1,Trat.2.
\DESIGN DEPENDENT = Escore.
        GROUPING = Trat.
\END
```

b.2) Sub-rotina GLM do SAS

DATA *Título;*
INFILE *'Nome do arquivo de dados';*
INPUT Escore Trat.;
PROC GLM;
MODEL Escore = Trat.;
RUN;

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Os autores agradecem a um avaliador anônimo pelos valiosos comentários e a Adriana Sañudo pelo auxílio computacional.

REFERÊNCIAS BIBLIOGRÁFICAS

BMDP (1992). BMDP *Statistical Software.* Release 7. Dixon, W.Y. Berkeley: University of California Press.

BOTTER, D.A., SANDOVAL, M.C. & VIANA, P.E. (1993).Comparação entre os efeitos do nitrogênio com os de uma dose aguda de álcool. RAE-CEA 9306. São Paulo: IME-USP.

BOTTER, D.A. & SINGER, J.M. (1991). Experiments with three-treatment three-period crossover design: analysis through the general linear model. *Biometrical Journal* 33, 401-410.

BROWN Jr., B.W. (1980). The crossover experiment for clinical trials. *Biometrics* 36, 69-79.

CARLINI, E.A. (1988). Uma abordagem científica da homeopatia. *Ciência Hoje* 7, 52-59.

COX, D.R. (1958). *Planning of Experiments.* New York: Wiley.

GOMEZ-MARIN, O. & McHUGH, R.B. (1983). Analysis of the unbalanced two-period crossover design with negligible residual effects. *Biometrical Journal 25*, 3-19.

GRIZZLE, J.E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics 21*, 467-480, e Corrigenda em *Biometrics* 30, 727 (1974).

JONES, B. \& KENWARD, M.G. (1989). *Design and Analysis of Cross-Over Trials.* London: Chapman and Hall.

MILLIKEN, G.A. & JOHNSON, D.E. (1984). *Analysis of Messy Data.* Vol.I. California: Lifetime Learning Publications.

NTIA. Software científico desenvolvido pelo Centro Nacional de Pesquisa em Informática para Agropecuária. CNPTIA/EMBRAPA - Caixa Postal 6162 - CEP 13100 - Campinas, SP. Email: ntiasup@cnptia.embrapa.br.

SAS Institute, Inc. (1988). *SAS Language Guide.* Release 6.03 ed. Cary, N.C.: SAS Institute Inc.

WALLENSTEIN, S. & FISHER, A.C. (1977). The analysis of the two-period repeated measurements crossover design with application to clinical trials. *Biometrics* 33, 261-269.

WINER, B.J. (1971). *Statistical Principles in Experimental Design.* New York: McGraw-Hill.

RESUMO

Consideramos dois modelos propostos por Grizzle (1965, Biometrics) associados ao planejamento com intercâmbio de dois tratamentos e dois períodos. Quando os efeitos residuais são diferentes, o teste de igualdade de efeitos dos tratamentos envolve apenas as observações do primeiro período e equivale ao do planejamento completamente casualizado; por outro lado, quando os efeitos residuais são iguais, a Análise de Variância resultante equivale à de experimentos do tipo *split-plot.* Ilustramos esses procedimentos com dois exemplos numéricos e discutimos sua implementação através de sub-rotinas computacionais apropriadas para Análise de Variância e análise do Modelo Linear Geral.

Palavras-chaves: Efeito residual de tratamento; medidas repetidas; modelo linear geral; planejamento com intercâmbio.

ABSTRACT

Two models proposed by Grizzle (1965, Biometrics) for the two-period crossover design are investigated. When the residual effects are different, the test of equality of treatment effects should be based only on the observations taken on the first period. This test is equivalent to that of a completely randomized design. Otherwise, when the residual effects are equal, the analysis corresponds to that of a split-plot design. These procedures are illustrated with two numerical examples and their implementation through computational routines appropriate to the analysis of variance and the analysis of the general linear model is considered.

Key words: crossover design; general linear model; repeated measures; residual effects

Métodos Automáticos de Previsão para Séries Temporais Multivariadas

Enivaldo Carvalho da Rocha*
Basilio de Bragança Pereira**

1. INTRODUÇÃO

A aplicação de séries temporais multivariadas em diversas áreas como Engenharia, Economia e especialmente na Teoria de Controle, onde uma política de controle ótimo do processo pode ser desenvolvida do modelo em forma de Espaço de Estado, tem-se tornado comum principalmente quando se deseja conhecer a dependência entre as várias séries envolvidas. Entretanto, existem poucas técnicas que permitam realizar, sem muita dificuldade, a análise e previsão de séries temporais multivariadas.

Uma metodologia automática para ajustamento e previsão de séries temporais multivariadas foi proposta por KITAGAWA e AKAIKE (1978), a qual é baseada no procedimento Bayesiano e denominada Multivariado Auto-Regressivo Bayesiano - BVAR. O modelo é estimado tomando-se a

* CCEN-UFPE.
** IME e COPPE-UFRJ.

média dos modelos de várias ordens ponderada por sua distribuição *a posteriori*.

Outra metodologia, também proposta por KITAGAWA e AKAIKE (1978), denominada Multivariado Auto-Regressivo Instantâneo - IMAR - consiste de variáveis que se relacionam no mesmo instante t como uma combinação linear das outras mais os termos auto-regressivos.

AKAIKE (1971) também propõe os modelos Auto-Regressivos Vetoriais - VAR -, onde sucessivos modelos são ajustados em ordem crescente e para cada modelo obtém-se o critério de informação de Akaike. Será escolhido, então, o modelo que fornece o menor critério.

Uma metodologia automática para estimação e previsão de modelos de séries temporais univariadas e multivariadas foi proposta por PANDIT e WU (1983), baseada fundamentalmente no ajustamento progressivo de modelos do tipo ARMAV$(n,n-1)$ para $n=1,2,\ldots,$ até que a redução no determinante da matriz da soma do quadrado e produtos cruzados dos resíduos não seja significativa. Através de um critério pode-se testar a significância da redução, criar uma regra de parada para os consecutivos ajustamentos e analisar os possíveis modelos alternativos.

Outra metodologia automática, a qual utiliza a representação em Espaço de Estados e estima o vetor de estado através da análise de correlação canônica, foi desenvolvida por AKAIKE (1976). O conceito de Estado, fundamental no desenvolvimento do método e que pode ser derivado de propriedades de entrada e saída do sistema, é de um vetor contendo o mínimo de informação presente e passada do processo, necessária para prever as respostas do sistema presente e futuro.

Neste artigo faz-se uma descrição destes métodos de modelagem de séries multivariadas. Três séries bivariadas são utilizadas para comparar estas metodologias quanto ao ajustamento dos dados e quanto à capacidade preditiva.

2. MÉTODOS DE MODELAGEM

2.1. Modelo Multivariado Auto-Regressivo Bayesiano (BVAR)

Esta metodologia descrita por KITAGAWA e AKAIKE (1978) considera estimativas Bayesianas para o modelo auto-regressivo, obtidas a partir da média ponderada de estimativas dos modelos de várias ordens. Esta ponderação é feita usando a distribuição de probabilidades *a posteriori* de que o *k*-ésimo modelo tenha sido escolhido.

Da inferência Bayesiana sabemos que a distribuição *a posteriori* é proporcional à verossimilhança vezes *a priori*. Então a probabilidade *a posteriori* de *k*-ésimo modelo será:

$$\prod(k/x) \quad \alpha \quad f(x/k)\cdot\prod_1(k), \tag{2.1}$$

onde $f(x/k)$ é a verossimilhança de um modelo auto-regressivo de ordem k e pode ser obtida a partir da definição do Critério de Informação de Akaike, isto é:

$$f(x/k) = \exp\left\{-\frac{1}{2}AIC(k)\right\}. \tag{2.2}$$

Segundo a abordagem dos Modelos Hierárquicos Bayesianos, a probabilidade *a priori* do *k*-ésimo modelo segue uma distribuição geométrica, isto é:

$$\prod_2(k/\rho) = (1-\rho)\rho^k, \quad k = 0,1,2,\cdots \tag{2.3}$$

e que

$$\prod_2(\rho) = \frac{1}{1-\rho}, \quad 0 \le \rho \le 1. \tag{2.4}$$

Logo, a distribuição *a priori* de k é dada por:

$$\prod_1(k) = \int_\rho \prod_2(k/\rho) \cdot \prod_2(\rho) d\rho = \frac{1}{1+k}. \tag{2.5}$$

Substituindo (2.2) e (2.5) em (2.1), temos:

$$\prod(k/x) \quad \alpha \quad \frac{1}{1+k} \exp\left\{-\frac{1}{2} AIC(k)\right\}, \tag{2.6}$$

Os estimadores dos coeficientes auto-regressivos são obtidos através das equações de Durbin, para k variando do 0 a m, onde m é o limite superior estabelecido para a ordem do modelo:

$$\begin{aligned} a_k(i) &= a_{k-1}(i) - b(k) a_{k-1}(k-1), \quad i = 1,\ldots,k-1 \\ a_k(k) &= b(k) \end{aligned} \tag{2.7}$$

onde $b(k) = \hat{b}(k)\left[\prod(k/x) + \cdots + \prod(m/x)\right]$ e $\hat{b}(k)$ é o estimador de mínimos quadrados do último coeficiente de correlação parcial.

O número de parâmetros deste modelo é $\sum_{j=1}^{m} d(i)^2 + 1$. A versão Bayesiana do critério AIC é:

$$BIC = (N_t - m)\log\hat{\sigma}_B^2 + 2\left(\sum_{j=1}^{m} d(i)^2 + 1\right). \tag{2.8}$$

A extensão multivariada é facilmente obtida substituindo $\hat{b}(k)$ pelos estimadores de máxima verossimilhança condicional da matriz de coeficientes de autocorrelação parcial de ordem mais alta.

O número de parâmetros é

$$\left[\sum_{j=1}^{m} d(i)^2\right] \cdot p + \frac{p(p+1)}{2}, \tag{2.9}$$

onde p é a dimensão do vetor de séries temporais.

2.2. Modelo Multivariado Auto-Regressivo Instantâneo (IMAR)

Esta metodologia descrita por KITAGAWA e AKAIKE (1978) considera o seguinte modelo:

$$X_t = A_0 X_t + A_1 X_{t-1} + \ldots + A_m X_{t-m} + u_t, \tag{2.10}$$

onde A_0 é da forma:

$$A_0 = \begin{bmatrix} 0 & 0 & \ldots\ldots\ldots\ldots & 0 & 0 \\ a_0(2,1) & 0 & \ldots\ldots\ldots\ldots & 0 & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots\ldots & \ldots & \ldots \\ a_0(\ell-1,1) & a_0(\ell-1,2) & \ldots\ldots\ldots\ldots & 0 & 0 \\ a_0(\ell,1) & a_0(\ell,2) & \cdots a_0(\ell,\ell-1) & & 0 \end{bmatrix}$$

denominado modelo de resposta instantânea, no sentido que pode existir uma relação contemporânea entre as variáveis no mesmo instante t, de maneira que o modelo auto-regressivo é dado por:

$$X_t = \sum_{j=1}^{m} (I - A_0)^{-1} A_i X_{t-i} + (I - A_0)^{-1} u_t. \tag{2.11}$$

Desde que as inovações $u_t(1), \cdots, u_t(\ell)$ são não correlacionadas, o ajustamento de modelo multivariado auto-regressivo é obtido aplicando mínimos quadrados via transformação de Householder aos ℓ modelos da forma:

$$X_t(i) = \sum_{j=1}^{\ell} \sum_{m-1}^{m(i,j)} a_m(i,j) X_{t-m}(j) + \varepsilon_t(i), i = 1,2,\ldots,\ell \tag{2.12}$$

onde $a_k(i,j)$ é um elemento da matriz $(I - A_0)^{-1} A_k, k = 1,\ldots,m$.

2.3. Modelo Auto-Regressivo Vetorial (VAR)

Esta metodologia proposta por AKAIKE (1971) considera que toda informação significante passada e presente para prever o futuro esteja contida no conjunto $Y_t, Y_{t-1}, \ldots, Y_{t-k}$, onde k é a ordem do modelo. Então sucessivos modelos auto-regressivos multivariados são ajustados em ordem crescente e para cada modelo obtém-se o AIC - Critério de Informação de Akaike - definido por:

$$AIC = -2\log \text{(verossimilhança)}+2\text{(número de parâmetros)}$$

Para um modelo $VAR(k)$ o critério é dado por:

$$AIC(k) = T\log\left(\left|\hat{\Sigma}_k\right|\right) + 2km^2 \qquad (2.13)$$

onde m é a dimensão do vetor Y_t, k é a ordem do modelo, T é o número de observações e $\left|\hat{\Sigma}_k\right|$ é o determinante da matriz de covariâncias dos resíduos para um modelo Auto-Regressivo Vetorial de ordem k.

Será escolhido então o k que fornece o menor AIC.

2.4. Modelo Auto-Regressivo - Médias Móveis Vetorial (ARMAV)

Este método, proposto por PANDIT e WU (1983), consiste basicamente em ajustar sucessivamente modelos $ARMAV(n, n-1)$ para $n = 1,2,\ldots$, até que uma redução não significativa no determinante da matriz da soma de quadrados e produtos cruzados (SQP) dos resíduos seja obtida.

Este procedimento será melhor entendido a partir das seguintes etapas:

a) Ajustar um modelo $ARMAV(n,n-1)$;

b) Calcular a matriz da soma de quadrados e produtos dos resíduos;

c) Repetir os passos a) e b) até que uma redução do determinante da matriz obtida em b) não seja significativa e o intervalo de confiança para os últimos parâmetros não inclua o valor zero; e

d) Se o modelo é considerado adequado mas tem alguns parâmetros cujos intervalos incluem o zero, reestimar o modelo excluindo estes parâmetros. Testar a adequação do novo modelo, para determinação do modelo final.

A redução do determinante da matriz SQP pode ser julgada através do teste de Wilks, também denominado critério F, que é dado pela razão de dois determinantes, ou seja:

$$\Lambda = \frac{|A_1|}{|A_0|}, \tag{2.14}$$

onde A_0 é a matriz SQP do modelo ajustado sem restrição (modelo com mais parâmetros) e A_1 é a matriz SQP do modelo ajustado com restrição (modelo com menos parâmetros).

O critério é baseado na seguinte estatística:

$$F = \frac{1-\Lambda^{1/\ell}}{\Lambda^{1/\ell}} \cdot \frac{(k\ell - 2\lambda)}{ms} \cdot F(ms, k\ell - 2\lambda), \tag{2.15}$$

onde $k = T - r - \frac{m-s+1}{2}$

$r = m \times$ número de matrizes parâmetros

$s = m \times$ número de matrizes parâmetros restritas a zero

$T =$ número de observações

$m =$ números de séries

$$\ell = \sqrt{\frac{m^2 \cdot s^2 - 4}{m^2 + s^2 - 5}}$$

$$\lambda = \frac{ms - 2}{4}$$

A aplicação deste critério pode ser ilustrada através do seguinte diagrama:

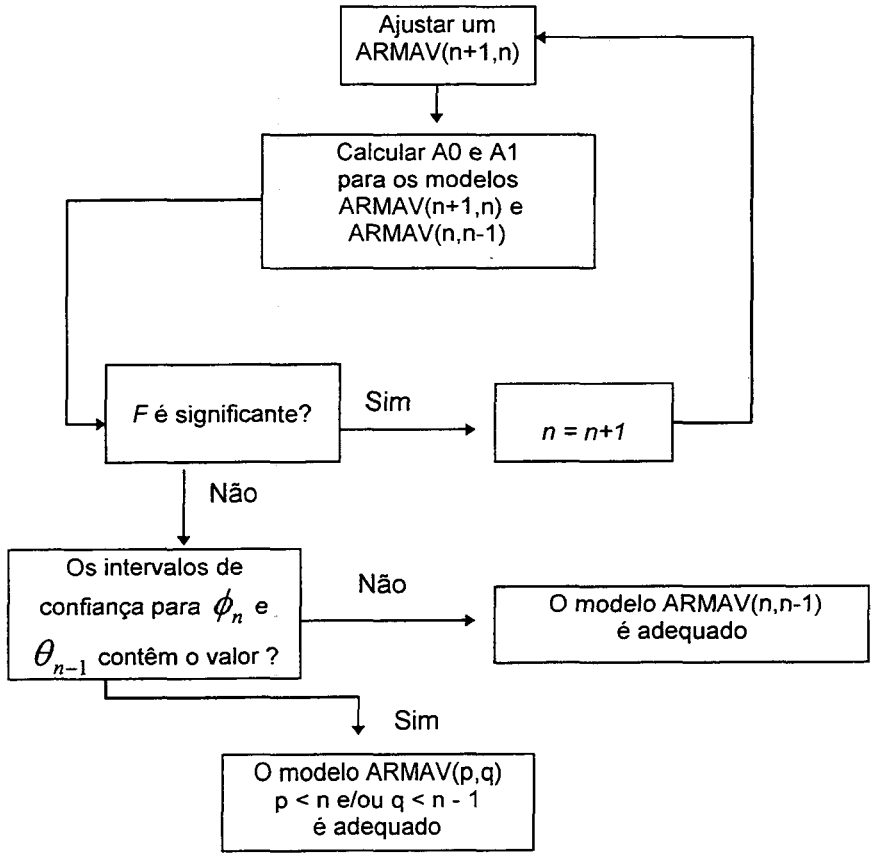


Figura 2.1 - Diagrama do ajustamento de modelos ARMAV (n,n-1).

Fazendo a busca de forma crescente em n temos:

1) Se os parâmetros θ não são zero, isto é, $ARMAV(2n,2n-1), ARMAV(2n+1,2n)$ para $n=1,2,\ldots,$ os modelos pesquisados serão de ordem $(2,1),(3,2),(4,3),\ldots$

2) Se os parâmetros θ podem ser zero, isto é, $ARMAV(n,m),\ n=1,2,\ldots,$ e $m=0,1,\ldots,n-1,$ os modelos pesquisados serão de ordem $(1,0),(2,0),(2,1),(3,0),(3,1),\ldots$

3) Se os parâmetros θ podem somente assumir o valor zero, isto é, $ARMAV(n,0),\ n=1,2,\ldots,$ os modelos pesquisados serão de ordem $(1,0),(2,0),\ldots$

A vantagem da implementação deste procedimento é que, caso seja escolhida a terceira opção, o programa evitará os procedimentos não lineares de estimação, utilizando, desta forma, mínimos quadrados lineares, tornando o algoritmo mais eficiente.

2.5. Modelo Espaço de Estados (STS)

Esta metodologia proposta por AKAIKE (1976) considera a representação de um modelo em espaço de estados dada pelas equações de atualização e observação, respectivamente:

$$Y_{t+1} = Fv_t + G\varepsilon_t$$

$$Y_t = Hv_t \qquad (2.16)$$

onde v_t é o vetor de estado $(p \times 1)$

F é a matriz de transição $(p \times p)$

G é a matriz de entrada $(p \times m)$

ε_t é um vetor de erros $(m \times 1)$

Y_t é o vetor que contém as séries de tempo $(m \times 1)$

H é a matriz de observação $(m \times p)$

Em geral a dimensão de v_t é maior do que a dimensão de Y_t, de maneira que v_t é dado por:

$$v_t = (Y_{t/t}, Y_{t+1/t}, \ldots) \tag{2.17}$$

Com o objetivo de determinar a dimensão do vetor v_t e devido à falta de unicidade da representação acima, AKAIKE (1974) introduziu o conceito de realização mínima, que é obtida através de uma análise de correlações canônicas entre o conjunto de observações presentes e passadas Y_t^P e presentes e futuras Y_t^F, onde

$$\begin{aligned} Y_t^P &= (Y_t, Y_{t-1}, \ldots, Y_{t-\ell})' \\ Y_t^F &= (Y_t, Y_{t+1}, \ldots, Y_{t+\ell})', \ldots \end{aligned} \tag{2.18}$$

2.5.1. Relação entre o Passado e o Futuro

A análise de correlação canônica seleciona uma combinação linear das variáveis dos vetores Y^F e Y^P, de dimensões k_1 e k_2, respectivamente, da forma

$$\varepsilon_i = a_i' Y^F$$

$$\eta_i = b_i' Y^P, \quad i = 1,\ldots,min(k_1,k_2) \tag{2.19}$$

onde $a_i' = (a_{1i},\ldots,a_{k_1 i})$ e $b_i' = (b_{1i},\ldots,k_{k_2 i})$

Calcula-se então os coeficientes de correlações $C_1, C_2,\ldots$ entre ε_1 e η_1, ε_2 e $\eta_2,\ldots$ respectivamente, até que no final temos a seguinte relação:

$$C_1 \leq C_2 \leq \cdots \leq C_{min}(k_1,k_2) \tag{2.20}$$

Suponhamos que a partir de k todos os coeficientes sejam significativamente nulos, $C_i \cong 0$ para $j = k+1,\ldots,\min(k_1,k_2)$, então somente um conjunto $(\varepsilon_1,\eta_1),\ldots,(\varepsilon_k,\eta_k)$ tem correlações positivas de maneira que a relação de dependência entre Y_t^P e Y_t^F estará totalmente contida neste conjunto.

2.5.2. Critério de Akaike para Determinar o Passado

Considere um modelo auto-regressivo de ordem M,

$$Y_t + \phi_1 Y_{t-1} + \ldots + \phi_M Y_{t-M} = \varepsilon_t \tag{2.21}$$

onde ε_t tem distribuição normal multivariada, com média zero e matriz de

covariância Σ. Neste caso, o critério de Akaike é dado por:

$$AIC = T\log\left|\hat{\Sigma}\right| + 2.Mm^2, \tag{2.22}$$

onde T é o número de observações de cada série e $\left|\hat{\Sigma}\right|$ é o determinante da matriz de covariância amostral.

2.5.3. Identificação do Vetor de Estados

O método para ajustar modelos de séries de tempo em espaço de estados pode ser descrito pelo seguinte procedimento:

1) Ajustar um modelo AR(m), através do AIC.

2) Definir o vetor Y_t^P.

3) Fazer uma análise de correlação canônica entre Y_t^P e U_t, onde U_t é um subvetor de Y_t^F, tal que $U_t = (U_t^*, Y_{ut+j})$. Se $U_t^* = (Y_{1t}, \ldots, Y_{mt})$, então os coeficientes de correlação entre Y_t^P e U_t^* serão todos positivos, pois $Y_{1t}, \ldots, Y_{mt}$ são linearmente independentes, logo o procedimento pode ser iniciado a partir da inclusão de $Y_{1(t+1)}$, ou seja:

$$U_t = (Y_{1t}, \ldots, Y_{mt}, Y_{1(t+1)}) = (U_t^*, Y_{(1t+1)}). \tag{2.23}$$

Portanto, os m primeiros elementos de v_t são $Y_{1t}, \ldots, Y_{mt}$.

4) Testar as correlações através do seguinte critério:

$$\text{DIC}(q) = -T\log\prod_{i=q+1}^{k_2}(1-C_i^2) - 2(k_1-q)(k_2-q) \qquad (2.24)$$

onde q é o número de coeficientes de correlação não nulos. Este critério DIC é obtido pela diferença entre o AIC do modelo de correlação canônica quando a última correlação é restrita a zero e o AIC do modelo sem restrição.

Desde que o posto da matriz F é $q < k_2$, o número de parâmetros do modelo restrito será:

$$F(q) = \frac{k_2(k_2+1)}{2} + \frac{k_1(k_1+1)}{2} + q(k_1+k_2-q), \qquad (2.25)$$

$$AIC(q) = T\log\prod_{i=1}^{q}(1-C_i^2) + 2(k_1+k_2-q) \qquad (2.26)$$

e o AIC do modelo sem restrição será:

$$AIC(k_2) = T\log\prod_{i=1}^{k_2}(1-C_i^2) + 2F(k_2), \qquad (2.27)$$

onde

$$F(k_2) = \frac{k_2(k_2+1)}{2} + \frac{k_1(k_1+1)}{2} + k_2 + k_1, \qquad (2.28)$$

logo

$$DIC(q) = -T\log\prod_{i=K_1+1}^{k_2}(1-C_i^2) - 2(k_1-q)(k_2-q). \qquad (2.29)$$

5) Se $DIC(j-1)<0$, então o menor coeficiente de correlação é significativamente nulo, Y_{ut+j} é linearmente dependente dos seus antecessores e o *j*-ésimo ciclo termina. Se ainda existirem elementos a serem testados, retornar à etapa 4 e testar a inclusão do próximo elemento $Y_{ut+(j+1)}$, caso contrário ir para etapa 7.

6) Se $DIC(j-1)>0$, então Y_{ut+j} é linearmente independente dos seus antecessores, faz-se $U_j = Y_{ut+j}$, e se o ciclo não terminou inclui-se $Y_{(u+1)t+j}$ e vai-se para etapa 4. Caso contrário inclui-se $Y_{ut+(j+1)}$ e vai-se para etapa 4, começando um novo ciclo. Finalmente, se todos elementos de Y_t^F foram incluídos, ir para etapa 7.

7) Determinar as matrizes F e G.

Então, terminando o procedimento, v_t será constituído pelos p primeiros elementos linearmente independentes de Y_t^F, ou seja:

$$v_t = \begin{bmatrix} Y_t' \\ Y_{t+1/t}' \\ \vdots \\ Y_{t+p-1/t}' \end{bmatrix} = \begin{bmatrix} Y_{1t} & Y_{2t} & \cdots & Y_{mt} \\ Y_{1t+1/t} & Y_{2t+1/t} & \cdots & Y_{mt+1/t} \\ \vdots & \vdots & & \\ Y_{1t+p_1-1/t} & Y_{2t+p_2-1/t} & \cdots & Y_{mt+p_m-1/t} \end{bmatrix} \quad (2.30)$$

tal que $\sum_{i-1}^{m} p_i = p$.

Um algoritmo mais recente, apresentado no Manual do Forecast Master (ver GOODRICH e STELLWAGEM, 1995), consiste basicamente na mesma formulação proposta por Akaike, diferindo apenas na estrutura do vetor de estado.

O vetor do passado é o mesmo, ou seja, contém o passado e o presente, porém o vetor do futuro não inclui o presente Y_t, isto é:

$$v_t = \begin{bmatrix} Y'_{t+1/t} \\ Y'_{t+2/t} \\ \vdots \\ Y'_{t+p/t} \end{bmatrix} \tag{2.31}$$

2.5.4. Determinação das Matrizes F e G

Suponhamos que a correlação canônica de *j*-ésima variável é nula, isto significa que Y_{t+j} é linearmente dependente e, portanto, pode ser escrita como uma combinação linear dos seus antecessores de maneira que os coeficientes da equação:

$$Y_{ut+j} = -\frac{a^1_{1t}}{a^1_{ut}} Y_{1t} - \cdots - \frac{a^h_{(u-1)t}}{a^h_{ut}} Y_{(u-1)t+j} \tag{2.32}$$

formarão a *u*-ésima linha da matriz F.

A matriz G será obtida do modelo ajustado na etapa 2 do algoritmo, dado por:

$$\Phi(B)Y_t = \varepsilon_t \tag{2.33}$$

Considerando a representação médias móveis temos:

$$Y_t = \psi(B)\varepsilon_t \tag{2.34}$$

Igualando as duas equações temos $\Phi^{-1}(B) = \psi(B)$, tal que:

$$\psi_0 = I$$
$$\psi_1 = \Phi_1$$
$$\psi_2 = \Phi_1\psi_1 + \Phi_2$$
.....
$$\psi_M = \Phi_1\psi_{M-1} + \Phi_2\psi_{M-2} + \cdots + \Phi_{M-1}\psi_1 + \Phi_M$$

e finalmente

$$G = \begin{bmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_M \end{bmatrix} \qquad (2.36)$$

Exemplo: JONES, 1982.

Jones considerou 144 observações de três variáveis $Y_t = (Y_{1t}, Y_{2t}, Y_{3t})$ obtidas da simulação de um ARMA(2,1) com

$$\Phi_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0{,}44 \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} 0{,}75 & 0 & 0 \\ 0 & 0{,}6 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\theta_1 = \begin{bmatrix} -0{,}75 & 0 & 0 \\ 0 & -0{,}6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ e } \quad \underset{\sim}{\varepsilon}_t \sim N(\underset{\sim}{0}, I).$$

Pelo critério AIC selecionou-se AR(3), de maneira que $Y_t^P = (Y_t, Y_{t-1}, Y_{t-2}, Y_{t-3})$

t	Y_t^P Y_t'	Y_{t-1}'	Y_{t-2}'	Y_{t-3}'
1	$Y_{11}\ Y_{21}\ Y_{31}$			
2	$Y_{21}\ Y_{22}\ Y_{32}$	$Y_{11}\ Y_{21}\ Y_{31}$		
3	$Y_{13}\ Y_{23}\ Y_{33}$	$Y_{12}\ Y_{22}\ Y_{32}$	$Y_{11}\ Y_{21}\ Y_{31}$	
4				$Y_{11}\ Y_{21}\ Y_{31}$
⋮	.....	.....	.....	.....
144	$Y_{1144}\ Y_{2144}\ Y_{3144}$	$Y_{1143}\ Y_{2143}\ Y_{3143}$	$Y_{1142}\ Y_{2142}\ Y_{3142}$	$Y_{1141}\ Y_{2141}\ Y_{3141}$

Y_t^F	Ciclo	i	C_i	DIC(q)
$Y_{1t}, Y_{2t}, Y_{3t} e Y_{1(t+1)}$	4	1	1.0	
		2	1.0	
	+	3	1.0	
+	+	4	0.45	14.4
Y_t, Y_{1t+1}, Y_{2t+1}	5	4	0.48	31.0
		5	0.43	14.0
$Y_t, Y_{1t+1}, Y_{2t+1}, Y_{3t+1}$	6	4	0.55	53.3
		5	0.44	23.5
		6	0.39	10.4
$Y_t, Y_{t+1}, Y_{1(t+2)}$	7	4	0.58	58.5
		5	0.46	22.9
		6	0.43	9.0
		7	0.22	-4.6*
$Y_t, Y_{t+1}, Y_{2(t+2)}$	8	4	0.60	58.5
		5	0.47	18.4
		6	0.40	2.2
		7	0.19	-6.8**
$Y_t, Y_{t+1}, Y_{3(t+2)}$	9	4	0.55	52.6
		5	0.48	23.8
		6	0.39	5.2
		7	0.24	-3.2***

Com base na análise acima:

$$v_t = (Y_t, Y_{t+1})' = (Y_{1t}, Y_{2t}, Y_{3t}, Y_{1t+1}, Y_{2t+1}, Y_{3t+1})'$$

$$F = \left[\begin{array}{ccc|ccc} & 0 & & & I & \\ 0.25 & 0.25 & 0.10 & -0.02 & -0.45 & -0.32 \\ -0.17 & -0.13 & -0.19 & 0.84 & 0.18 & 0.47 \\ 0.61 & 0.15 & 0.11 & -0.18 & 0.47 & 0.33 \end{array}\right]$$

$$G = \left[\begin{array}{c} I \\ \hline \cdots\cdots\cdots\cdots\cdots\cdots \\ \begin{array}{ccc} 0.65 & -0.02 & -0.12 \\ 0.50 & 0.44 & -0.03 \\ -0.07 & -0.03 & 0.39 \end{array} \end{array}\right]$$

2.5.5. Previsão

Um procedimento para previsão usando o filtro de Kalman consiste em:

1) Ajustar um modelo de espaço de estados usando análise de correlação canônica para determinar a dimensão do vetor de estado e estimar as matrizes F, G e Σ_ε.

2) Depois de determinar as matrizes F, G e Σ_ε, os preditores são calculados recursivamente usando o filtro de Kalman.

As previsões h passos a frente são facilmente geradas por

$$\hat{Y}_{t+h/t} = H\hat{v}_{t+h/t},$$

onde

$$\hat{v}_{t+h/t} = F^h v_{t/t}, \quad h = 1,2,\cdots \tag{2.37}$$

A matriz de covariância dos erros de previsões é dada por:

$$P^v_{t+h/t} = \text{cov}(Y_{t+h} - \hat{Y}_{t+h/t}) = H\text{cov}(v_{t+h} - \hat{v}_{t+h/t})H = HP^v_{t+h/t}H' \tag{2.38}$$

onde $P^v_{t+h/t}$ é a matriz de covariância do erro de predição, que é calculada recursivamente a partir de

$$\begin{aligned} P^v_{t+1/t} &= FP^v_{t/t}F' + G\sum_{\varepsilon} G' \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ P^v_{t+h/t} &= FP^v_{h-1/t}F' + G\sum_{\varepsilon} G', \end{aligned} \tag{2.39}$$

onde $P^v_{t/t}$ é a matriz de covariância dos erros de estimação.

Os elementos de $v_{t/t}$ e $P_{t/t}$ são calculados recursivamente através do filtro de Kalman, dado pelas seguintes equações:

$$\hat{v}_{t+1/t+1} = \hat{v}_{t+1/t} + K_{t+1}(Y_{t+1} - H\hat{v}_{t+1/t}), \tag{2.40}$$

e K_{t+1} é denominada matriz ganho de Kalman, dada por:

$$K_{t+1} = P^v_{t+1/t}H'(HP^{\upsilon}_{t+1/t}H' + R_{\delta})^{-1} \tag{2.41}$$

onde R_{δ} é a matriz de covariâncias dos distúrbios estocásticos associados à equação de observação:

$$Y_{t+1} = Hv_t + \delta_t \tag{2.42}$$

que, em geral, é identicamente nula.

Exemplo: Suponhamos que estamos interessados em calcular $\hat{v}_{1/1}$ e $P^{v}_{1/1}$, então:

$$\hat{v}_{1/1} = \hat{v}_{1/0} + k_1(Y_1 - H\hat{v}_{1/0}) \tag{2.43}$$

onde $\hat{v}_{1/0} = F v_0$

$$k_1 = P^{v}_{1/0} H'(HP^{v}_{1/0} H' + R_\delta)^{-1},$$

$$P^{v}_{1/0} = FP_0 F' + G\Sigma_\varepsilon G$$

finalmente

$$P^{v}_{1/1} = P^{v}_{1/0} - K_1 HP^{v}_{1/0} \tag{2.44}$$

Os valores iniciais v_0 e P_0 devem ser conhecidos a priori e as matrizes F, G e Σ_ε são determinadas a partir do modelo ajustado.

2.5.6. Equivalência com os Modelos de Box-Jenkins

Considere um modelo ARMAV(p,q)

$$\Phi(B) Y_t = \theta(B)\varepsilon_t \tag{2.45}$$

ou em sua representação de médias móveis infinita:

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \tag{2.46}$$

com $\psi_0 = I$.

A equação de previsão será dada por:

$$E(Y_{t+k}) = Y_{t+k/t} = \sum_{j=k}^{\infty} \psi_j \varepsilon_{t+k-j}, \tag{2.47}$$

$$Y_{t+k/t+1} = \sum_{j=k}^{\infty} \psi_{l-1} \varepsilon_{t+1+k-j} = \psi_{k-1} \varepsilon_{t+1} +$$
$$+ \sum_{j=k+1}^{\infty} \psi_{j-1}\ \varepsilon_{t+k-(j-1)} = Y_{t+k/t} + \psi_{k-1}\ \varepsilon_{t+1}, \tag{2.48}$$

se $k = \max(p; q+1)$; então,

$$Y_{t+k/t} = \sum_{j=1}^{t} \phi_j Y_{t+k-j/t} \tag{2.49}$$

Suponhamos, sem perda de generalidade, que $p > q$:

$$\begin{aligned} Y_{t+1/t+1} &= Y_{t+1/t} + \psi_0\, \varepsilon_{t+1} \\ Y_{t+2/t+1} &= Y_{t+2/t} + \psi_1\, \varepsilon_{t+1} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_{t+p/t+1} &= \phi_p Y_{t/t} + \phi_{p-1} Y_{t+1/t} + \cdots + \phi_1 Y_{t+p-1/t} + \psi_{p-1} \varepsilon_{t+1} \end{aligned} \tag{2.50}$$

ou na forma matricial:

$$\begin{bmatrix} Y_{t+1/t+1} \\ Y_{t+2/t+1} \\ \vdots \\ Y_{t+p/t+1} \end{bmatrix} = \begin{bmatrix} 0 & \\ 0 & I \\ & \vdots \\ \phi_p \cdots & \phi_1 \end{bmatrix} \begin{bmatrix} Y_{t/t} \\ Y_{t+1/t} \\ \vdots \\ Y_{t+p-1/t} \end{bmatrix} + \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_{p-1} \end{bmatrix} \varepsilon_{t+1} \tag{2.51}$$

que é a representação em espaços de estados.

Consideremos o modelo em espaço de estados de ordem p:

$$\begin{bmatrix} v_{1t} \\ v_{2t} \\ \vdots \\ v_{pt} \end{bmatrix} = \begin{bmatrix} 0 & \\ 0 & I \\ & \vdots \\ \phi_p \cdots & \phi_1 \end{bmatrix} \begin{bmatrix} v_{1t-1} \\ v_{2t-1} \\ \vdots \\ v_{pt-1} \end{bmatrix} + \begin{bmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{p-1} \end{bmatrix} \varepsilon_t \tag{2.52}$$

$$
\begin{aligned}
v_{1t} &= v_{1t-1} + \psi_0 \varepsilon_t \\
v_{2t} &= v_{2t-1} + \psi_1 \varepsilon_t \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
v_{pt} &= \sum_{j=0}^{p-1} \phi_{p-j} v_{j+1(t-1)} + \psi_{p-1} \varepsilon_t .
\end{aligned}
\tag{2.53}
$$

Fazendo sucessivas substituições na primeira equação temos:

$$
v_{1t} = \sum_{k=1}^{p} \phi_k v_{1t-k} - \sum_{k=1}^{p-1} \phi_k \varepsilon_{t-k} + \psi_0 \varepsilon_t . \tag{2.54}
$$

Fazendo $v_{1t} - Y_t$ e $\psi_0 = I$, temos:

$$
Y = \sum_{k=1}^{p} \phi_k \, Y_{t-k} - \sum_{k=1}^{p-1} \phi_k \varepsilon_{t-k} + \varepsilon_t \tag{2.55}
$$

que é a representação tradicional de um modelo ARMAV$(p, p-1)$.

Se v_t não contém o presente, então para um modelo em forma de espaços de estados de ordem p temos:

$$
\begin{bmatrix} v_{1t} \\ v_{2t} \\ \vdots \\ v_{pt} \end{bmatrix} = \begin{bmatrix} Y_{t+1/t} \\ Y_{t+2/t} \\ \vdots \\ Y_{t+1-p/t} \end{bmatrix}. \tag{2.56}
$$

Substituindo v_{1t} por $Y_{t+1/t}$ na equação acima

$$
Y_{t+1/t} = \sum_{k=1}^{p} \phi_k Y_{t+1-k/t} - \sum_{k=1}^{p-1} \phi_k \varepsilon_{t-k} + \varepsilon_t , \tag{2.57}
$$

$$Y_t = \sum_{k=1}^{p} \phi_k Y_{t-k/t} - \sum_{k=1}^{p-1} \phi_k \varepsilon_{t-1-k} + \varepsilon_{t-1} = \\ = \sum_{k=1}^{p} \phi_k Y_{t-1-k/t} - \sum_{k=1}^{p-1} \phi_{k-1} \varepsilon_{t-k}, \quad (2.58)$$

onde $\phi_0 = I$, que é um modelo ARMAV(p,p).

3. CRITÉRIOS DE AJUSTAMENTO E PREVISÃO

Desde o trabalho pioneiro de AKAIKE (1969) vários critérios para seleção de modelos foram desenvolvidos, reduzindo significativamente a necessidade do julgamento do analista e ao mesmo tempo podendo ser automatizados facilmente em computador. Maiores referências (ver PEREIRA, 1984).

A idéia geral é que todos os critérios de qualidade do ajustamento melhoram mais ou menos regularmente quando aumenta-se o número de parâmetros do modelo. Entretanto, aumentar demasiadamente o número de parâmetros é indesejável pois a precisão das estimativas diminui e o modelo torna-se de pouca utilidade. Para resolver esses dois objetivos, em princípio conflitantes, os métodos de escolha procuram penalizar os critérios de ajustamento pelo número de parâmetros incluídos no modelo.

Apresentamos nesta seção os principais critérios de ajustamento, e de erros de previsão, como o erro percentual médio absoluto (MAPE) e o erro médio quadrático (RMSE).

A escolha do melhor modelo consiste em ajustar consecutivos modelos, em ordem crescente, e selecionar o modelo que fornecer o menor valor do critério. Este critério é uma estatística que consiste de um termo da penalidade e outro da soma de quadrado dos resíduos, onde a penalidade do número de observações é uma função crescente do número de

parâmetros do modelo. O problema da escolha de um bom critério é então escolher uma estatística que forneça a melhor forma para a penalidade.

ENGLE e BROWN (1986) fazem uma descrição dos critérios apresentados aqui e analisam empiricamente os erros de previsão para os modelos selecionados por diferentes critérios tais como aqueles a seguir.

O critério de Informação de Akaike (AIC) é um procedimento que consiste em escolher o modelo que minimiza

-2 log(verossimilhança)+2(número de parâmetros).

Para um modelo auto-regressivo de ordem k o AIC assume a seguinte forma:

$$\text{AIC}(k) = \log(\hat{\sigma}^2) + \frac{2k}{T} \tag{3.1}$$

onde k é o número de parâmetros do modelo e $\hat{\sigma}^2$ é o estimador de máxima verossimilhança da variância residual. No caso multivariado substitui-se a variância residual pelo determinante da matriz de covariâncias dos resíduos $\left|\hat{\Sigma}\right|$ e o termo da penalidade é proporcional à dimensão ao quadrado do vetor Y_t, ou seja, a versão multivariada do AIC é:

$$\text{AIC}(k) = \log(\left|\hat{\Sigma}\right|) + \frac{2km^2}{T}. \tag{3.2}$$

Como m é fixo, a relação entre os dois termos permanece a mesma em função de k e T, respectivamente.

O AIC pode ser escrito em sua forma produto, dada por:

$$\text{AIC} = \hat{\sigma}^2 \exp\{2k / T\}. \tag{3.3}$$

O critério do erro de predição final (FPE) foi também desenvolvido por Akaike para selecionar a ordem de modelos auto-regressivos. O FPE é definido como sendo o erro de predição do modelo ajustado. O procedimento consiste em ajustar auto-regressões de ordem $k = 0,1,\cdots, L$ a uma série Y_t usando mínimos quadrados. Então a estatística:

$$\text{FPE}(k) = \left(1 + \frac{k+1}{T}\right) \frac{T}{(T-1-k)} \frac{\hat{\sigma}^2}{T} \tag{3.4}$$

é calculada para cada ordem. A ordem com o menor valor de $\text{FPE}(k)$ é então selecionada como a melhor. Em termos do problema de determinar o número correto de variáveis a serem incluídas no modelo (auto-regressão stepwise), o FPE pode simplesmente ser reescrito na forma produto, dada por:

$$\text{FPE} = \hat{\sigma}^2 (T+k) / (T-k). \tag{3.5}$$

O AIC é obtido como um caso particular do critério de informação generalizado, dado por:

$$\text{GIC} = \log(\hat{\sigma}^2) + \frac{Ck}{T}. \tag{3.6}$$

Um critério consistente, denominado critério de Schwarz (ver SCHWARZ, 1978), é obtido fazendo-se $C = \log T$ e substituindo na equação acima, temos:

$$\text{SCHWARZ} = \log(\hat{\sigma}^2) + \frac{\log(T)k}{T} \tag{3.7}$$

ou

$$\text{SCHWARZ} = \hat{\sigma}^2 T^{K/T}. \tag{3.8}$$

Substituindo C por uma função decrescente de T, tal que C/T converge para zero quando T vai para infinito, obtém-se critérios consistentes. Tomando o limite inferior da taxa de crescimento de C, como

sendo $\log\ \log(T)$, ENGLE e BROWN (1986), descrevem o critério de Hannan-Quinn (ver HANNAN, 1980) dado por:

$$HQ = \log(\hat{\sigma}^2) + \frac{\log\log(T)k}{T} \tag{3.9}$$

ou

$$HQ = \hat{\sigma}^2 (\log T)^{K/T}. \tag{3.10}$$

O critério de Shibata (ver SHIBATA, 1980) é definido por:

$$\text{SHIBATA} = \hat{\sigma}^2\left(1 + \frac{2k}{T}\right). \tag{3.11}$$

O critério de validação generalizada é dado por:

$$\text{GCV} = \hat{\sigma}^2\left(1 - \frac{k}{T}\right)^{-2} \tag{3.12}$$

e o RICE (ver RICE, 1984) é:

$$\text{RICE} = \hat{\sigma}^2\left(1 - \frac{2k}{T}\right) \tag{3.13}$$

Existe muita semelhança entre os critérios AIC, FPE, GCV, RICE e SHIBATA para pequenos valores de K/T. Basicamente, eles diferem somente em relação ao termo da penalidade. Expandindo em séries de Taylor o termo da penalidade dos critérios tem-se

$$\text{AIC} = 1 + 2\frac{k}{T} + 2\left(\frac{k}{T}\right)^2 + \frac{8}{3}\left(\frac{k}{T}\right)^3 + \cdots \tag{3.14}$$

$$\text{FPE} = 1 + 2\frac{k}{T} + 2\left(\frac{k}{T}\right)^2 + 4\left(\frac{k}{T}\right)^3 + \cdots \tag{3.15}$$

$$\text{SCHWARZ} = 1 + \frac{(\log T)}{T} k + \cdots \tag{3.16}$$

$$\text{HQ} = 1 + \frac{2(\log \log T)}{T} k + \cdots \tag{3.17}$$

$$\text{SHIBATA} = 1 + 2\frac{k}{T} \tag{3.18}$$

$$\text{GCV} = 1 + 2\frac{k}{T} + 3\left(\frac{k}{T}\right)^2 + \cdots \tag{3.19}$$

$$\text{RICE} = 1 + 2\frac{k}{T} + 6\left(\frac{k}{T}\right)^2 + \cdots \tag{3.20}$$

Os dois primeiros termos dos critérios AIC, FPE, SHIBATA, GCV e RICE são todos iguais a $1 + 2\frac{k}{T}$. A diferença entre eles é somente nos termos de ordem mais elevada. Os critérios Hannan-Quinn e Schwarz têm penalidade muito alta comparado com os outros. A ordem dos critérios, de acordo com a penalidade, em ordem crescente. é: SHIBATA, AIC, FPE, GCV e RICE. No entanto, quando k se aproxima de $T/2$, o critério de RICE passa a ter uma penalidade superior a do critério de Hannan-Quinn.

Através de simulação diversos estudos têm verificado as propriedades destes critérios. Por ser um critério consistente, neste trabalho a escolha dos modelos será feita através do critério de Schwarz (também conhecido como BIC - Bayesian Information Criterion).

Para avaliar o erro de previsão é usado o erro percentual médio absoluto (MAPE) e a raiz quadrada do erro médio quadrático, definidos por:

$$\text{MAPE}(h) = \left(\frac{1}{NP}\sum_{t=1}^{NP} |e_t(h)| / Y_{t+h}\right) \times 100 \tag{3.21}$$

$$\text{RMSE}(h) = \left(\frac{1}{NP} \sum_{t=1}^{NP} e_t^2(h) \right)^{1/2} \qquad (3.22)$$

onde NP é o número de períodos previsto e $e_t(h)$ é o erro de previsão h passos à frente:

$$e_t(h) = Y_{t+h} - \hat{Y}_t(h)$$

4. ASPECTOS COMPUTACIONAIS

Os programas para ajustamento e previsão dos modelos BVAR, IMAR, VAR e STS foram adaptados para microcomputador a partir dos programas originais TIMSAC-78, escritos em linguagem FORTRAN e publicados por AKAIKE, KITAGAWA, ARAHATA e TADA (1979) onde são indicadas as referências sobre os métodos e algoritmos. Para o modelo ARMAV foi implementado o teste F utilizando o SAS no cálculo dos estimadores de máxima verossimilhança.

Os programas são os seguintes:

MULCOR - Calcula a função de covariância cruzada multivariada.

CANOCA - Análise de correlação canônica de um vetor de séries temporais.

PRDCTR - Calcula as previsões para um modelo ARMAV.

MULMAR - Caso multivariado do método de mínimo AIC para ajustar modelos auto-regressivos.

MULBAR - Caso multivariado do método Bayeisiano para ajustar modelos auto-regressivos.

5. ANÁLISE DO AJUSTAMENTO E PREVISÃO

O objetivo desta seção é descrever os resultados do ajustamento e previsão em três séries bivariadas.

Vários critérios de ajustamento são utilizados, embora a análise possa se restringir apenas ao critério de Schwarz, pois os demais critérios fornecem resultados semelhantes. As previsões são avaliadas através dos erros médios quadráticos.

As previsões serão calculadas quatro períodos à frente com a origem variando de t à $t+19$.

	$t+1$	$t+2$	$t+3$	$t+4$	...	$t+19$	$t+20$	$t+21$	$t+22$	$t+23$
t	x	x	x	x						
$t+1$		x	x	x	x					
⋮					⋮					
$t+18$						x	x	x	x	
$t+19$							x	x	x	x

Se a origem é $t+18$, calcula-se as previsões para os períodos $t+19$ à $t+22$, em seguida mudamos a origem para $t+19$ e obtém-se os próximos quatro períodos até $t+23$. Então, se observamos o bloco diagonal do diagrama acima, teremos dois valores para o cálculo dos erros de previsão considerando os períodos $t+18$ e $t+19$. Logo, a medida que a origem se desloca para esquerda, mais elementos teremos para a avaliação do MAPE e RMSE; porém, se a série for pequena, é interessante

saber quantos valores podem ser retirados da série de maneira a não comprometer o processo de estimação.

5.1. A Série Mink e Muskrat

A série de dados Mink e Muskrat refere-se à quantidade de vendas anuais das peles destes animais pela Hundson's Bay Company para o período de 1767 a 1911.

BULMER (1974) através da análise espectral das duas séries para o período de 1848 a 1911 encontrou um período de 19 anos para ambas, estimando que o ciclo dos muskrats é cerca de 2,3 anos adiante do ciclo dos minks, pois a população dinâmica de ambos é afetada devido ao fato do mink ser o maior predador do muskrat.

A maioria dos autores, tais como: JENKINS (1975), CHAN e WALLIS (1978), JENKINS e ALAVI (1981), COOPER e WOOD (1982), LIM e TONG (1983) e TERASVIRTA (1985) tem analisado a série para o período compreendido entre 1848 e 1911 desprezando o período anterior, pois houve uma mudança brusca no processo por volta de 1840, (ver Figura 5.1a). A Figura 5.1b apresenta o logaritmo das séries, analizado nos artigos mencionados e neste trabalho.

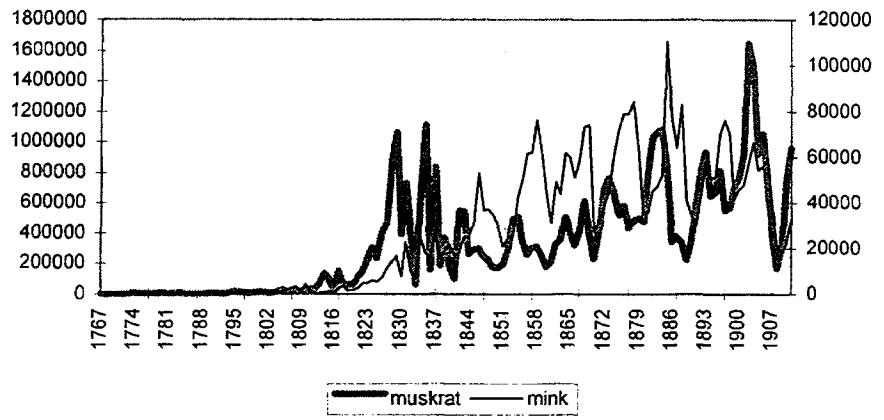


Figura 5.1a - *Séries dos Minks e Muskrat (1767-1911)*

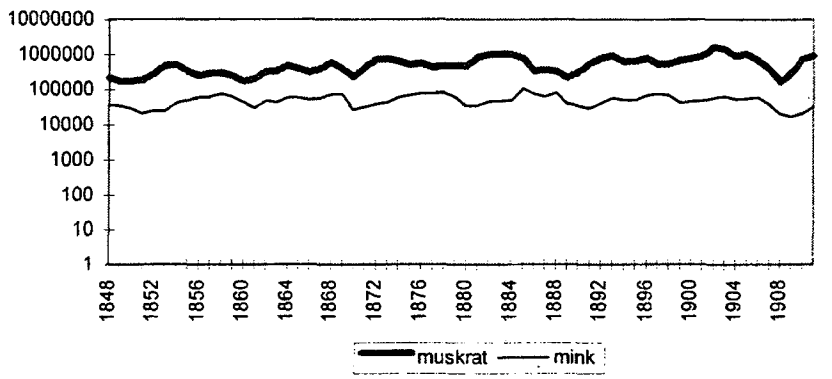


Figura 5.1b - *Logaritmo das Séries (1848-1911)*

Como foi visto anteriormente, todos os modelos descritos neste trabalho são automáticos, uma vez que determinam a ordem do modelo sem intervenção do analista. Mesmo utilizando critérios que penalizam o acréscimo de parâmetros no modelo, estas metodologias tendem a selecionar modelos de ordem elevadas, especialmente os Auto-Regressivos BVAR e IMAR. Os modelos ARMAV e STS (Espaços de Estados) fornecem modelos mais parcimoniosos. As Tabelas 5.1a,b

apresentam os modelos ajustados a série Mink e Muskrat e os diferentes critérios de ajustamento utilizados.

Tabela 5.1a - Critérios de ajustamento para os modelos - séries Mink e Muskrat

MODELO	CRITÉRIOS NA FORMA PRODUTO						
	AIC	FPE	SCHWARZ	HQ	SHIBATA	GCV	RICE
BVAR (3)	.462	.465	.788	.162	.241	.486	.522
IMAR (6)	.181	.182	.278	.783E-2	.108	.187	.195
VAR (2)	.170	.171	.206	.118	.137	.171	.172
ARMAV (2,1)	.661E-2	.661E-2	.735E-2	.536E-2	.582E-2	.662E-2	.663E-2
STS (4)	.253E-2	.254E-2	.349E-2	.135E-2	.173E-2	.258E-2	.263E-2

Tabela 5.1b - Critérios de ajustamento para os modelos - séries Mink e Muskrat

MODELO	CRITÉRIOS NA FORMA NORMAL	
	AIC	SCHWARZ
BVAR (3)	-74	-23
IMAR (6)	-164	-123
VAR (2)	-170	-152
ARMAV (2,1)	-481	-472
STS (4)	-574	-543

Analisando as Tabelas 5.1a,b observa-se que o critério de Schwarz é o mais rigoroso, pois penaliza fortemente os modelos de ordem mais elevada. Todos os critérios foram consistentes e selecionaram o modelo STS como sendo o que melhor se ajustou à série bivariada Mink e Muskrat. As duas últimas colunas da Tabela 5.1b contêm os valores dos critérios AIC e Schwarz em suas formas normais. Segundo o critério de Schwarz a seleção do melhor modelo obedeceu à seguinte ordem: STS, ARMAV, VAR, IMAR e BVAR.

Os critérios para analisar as previsões foram o erro percentual médio absoluto (MAPE) e a raiz quadrada do erro médio quadrático (RMSE). A Tabela 5.2 apresenta os erros de previsão 1,2,3 e 4 passos à frente para a série Mink, considerando os modelos ajustados anteriormente.

Tabela 5.2 - Critérios de previsão para os modelos - série Mink

	MAPE				RMSE			
MODELO	1	2	3	4	1	2	3	4
BVAR	34,307	66,106	33,246	44,305	21982,14	35429,55	14857,95	19513,16
IMAR	14,486	24,910	35,937	33,507	8870,118	10695,65	13178,70	12420,00
VAR	17,784	147,52	280,73	222,86	10267,06	57590,25	98539,77	76679,64
ARMAV	23,137	42,877	53,880	54,449	12038,02	16958,76	18478,94	18413,61
STS	21,860	24,637	26,178	39,058	15227,10	14824,50	10650,50	13624,90

Tanto o MAPE como o RMSE um passo à frente indicam o modelo IMAR como sendo o melhor para prever a série Mink, seguido do VAR, STS, ARMAV e BVAR. Observa-se que a ordem de seleção dos modelos quanto ao ajustamento é diferente da classificação com relação aos erros de previsão.

A Tabela 5.3 apresenta os erros de previsão MAPE e RMSE 1, 2, 3 e 4 passos à frente para a série Muskrat.

Tabela 5.3 - Critérios de previsão para os modelos - série Muskrat

	MAPE				RMSE			
MODELO	1	2	3	4	1	2	3	4
BVAR	59,773	85,192	94,676	97,851	628519,3	823836,5	893482,9	929642,1
IMAR	62,878	130,84	213,50	302,59	425164,8	746847,9	114127,0	158644,6
VAR	408,196	2256,2	5823,6	158,78	285350,9	13340020	32398630	86572540
ARMAV	44,992	73,555	113,88	148,80	345722,5	585231,4	677464,9	914518,0
STS	43,759	109,08	175,07	200,06	300523,0	557001,0	837546,0	999843,0

Os erros de previsão um passo à frente para séries Muskrat indicam o modelo de Espaço de Estados (STS) como o melhor para fazer previsões, o que difere da escolha para a variável Mink.

5.2 A Série Lydia E. Pinkham

A série Lydia E. Pinkham representa as vendas e os gastos com propaganda de um medicamento durante o período de 1907 a 1960. O composto vegetal da Lydia E. Pinkham é um extrato erval em solução alcoólica considerado eficaz contra dores menstruais e indisposição da menopausa. A história deste produto ganhou grande publicidade em várias ocasiões devido à controvérsia em torno de seus componentes, obrigando o fabricante a fazer algumas mudanças nos folhetos de propaganda, ocasionando quatro períodos distintos de comportamento: nos períodos de 1907 a 1944 e de 1926 a 1940 foi considerado um remédio de efeitos gerais, de 1915 a 1925 e de 1941 a 1980, um remédio para cólicas menstruais.

Diversos autores têm analisado esta série com o principal objetivo de estudar os efeitos das vendas com futuros gastos com propaganda. Modelos econométricos e de séries temporais univariado, multivariado e de função de transferência têm sido considerado por diversos autores (ver B.B. PEREIRA, E.C. ROCHA, J.G.C. ROCHA e S. DRUCK, 1987a).

A série, em alguns casos, têm sido diferenciada uma vez, o que foi suficiente para torná-la estacionária, muito embora um comportamento cíclico de período aproximadamente de 13 anos tenha sido detectado para os dados anuais.

Os gráficos das séries anual e mensal são apresentados nas Figuras 5.2 e 5.3.

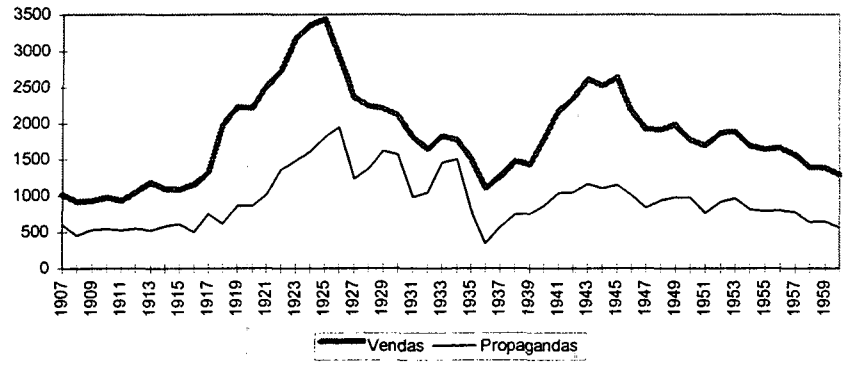


Figura 5.2 - *Série de vendas e propaganda anual da Lydia E. Pinkham (1907-1960)*

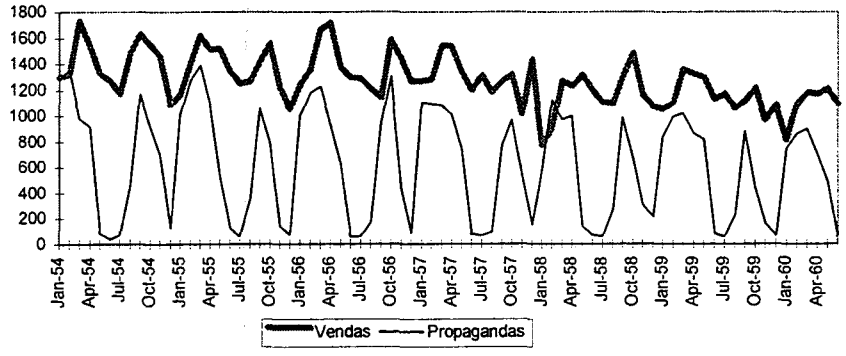


Figura 5.3 - *Série de vendas e propaganda mensal da Lydia E. Pinkham (jan/1954-jun/1960)*

5.2.1. Ajustamento e Previsão para a Série Lydia E. Pinkham Anual

As Tabelas 5.4a,b apresentam os modelos ajustados à série Lydia E. Pinkham anual e diferentes critérios de ajustamento utilizados.

Tabela 5.4a - Critérios de ajustamento para os modelos - série Lydia E. Pinkham anual

MODELO	CRITÉRIOS NA FORMA PRODUTO						
	AIC	FPE	SCHWARZ	HQ	SHIBATA	GCV	RICE
BVAR (10)	.180E+10	.200E+10	.602E+10	.142E+9	.314E+9	.217E+10	.251E+10
IMAR (2)	.129E+10	.129E+10	.141E+10	.107E+10	.115E+10	.129E+10	.129E+10
VAR (2)	.140E+10	.140E+10	.153E+10	.115E+10	.124E+10	.140E+10	.140E+10
ARMAV (2,1)	.404E+10	.404E+10	.499E+10	.259E+10	.307E+10	.407E+10	.411E+10
STS (3)	.243E+10	.243E+10	.274E+10	.188E+10	.208E+10	.243E+10	.244E+10

Tabela 5.4b - Critérios de ajustamento para os modelos - série Lydia E. Pinkham anual

MODELO	CRITÉRIOS NA FORMA NORMAL	
	AIC	SCHWARZ
BVAR (10)	1662	1756
IMAR (2)	1636	1643
VAR (2)	1643	1649
ARMAV (2,1)	1725	1742
STS (3)	1685	1695

Analisando as Tabelas 5.4a,b observa-se que o critério de Schwarz seleciona o modelo IMAR como o melhor quanto ao ajustamento, seguido pelo VAR, STS, ARMAV e BVAR. Observa-se ainda que o modelo auto-regressivo BVAR não foi parcimonioso, o que significa sua última classificação.

A Tabela 5.5 apresenta os erros de previsão MAPE e RMSE para a série Sales, considerando os modelos ajustados anteriormente.

Tabela 5.5 - Critérios de previsão para os modelos - série Sales

MODELO	MAPE				RMSE			
	1	2	3	4	1	2	3	4
BVAR	7,074	13,454	11,673	13,285	158,373	253,455	236,455	248,597
IMAR	6,081	12,080	11,604	10,600	135,495	230,306	217,781	210,520
VAR	6,296	12,930	12,776	10,888	137,026	243,292	235,175	217,584
ARMAV	5,866	11,390	10,574	11,013	130,775	219,630	205,201	209,155
STS	6,052	12,085	11,624	11,422	133,775	230,897	221,459	221,894

Os erros de previsão um passo à frente para a série Sales indicam o modelo ARMAV como o melhor quanto à previsão.

A Tabela 5.6 apresenta os erros de previsão para a série Advertising.

Tabela 5.6 - Critérios de previsão para os modelos - série Advertising

MODELO	MAPE				RMSE			
	1	2	3	4	1	2	3	4
BVAR	7,138	13,225	13,703	15,506	69,147	137,698	141,554	146,735
IMAR	9,410	16,262	15,568	14,991	98,550	149,430	140,340	135,826
VAR	11,297	19,464	17,646	15,401	116,124	182,961	160,496	148,608
ARMAV	11,743	15,182	14,957	18,499	118,694	156,361	133,046	155,087
STS	10,231	15,382	14,184	16,594	98,577	140,317	127,761	137,925

Os erros de previsão um passo à frente para a série Advertising indicam o modelo BVAR como o melhor quanto à previsão.

5.2.2. Ajustamento, Previsão e Simulação para a Série Lydia E. Pinkham Mensal

As Tabelas 5.7a,b apresentam os modelos ajustados à série Lydia E. Pinkham mensal e os diferentes critérios de ajustamento utilizados.

Tabela 5.7a - Critérios de ajustamento para os modelos - Lydia E. Pinkham mensal

MODELO	CRITÉRIOS NA FORMA PRODUTO						
	AIC	FPE	SCHWARZ	HQ	SHIBATA	GCV	RICE
BVAR (10)	.428E+10	.469E+10	.139E+11	.380E+9	.827E+9	.615E+10	-.662E+11
IMAR (8)	.489E+10	.492E+10	.805E+10	.175E+10	.256E+10	.514E+10	.552E+10
VAR (2)	.208E+10	.208E+10	.234E+10	.164E+10	.180E+10	.209E+10	.209E+10
ARMAV (3,2)	.190E+11	.190E+11	.234E+11	.125E+11	.147E+11	.192E+11	.193E+11
STS (3)	.175E+10	.175E+10	.208E+10	.121E+10	.140E+10	.176E+10	.177E+10

Tabela 5.7b - Critérios de ajustamento para os modelos - Lydia E. Pinkham mensal

MODELO	CRITÉRIOS NA FORMA NORMAL	
	AIC	SCHWARZ
BVAR (10)	1818	1915
IMAR (8)	1829	1870
VAR (2)	1759	1769
ARMAV (3,2)	1941	1958
STS (3)	1745	1759

Analisando as Tabelas 5.7a,b observa-se que o critério de Schwarz seleciona o modelo STS como o melhor quanto ao ajustamento, seguido pelos modelos VAR, IMAR, BVAR e ARMAV.

A Tabela 5.8 apresenta os erros de previsão MAPE e RMSE para a série Sales, considerando os modelos ajustados anteriormente.

Tabela 5.8 - Critérios de previsão para os modelos - série Sales

	MAPE				RMSE			
MODELO	1	2	3	4	1	2	3	4
BVAR	16,486	15,702	15,928	16,814	233,411	201,554	214,284	208, 529
IMAR	18,985	17,371	20,277	20,903	265,263	250,310	247,390	265,254
VAR	13,367	15,785	16,259	16,503	195,144	202,617	231,025	211,799
ARMAV	16,061	16,742	23,022	35,808	235,701	241,534	309,117	406,365
STS	11,981	17,307	16,793	13,490	179,276	242,833	225,671	178,853

Os erros de previsão um passo à frente para a série Sales indicam o modelo STS como o melhor modelo quanto à previsão.

A Tabela 5.9 apresenta os erros de previsão para série Advertising.

Tabela 5.9 - Critérios de previsão para os modelos - série Advertising

	MAPE				RMSE			
MODELO	1	2	3	4	1	2	3	4
BVAR	129,570	46,299	74,900	104,666	282,702	289,258	247,157	229,224
IMAR	179,600	74,649	67,355	143,916	398,833	285,008	360,967	255,613
VAR	161,289	41,612	84,267	110,696	312,974	299,749	264,578	263,240
ARMAV	330,717	101,725	205,327	219,978	543, 821	557,944	629,815	462,536
STS	133,777	51,744	74,036	104,877	312,411	280,465	267,029	284,410

Os erros de previsão um passo à frente para série Advertising indicam o modelo BVAR como o melhor quanto à previsão.

CONCLUSÃO

Os critérios para selecionar o melhor modelo quanto ao ajustamento apresentam-se bastante semelhantes, com exceção do critério de Schwarz, que tende a penalizar fortemente modelos mais complexos. No

entanto, o critério RICE se aproxima bastante do Schwarz quando o número de parâmetros é aproximadamente a metade do número de observações. Este fato foi verificado em séries de tempo pequenas, como a série Mink e Lydia E. Pinkham.

A escolha do melhor modelo de previsão é uma tarefa árdua devido à falta de consistência entre os erros de previsão de uma série para outra, pois um erro pode ser menor em relação a uma delas e maior em relação à outra. Uma solução para este problema é um modelo de previsão que combine todas as situações. Para isto, deve ser introduzida uma métrica que leve em consideração a estrutura de correlação entre as séries.

Os critérios para selecionarmos o melhor modelo de previsão são consistentes, e em geral, nos levam a selecionar o mesmo modelo. No entanto, o MAPE tem uma interpretação mais simples do que o RMSE.

Finalmente, a comparação de métodos de identificação de modelos ARMA pode ser vista em DE GOOIJER, J.H., ABRAHAM, B., GOULD, A e ROBINSON, L. (1985); KOREISHA, S. e YOSHIMOTO, G. (1991) e PAPARODITIS, E. (1993).

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. (1969). Fitting autorregressive, models for prediction. Annals Institute Statistical Mathematics, 21, pp. 243-247.

AKAIKE, H. (1971). Autorregressive model-fitting for control. Annals Institute Statistical Mathematics, 23, 163-180.

AKAIKE, H. e NAKAGAWA, T. (1972). Statistical analysis and control of dynamic systems. Saiensu-Sha, Tokyo.

AKAIKE, H. (1974). Stochastic theory of minimal realization. IEEE Transaction on Automatic Control, 19, pp. 667-674.

AKAIKE, H., ARAHATA, E. e OZAKI, T. (1975). TIMSAC-74, A time series analysis and control program package (1). Computer Science Monographs, no. 5, The Institute of Statistical Mathematics, Tokyo.

AKAIKE, H. (1976). Canonical correlation analysis and the use of an information criterion. Advances and case studies in system indentification. Ed. R. Mehra and D. G. La Iniotis, pp. 27-96, Academic Press.

AKAIKE, H., ARAHATA, E. e OZAKI, T. (1976). TIMSAC-74, A time series analysis and control program package (2). Computer Science Monograph, no. 6, The Institute of Statistical Mathematics, Tokyo.

AKAIKE, H., KITAGAWA, G., ARAHATA, E. e TADA, E. (1979). TIMSAC-78, Computer Science Monograph, no. 11, The Institute of Statistical Mathematics, Tokyo.

BOX, G. E. P. e JENKINS, G. H. (1970). Time series analysis. Forecasting and Control, Holden Day, San Francisco.

BULMER, M. G. (1974). A statistical analysis of the 10 year cycle in Canada. J. Animal Ecology, 43, pp. 701-718.

CHAN, W. Y. e WALLIS, K. F. - (1978). Multiple time Series Wodelling: another look at The mink - muskrat interacction. Applied Statistics, 27, no. 2 pp 168-175.

COOPER, D. M. e WOOD, E. F. (1982). Identifying multivariate time series models. J. Time Series Analysis, 3, pp. 153-164.

GOOIJER, J.G. ABRAHAM, B., GOULD, A. e ROBINSON, L. (1985). Methods for determining the order of an autoregressive-moving average process: a survey. International Statistical Review, 53, 3, pp. 301-329.

ENGLE, R. F. e BROWN, S. J. (1986). Model selection for forecasting. Applied Mathematics and Computation, 20, pp. 313-327.

HANNAN, E. (1980). The estimation of the order of an ARMA process. Annals of Statistics, 8, pp. 1071-1081.

GOODRICH, R.L. e SLELLWAGEN, E.A. (1985). Forecast Master Multivariate Time Series Forecasting. (Scientific Systems, Cambridge, MA).

JENKINS, G. H. (1975). The interaction between the muskrat and mink cycles in North Canada. In Proceedings of the 8th International Biometric Conference, Edited by LCA Corten and T. Postelnicil, Ed. A.R.S. Romania Bukarest.

JENKINS, G. H. e ALAVI, A. S. (1981). Some aspects of modelling and forecasting multivariate time series, J. Time Series Analysis, 2, pp. 1-48.

JONES, K. J. (1982). A comparison of state transition and vector ARMA modeling procedure. Joint Statistical Meeting, Cincinnat, Ohio.

KITAGAWA, G. e AKAIKE, H. (1978). A procedure for the modeling of non-stationary time series. Annals Institute Statistical Mathematics, 30, pp. 351-363.

KOREISHA, S. e YOSHIMOTO, G. (1991). A comparison of among identification procedures for autoregressive moving average models. International Statistical Review, 59, pp. 37-57.

LIM, K. S. e TONG, H. (1983). A statistical approach to difference - delay equation modelling in ecology - Two case Studies. J. Time Series Analysis, 4, pp. 239-267.

OSBORN, D. R. (1977). Exact and approximate maximum likehood estimators for vector moving average processes. J. Royal Statistical Society, s.B., 39, pp. 114-118.

PANDIT, S. M. e WU, S. W. (1983). Time series and system analysis with applications. Wiley.

PAPARODITIS, E. (1993). A comparison of some autocovariance-based methods of ARMA model selection: a simulation study. Journal Statistical Computation Simulation, 45, pp. 97-120.

PEREIRA, B. B. (1984). Séries temporais multivariadas. Curso do 6º Simpósio Nacional de probabilidade e Estatística, Rio de janeiro, RJ.

PEREIRA, B. B., E. c. ROCHA, J.G. C. ROCHA e S. DRUCK (1987a). Lydiametrics revisited. II Escola de Séries Temporais e Econometria, ENCE, RJ.

PEREIRA, B. B., E. c. ROCHA (1987b). Uma re-análise das previsões de vendas e propaganda da Lydia Pinkham. II Escola de Séries Temporais e Econometria, ENCE, RJ.

RICE, J.A. (1984). Bandwidth choice for nonparametric kernel regression. Annals of Statistics., 12, pp. 1215-1230.

SCHWARZ, C. (1978). Estimating the dimension of a model. Annals of Statistics, 6, pp. 461-464.

SHIBATA, R. (1980). Asymptotically efficient selection of process. Annals of Statistics, 8, pp. 147-164.

TERASVIRTA, T. (1985). Mink and muskrat interaction: a structural analysis. J. of Time Series Analysis, 6, no, 3, pp. 171-180.

RESUMO

Neste artigo cinco procedimentos automáticos de previsão para séries temporais multivariadas são apresentados. A comparação dos métodos é feita através de critérios de escolha de modelos e comparando a precisão das previsões através de medidas de erros de previsão. Três séries bivariadas conhecidas são utilizadas na comparação.

ABSTRACT

In this paper five automatic procedures for forecasting multivariate time series are presented. The methods are compared through model choices criterias and forecasting erros measures. Three well known bivariate time series data are used in the comparison.

POLÍTICA EDITORIAL

A RBEs objetiva promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais através da apresentação, descrição e discussão desses métodos e de suas aplicações, num formato de fácil assimilação pelos membros da comunidade científica. Destina-se também a servir de veículo para troca de idéias entre os especialistas e todos os interessados em análise e desenvolvimento de metodologia estatística.

A RBEs tem periodicidade semestral e publica artigos teóricos e/ou aplicados de métodos estatísticos, com ênfase na análise de fenômenos econômicos e sociais. São também aceitos artigos abordando os diversos aspectos do desenvolvimento metodológico relevantes para órgãos produtores de estatísticas, assim como artigos de revisão do estado da arte em temas específicos:

a) delineamento de pesquisas;

b) avaliação de pesquisas e mensuração de erros;

c) uso e combinação de fontes alternativas de informações;

d) novos desenvolvimentos em metodologia de pesquisa;

e) análise de séries de tempo;

f) estudos demográficos;

g) integração de dados;

h) amostragem e estimação;

i) análises de dados;

j) crítica e imputação de dados;

l) disseminação e confiabilidade de dados; e

m) modelos econométricos.

Todos os artigos submetidos serão avaliados pelo Comitê Editorial da RBEs quanto a sua qualidade e relevância, devendo os mesmos serem inéditos. Além disto, não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional.

A RBEs publicará também resenhas de livros, artigos escritos a convites e ensaios sobre o ensino de Estatística.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos remetidos para publicação deverão ser submetidos em 3 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 à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

- Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

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) autores, indicando-se para cada um a filiação e endereço permanente para correspondência. Agradecimentos a colaboradores, outras 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 parágrado, sem fórmulas, com no máximo 150 palavras.

3.O artigo deve ser divido em seções numeradas progressivamente, com títulos concisos e apropriados. Subseções, todas, inclusive a primeira, 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 apresentados em folhas separadas, 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 traçados em papel branco, com 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 nos títulos.

7. Serão preferidos originais processados por editores de texto, tais como: Microsoft Word, Word Perfect e LATEX.

Se o assunto é Brasil, procure o IBGE

O IBGE põe à disposição da sociedade milhares de informações de natureza estatística (demográfica, social e econômica), geográfica, cartográfica, geodésica e ambiental, que permitem conhecer a realidade física, humana, social e econômica do País.

ATENDIMENTO TELEFÔNICO

Ligação Direta Gratuita: 0800-218181

INTERNET

http://www.ibge.gov.br
http://www.ibge.org

PONTOS DE ATENDIMENTO

Rio de Janeiro

Centro de Documentação e Disseminação de Informações - CDDI
Rua General Canabarro, 706 - 20271-201 - Maracanã
Fax: (021)284-1109

Livraria do IBGE
Avenida Franklin Roosevelt, 146 - loja - 20021-120 - Castelo
Tel.: (021)220-9147
Avenida Beira Mar, 436 - 2° andar - 20201- 060 - Castelo
Tel.: (021)210-1250 Ramais: 41 / 420 / 422 / 425 e 427
Fax: (021)240-0012

Norte

RO - Porto Velho - Rua Tenreiro Aranhã, 2643 - Centro - 78900-750
Telefax: (069)221-3658

AC - Rio Branco - Rua Benjamin Constant, 506 - Centro - 69900-160
Tels.: (068)224-1540/1490 - Ramal 6; Fax: (068)224-1382

AM - Manaus - Rua Afonso Pena, 38 - Centro - 69020-160
Telefax: (092)232-1372 PABX: (092) 633-2433 Ramais 48 e 49

RR - Boa Vista - Av. Getúlio Vargas, 76-E - Centro - 69301-031
Tel.: (095)224-4103 - Ramal 22 Telefax: (095)623-9399

PA - Belém - Av. Gentil Bittencourt, 418 - Batista Campos
66035-340 - Tel.: (091)242-0234; Fax: (091)241-1440

AP - Macapá - R. Leopoldo Machado, 2466 - Bairro Central
68908-120 - Telefax: (096)223-2696

TO - Palmas - ACSE 01 - Conjunto 03 - Lote 6/8 - Centro
77100-040 - Tel.: (063)215-1907 - Ramal 308; Fax: (063)215-1829

Nordeste

MA - São Luís - Av. Silva Maia, 131 - Praça Deodoro - 65020-570
Tel.: (098)221-5121; Fax: (098)232-3226

PI - Teresina - Rua Simplício Mendes, 436 - Centro - 64000-110
Tel.: (086)221-4161; Fax: (086)221-6308

CE - Fortaleza - Av. 13 de Maio, 2901 - Benfica - 60040-531
Tel.: (085)243-6941 Fax: (085)281-3353

RN - Natal - Av. Prudente de Moraes, 161 - Petrópolis - 59020-400
Tel.: (084)211-5310 - Ramal 13 Fax: (084)221-3025

PB - João Pessoa - Rua Irineu Pinto, 94 - Centro - 68010-100
Tel.: (083)241-1560 - Ramal 219 e 220 Fax: (083)241-7255

PE - Recife - Rua do Hospício,387 - 4° andar - Boa Vista - 50050-050
Tel.: (081)231-0811 - Ramal 215; Telefax: (081)423-0056 / 423-0355
Ramais 215 e 224

AL - Maceió - Praça dos Palmares, s/n° - Edifício do INAMPS 3° e 4°
and 57020-000 - Tel.: (082)221-2385 221-1531; Fax: (082)326-1754

SE - Aracajú - Rua Riachuelo, 1017 - Térreo - São José - 49015-160
Telefax: (079)222-3122 / 8197 / 8198

BA - Salvador - Av. Estados Unidos, 476 - 4° andar - Comércio
Edifício Sesquicentenário - 40013-900 - Tel.: (071)243-9277 - Ramais
2005 e 2008; Telefax: (071)241-2502

Sudeste

MG - Belo Horizonte - Rua Oliveira, 523 - 1° andar - Cruzeiro
30310-150 - Tel.: (031)223-0554 - Ramais 1112 e 1113
Telefax: (031)223-3381

ES - Vitória - Avenida dos Navegantes, 675 - 9° andar - Enseada do
Suá - 29056-900 - Tel: (027) 324-4016; Fax: (027) 325-3857

SP - São Paulo - Rua Urussuí, 93 - 3° andar - Itaim Bibi - 04542-050
Tels.: (011)822-2106 / 0077 - Ramal 281; Fax: (011)822-5264

Sul

PR - Curitiba - Alameda Dr. Carlos de Carvalho, 625 - Térreo - Centro
80430-180 - Tel.: (041) 322-5500 - Ramais 253 e 254;
Telefax: (041)222-5764

SC - Florianópolis - Rua Victor Meirelles, 170 - Centro - 88010-440
PABX: (048)224-0733 - Ramais 155, 144 e 140
Telefax: (048)222-0369

RS - Porto Alegre - Avenida Augusto de Carvalho, 1205 - Térreo
Praia de Belas - 90010-390 - Tel.: (051)228-6444 - Ramais 211, 213
e 225; Fax: (051)228-8507; Telefax: (051)228-6444 - Ramal 212

Centro-Oeste

MS - Campo Grande - Rua Barão do Rio Branco, 1431 - Centro
79002-174 - Tels.: (067)721-1163/1902/1525 - Ramais 32 e 42;
Fax: (067)721-1520

MT - Cuiabá - Avenida Tenente Coronel Duarte, 407 - 1°/ 2° andares
Centro - 78005-750 - Tels: (065)623-7121 / 7255
Fax: (065)623-0573

GO - Goiânia - Avenida Tocantins, 675 - Setor Central - 74015-010
Tel.: (062)223-3121; Telefax: (062)223-3106

DF - Brasília - SDS - Ed. Venâncio II - Bl H - Quadra 06 / 1° andar
70393-900 - Tels.: (061)223-1359 / 321-7702 - Ramal 124;
Fax: (061)226-9106

O IBGE possui, ainda, agências localizadas nos principais municípios.

IBGE
Instituto Brasileiro de Geografia e Estatística

Se o assunto é Brasil,
procure o IBGE

http://www.ibge.gov.br
http://www.ibge.org

atendimento
0800 21 81 81