MINISTÉRIO DA ECONOMIA, FAZENDA E PLANEJAMENTO
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034 - 7175

R. bras. Estat., Rio de Janeiro, v.51, n.195, p. 1 -117 , jan./jun. 1990.

REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE

Publicação semestral, editada pelo IBGE, 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 3 vias (que não serão devolvidas) para:

Djalma G. C. Pessoa
Editor Responsável - RBEs

ENCE
Rua André Cavalcante, 106 - Bairro de Fátima
CEP 20 231 - 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.

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

Capa
Pedro Paulo Machado

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística. - v.1, n.1 (jan./mar. 1940)- . - Rio de Janeiro: IBGE, 1940-
v.
Trimestral (1940-1986), semestral (1987-)
Órgão oficial do IBGE.
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989)

ISSN 0034-7175 - Revista brasileira de estatística

1. Estatística - Periódicos. I. IBGE.

IBGE. CDDI. Dep. de Documentação e Biblioteca
RJ-IBGE/88-05 rev.

CDU 31(05)

Impresso no Brasil/Printed in Brazil

SUMÁRIO

ARTIGOS

PRODUÇÃO DE ESTATÍSTICA E SISTEMA ESTATÍSTICO

Eduardo Augusto Guimarães*

I

A expressão do sistema estatístico nacional é usualmente empregada em duas acepções, relacionadas mas distintas.

A primeira acepção refere-se ao conjunto de informações estatísticas relativas à realidade econômica e social do País, estruturadas segundo regras e critérios próprios. A segunda, de natureza institucional, refere-se ao conjunto de entidades produtoras de estatísticas e está preocupada com a coordenação e a articulação entre essas entidades, de forma a garantir uma atuação eficiente. Para distinguir entre essas duas acepções, recorrer-se-á nesse texto a iniciais maiúsculas (Sistema Estatístico Nacional) para indicar a segunda acepção.

É certamente possível, em princípio, conceber um Sistema Estatístico Nacional nessa segunda acepção, vale dizer, é certamente possível formular o desenho de uma estrutura institucional que organize e articule as atividades das entidades produtoras de estatística. As dúvidas que podem se manifestar aqui referem-se, por exemplo, ao melhor formato desse sistema e ao modo mais eficiente de atuação; à viabilidade prática (política e operacional) de propostas alternativas de organização do sistema; e, eventualmente, até mesmo à conveniência de estruturar-se tal sistema. Não cabe, no entanto, questionar a possibilidade de sua existência.

A concepção de um sistema estatístico na primeira acepção indicada acima não é tão simples. Nesse caso, cabe questionar a própria idéia de um sistema estatístico nacional ou, mais precisamente, a possibilidade de estruturar e organizar a totalidade da produção estatística do País como um sistema.

*Economista, professor do IEI/UFRJ e atual presidente do IBGE

A esse respeito, é interessante observar que a legislação brasileira da década de 70 se refere a um Sistema Estatístico Nacional quando trata do aparato institucional da produção de estatística, mas formula apenas um Plano Geral de Informações Estatísticas (e Geográficas) quando se propõe a indicar e organizar o conjunto de informações necessárias ao conhecimento da realidade econômica e social do País.

Recorrendo à terminologia desses textos legais, a questão sugerida acima corresponde a indagar da possibilidade de conceber um Plano Geral de Informações Estatísticas como um sistema – vale dizer, como um conjunto ou uma totalidade de objetos, reais ou ideais, reciprocamente articulados e interdependentes (as noções de unidade, totalidade e interdependência entre os elementos constitutivos são três noções fundamentais implícitas no conceito de sistema).

Parece lícito afirmar que essa possibilidade depende, antes de mais nada, de que o mundo real ao qual as estatísticas se referem seja apreendido como um sistema, o que pressupõe a existência de esquema teórico que construa esse sistema. Depende também de que esse esquema teórico e os fenômenos focalizados sejam passíveis de quantificação.

Tais considerações sugerem que a principal dificuldade ou, mais do que isso, a impossibilidade mesmo da construção de um sistema estatístico nacional decorre da inexistência de um esquema teórico que capte a totalidade da realidade econômica e social. Na verdade, em face da ausência dessa teoria geral, cada campo teórico específico tem como contrapartida um recorte particular, parcial e próprio do mundo real. A esse respeito, cabe observar, aliás, que tampouco as chamadas abordagens multidisciplinares permitem avançar na direção da construção de um sistema mais geral; de fato, consistindo basicamente da superposição de abordagens disciplinares, o enfoque multidisciplinar, mesmo quando útil para a compreensão dos fenômenos do mundo real, é incapaz de estruturá-lo em um sistema.

Nesse contexto, a combinação de um campo específico de conhecimento com o recorte do mundo real correspondente delimita um espaço particular para a produção de estatística e aponta para a estruturação de um sistema estatístico específico. Assim, o sistema estatístico nacional aparece como uma virtualidade, cuja realização, de resto, depende de desenvolvimentos teóricos que escapam ao âmbito do próprio Sistema Estatístico Nacional.

Isto não exclui, no entanto, a possibilidade de conceber e estruturar a produção estatística como sistema. É certamente possível a construção de sistemas específicos e parciais – seja como sistemas inteiramente isolados e independentes, seja como sistemas que se articulam a outros sistemas igualmente específicos e parciais na constituição de sistemas mais amplos. Tais sistemas específicos podem eventualmente ser considerados como subsistemas de um virtual sistema estatístico nacional.

É interessante notar que esse processo de construção de sistemas mais amplos a partir da agregação e articulação de blocos menores – que se sugere como o caminho

possível para a produção de estatística – não difere do próprio processo de produção teórica. De fato – embora, do ponto de vista lógico, teorias parciais se construam dedutivamente, desdobrando-se de um corpo teórico mais geral – do ponto de vista da prática de produção do conhecimento, os grandes sistemas teóricos se desenvolvem tanto pelos desdobramentos dedutivos de novos segmentos teóricos, sugeridos pela lógica interna do próprio sistema, quanto pela agregação de blocos parciais que induzem construções teóricas mais envolventes a nível mais geral.

II

A distinção usual entre estatísticas econômicas e estatísticas sociais é freqüentemente caracterizada como uma segmentação de um virtual sistema estatístico nacional em dois subsistemas. Tal enfoque não parece, no entanto, pertinente: embora o conjunto das "estatísticas econômicas" seja passível de estruturação em um sistema, o mesmo não sucede com as "estatísticas sociais". Convém, portanto, examinar aqui os fatores que explicam essas diferentes situações.

Considerem-se inicialmente as estatísticas econômicas. Aqui, a produção de estatística tem como marcos de referência um campo específico de conhecimento (a teoria econômica e/ou a economia política) e o recorte do mundo real correspondente (o sistema econômico capitalista).

Tais marcos de referência dão origem a diversos subsistemas estatísticos, referidos cada um deles a segmentos específicos do sistema econômico e/ou a corpos teóricos particulares. Por outro lado, na medida mesmo em que a teoria econômica constrói o conceito de sistema econômico, ela contém os elementos necessários à integração sucessiva dos subsistemas estatísticos que ela suscita em sistemas mais amplos, convergindo para um sistema unificado de estatísticas econômicas. Na verdade, algumas categorias econômicas (por exemplo, moeda e preços, capital e lucro, empresa e mercado) e algumas características do sistema econômico (por exemplo, a natureza monetária de todas as relações econômicas, a produção para mercado) cumprem um papel fundamental, ainda que nem sempre explícito, do ponto de vista da possibilidade da estruturação de um sistema de estatísticas econômicas. Tal processo de estruturação é, de resto, facilitado pela natureza quantitativa dos conceitos utilizados e das relações que se estabelecem no âmbito do sistema econômico.

É importante assinalar, no entanto, que, não obstante a existência desses fatores favoráveis à integração crescente das estatísticas econômicas em sistemas articulados, a sistematização final dessas estatísticas – traduzida na matriz de relações interindustriais e, sobretudo, no sistema de contas nacionais – não foi espontânea mas sim precedida de desenvolvimentos teóricos que apontaram, a partir de suas formulações analíticas, para o formato de tais sínteses estatísticas. Por conseguinte, o sistema estatístico síntese assim estruturado (como de resto qualquer sistema) depende, enquanto sistema, do esquema teórico a que está associado.

Considerem-se agora as estatísticas sociais, como caracterizadas no contexto de uma segmentação dicotômica de um virtual sistema estatístico nacional. Nesse caso, as estatísticas sociais são definidas, de certo modo, como resíduo, a partir de sua natureza não-econômica. Não é de surpreender, portanto, que esse conjunto de estatísticas seja dificilmente articulável em um sistema.

Há, no entanto, razões menos contingentes que obstaculizam a estruturação de um sistema de estatísticas sociais: a inexistência de um corpo teórico que dê conta da totalidade do social e que, portanto, ao estruturá-lo como sistema, permitisse também a construção de um sistema de estatísticas sociais; e a natureza freqüentemente não quantitativa das categorias analíticas utilizadas e das relações sociais.

Tais limitações explicam por que mesmo o esforço mais sistemático de organização das estatísticas sociais, a construção dos indicadores sociais, está longe de caracterizar um sistema como definido inicialmente. Na verdade, os indicadores sociais implicam uma certa organicidade na utilização e nos propósitos das diversas estatísticas sociais, mas não parecem assegurar a integração e unidade conceitual do processo de produção dessas estatísticas.

Não obstante, também aqui é possível construir sistemas estatísticos parciais (ou, mais uma vez, subsistemas de um virtual sistema estatístico nacional) que focalizem segmentos específicos do social, ao mesmo tempo em que se procura articular progressivamente esses diversos subsistemas. As estatísticas demográficas constituem um exemplo de um subsistema estatístico particularmente bem estruturado.

III

Convém ressaltar aqui o ponto central das considerações anteriores. Sugeriu-se a impossibilidade de conceber um Plano Geral de Informações Estatísticas para o País como um sistema que abarque a totalidade da realidade econômica e social. Neste sentido, a estratégia de estruturação da produção nacional de estatística deve contemplar a construção de sistemas parciais e específicos e a progressiva articulação desses sistemas como subsistemas de sistemas mais amplos. Esse processo de integração progressiva apresenta, hoje, perspectivas mais promissoras no tocante às estatísticas econômicas, inclusive porque a matriz de relações interindustriais e o sistema de contas nacionais constituem um ponto de convergência e um princípio organizador de todo o conjunto das estatísticas econômicas.

Cabe enfatizar, no entanto, que essa necessidade de integrar os subsistemas estatísticos em sistemas mais amplos, sugerida pela estratégia proposta, não exclui o fato de que os subsistemas têm existência, objetivos e motivações próprias e independentes dos sistemas mais amplos, que cabe respeitar e preservar. Essa existência própria e específica pode ter, aliás, o mesmo *status* que os sistemas mais amplos, já que se apóia, freqüentemente, em corpos teóricos bem estabelecidos que focalizam aquele recorte particular do real ao qual o subsistema está referido. Assim, é importante

evitar que a preocupação totalizante dê lugar a uma postura que conduza à negação e à destruição da especificidade dos subsistemas.

Não obstante, é também verdade que o objetivo de uma integração progressiva dos diferentes subsistemas impõe que cada um deles se ajuste às exigências impostas por sua inserção em sistemas mais amplos. Assim, é igualmente importante evitar que a necessária autonomia de cada subsistema particular seja um fator disruptivo da construção de sistemas mais gerais.

Nesse contexto, o processo de estruturação da produção nacional de estatística – que deve se traduzir no Plano Geral de Informações Estatísticas – envolve dois movimentos distintos mas simultâneos e inseparáveis. O primeiro consiste na construção de cada um dos subsistemas e no aprimoramento de suas estruturas internas. O segundo corresponde à articulação entre os diversos subsistemas, tendo em vista a integração em sistemas mais amplos. A simultaneidade desses dois movimentos implica que nem um deles pode avançar de forma inteiramente independente, sendo o desenvolvimento proposto em relação a cada um desses movimentos constantemente questionado pelas exigências formuladas no âmbito do outro. Trata-se, portanto, de compatibilizar, a cada instante, o que é específico do subsistema com o que é comum aos demais subsistemas. Acrescente-se ainda que, freqüentemente, um dado subsistema se integra a mais de um sistema mais amplo, o que requer a compatibilização simultânea do específico e da demanda desses vários sistemas mais amplos.

Considerem-se, a título de exemplo das questões aqui tratadas, os desenvolvimentos recentes na área das estatísticas relativas aos diversos setores produtivos. O IBGE vem produzindo, há várias décadas, estatísticas relativas à agricultura, à indústria, ao comércio e aos serviços. Embora em relação a esses dois últimos setores a produção do IBGE tenha se restringido aos censos, é possível identificar a existência de um sistema de estatísticas agropecuárias e de um sistema de estatísticas industriais, ainda que deficientemente estruturados. Tais sistemas, bem como os censos comercial e de serviços, apresentam-se, no entanto, em sua origem, bastante desarticulados.

A partir do final dos anos 70, a matriz de insumo-produto, como um sistema mais amplo, passa a constituir um ponto de convergência desses distintos sistemas, a lhes dirigir demandas concretas e a exigir respostas comuns. A proposta, esboçada no Censo de 1985, de construir um sistema de estatísticas sobre a totalidade do sistema produtivo, estruturando-o em dois níveis distintos – o primeiro associado aos conceitos de indústria e mercado, o segundo ao conceito de empresa – veio gerar uma nova demanda de integração em relação aos subsistemas de estatísticas industriais, comerciais e de serviços, apresentando exigências freqüentemente distintas daquelas formuladas pela matriz de insumo-produto.

Assim, o movimento em curso no sentido de consolidar a estruturação de um subsistema de estatísticas industriais, além de dar conta de questões que lhe são específicas, deve também responder às exigências que lhe são dirigidas pela matriz de insumo-

produto e pelo sistema integrado de estatística sobre o sistema produtivo. Da mesma forma, o esforço recente de construção de um subsistema de estatísticas comerciais deve responder também a essas múltiplas demandas. Por fim, o subsistema de estatísticas agropecuárias, em que pese sua especificidade e sem o sacrifício dela, está também certamente chamado a integrar-se nessas iniciativas de construção de sistemas estatísticos mais abrangentes.

Enfatizou-se que o processo de estruturação da produção nacional de estatística envolve um duplo movimento – o da estruturação interna dos diversos subsistemas e o da articulação entre esses subsistemas. Cabe indicar aqui, também a título de exemplo, algumas questões relevantes associadas a esses movimentos a serem explicitados na definição de um Plano Geral de Informações Estatísticas.

Assim, do ponto de vista da definição do elenco de pesquisas em torno das quais se estrutura um subsistema específico, são pontos relevantes: a definição teórica das variáveis a serem investigadas e os mecanismos operacionais para captá-las; a identificação dos instrumentos de pesquisa (levantamentos censitários, pesquisas por amostra, *surveys* localizados, registros administrativos) adequados à natureza e às características das unidades e das variáveis investigadas; a periodicidade das distintas pesquisas e a articulação entre elas; a articulação entre o nacional e o regional; a questão da dimensão dos diversos segmentos do universo pesquisado e a relevância das variáveis; as necessidades de dar tratamentos distintos a questões ou informantes distintos. Do ponto de vista do movimento de articulação entre os diversos subsistemas, destacam-se como questões centrais a compatibilização dos sistemas de classificação e a busca de consistência conceitual entre as variáveis utilizadas em distintas pesquisas.

IV

A construção dos sistemas estatísticos vem sendo tratada nesse texto de forma impessoal, como se as demandas formuladas e as motivações que inspiram o desenvolvimento desses sistemas se explicitassem por si mesmas. Convém, portanto, focalizar aqui a participação nesse processo dos agentes a ele associados – as entidades produtoras de estatísticas e os usuários.

É certamente trivial a afirmação de que produção de estatísticas deve decorrer da interação entre esses dois agentes, de modo a combinar o *know-how* dos produtores de estatística com a competência dos usuários para definir as necessidades de informações estatísticas. Nesse sentido, o diálogo entre esses agentes permitiria identificar, dentre o que é desejável, aquilo que é possível.

Por outro lado, uma versão mais mal humorada da relação produtor × usuário sugeriria que o usuário é aquele que quer tudo, para todo o universo, com o maior detalhe e no menor prazo, enquanto que o produtor é aquele que nega sempre a possibilidade de atender a demanda do usuário na forma desejada e que sistematicamente desrespeita prazos.

Na verdade, tal interação produtores e usuários não é simples, nem são os respectivos papéis tão claramente definidos.

Destaque-se, em primeiro lugar, uma certa assimetria entre a postura desses distintos agentes, já que o produtor de estatística assume freqüentemente a perspectiva do usuário – muitas vezes, aliás, apenas a perspectiva de usuário – enquanto o usuário dificilmente se dispõe ou dispõe da informação e conhecimento necessários para avaliar suas demandas do ponto de vista do processo de produção de estatística.

Ressalte-se aliás que "avaliar uma demanda por informações estatísticas do ponto de vista do processo de produção de estatística" significa não apenas determinar a viabilidade técnica e operacional de atender essa demanda mas também determinar o custo desse atendimento. Curiosamente, a questão dos custos parece ser um fator de tensão entre produtores e usuários de importância muito menor do que se poderia esperar. Infelizmente, essa harmonia não parece decorrer da aceitação de parte do usuário dos limites impostos pelos custos mas sim da usual despreocupação dos produtores quanto a esse aspecto do processo de produção de estatística.

Essa despreocupação com custos tem várias causas. Reflete, por exemplo, uma postura "técnica" freqüentemente encontrada nos produtores de estatística que, de um lado, adotando a perspectiva do usuário, enfatiza a utilidade da informação a ser produzida e, de outro, se vendo como produtor, se entusiasma com o desafio implícito na dificuldade ou na grandiosidade do projeto. Uma segunda causa dessa despreocupação com custo é o despreparo e a ineficiência administrativa e gerencial de inúmeras instituições produtoras de estatísticas que implicam a total ignorância quanto aos custos dos projetos desenvolvidos. Por fim, mas não menos importante, há o fato de que a produção de estatística é uma atividade governamental, financiada com recursos governamentais.

Nesse contexto, estabelecem-se, com freqüência, entre produtores e usuários, comunhões de interesses técnicos e entusiasmos profissionais, bem intencionadas mas, muitas vezes, extremamente dispendiosas.

Cabe ainda observar que a relação produtor × usuário não é freqüentemente uma relação bilateral, mas pode envolver uma multiplicidade de usuários com demandas diferenciadas que, ainda que não exatamente conflitantes, tendem a superar, quando agregadas, a possibilidade de atendimento da instituição estatística.

Ao lado da relação produtor × usuário, outro aspecto importante da definição e construção de sistemas estatísticos é a interação entre as diversas instituições produtoras de estatística. O Sistema Estatístico Nacional (agora com iniciais maiúsculas), previsto em lei, constitui a estrutura institucional no interior do qual deve se estabelecer essa interação. A caracterização desse espaço institucional como um Sistema pressupõe, desde logo, que a produção de estatística no País deve se estruturar com um certo grau de descentralização, envolvendo diversas entidades; pressupõe também, no entanto, que essa produção deve se processar de forma coordenada e articulada.

É sabido que, embora previsto em lei, esse Sistema jamais se estruturou; tampouco tem sido efetivamente exercida a função de coordenação do Sistema atribuída ao IBGE.

Não é intenção desse texto abordar a questão da estruturação institucional do Sistema Estatístico Nacional. Cabe apontar, no entanto, distintas direções e características que pode assumir, na prática, o processo de implantação desse Sistema. Uma primeira alternativa é constituir um processo meramente político-administrativo que distribui áreas de competência e procura evitar conflitos de interesses, criando áreas autônomas de atuação. Essa alternativa, embora indesejável, representa certamente a linha de menor resistência e significaria, provavelmente, apenas referendar as especializações que foram espontaneamente se definindo ao longo do tempo. Alternativamente, é possível, em tese, admitir uma estratégia em que se redefinem as atribuições das entidades participantes do Sistema e o órgão coordenador – o IBGE – passa a gerir de forma centralizada suas atividades. Essa segunda alternativa, além de indesejável, corresponde certamente à linha de maior resistência e não apresenta nenhuma viabilidade política.

Por fim, uma terceira alternativa consiste em preservar a autonomia das diversas entidades produtoras de estatísticas, procurando no entanto integrá-las no processo de construção dos diversos subsistemas estatísticos. Essa integração pressupõe a articulação e a coordenação – eventualmente operacional mas sobretudo metodológica – das entidades produtoras em torno de projetos específicos inseridos no âmbito de um Plano Geral de Informações Estatísticas, sem prejuízo evidentemente de sua capacidade de levar adiante empreendimentos próprios.

Tal esquema de organização e operação do Sistema Estatístico Nacional constitui, na verdade, a contrapartida institucional e operacional da estratégia de estruturação da produção nacional de estatísticas proposta anteriormente, que contempla a construção de sistemas parciais e específicos e a progressiva articulação desses sistemas como subsistemas de sistemas mais amplos.

Nesse sentido, a definição de um Plano Geral de Informações Estatísticas aparece como condição preliminar também para a estruturação efetiva do Sistema Estatístico Nacional.

DESAFIOS À ESTATÍSTICA E AOS ESTATÍSTICOS*

Ivan P. Fellegi**

Minha primeira e mais agradável tarefa é agradecer ao Governo da França por ter convidado o **Instituto Internacional de Estatística** para promover, em Paris, sua Sessão Bianual de 1989. Há uma significação simbólica para este evento.

O período que conduziu à Revolução Francesa há 200 anos atrás é particularmente expressivo na história da Estatística. O uso pertinente da palavra "Estatística" surgiu no curso desse período e no contexto de estudos quantitativos de populações humanas. Como ressalta *Porter*, em seu recente livro sobre o Crescimento do Pensamento Estatístico, "(...) a estatística tendeu a equalizar indivíduos. Não faz sentido contar pessoas, se o que há de comum no fato de elas serem pessoas não for, de algum modo, considerado mais significante do que suas diferenças. O Antigo Regime viu nas pessoas não seres autônomos, mas membros de comunidades específicas. Eles possuíam não direitos individuais, mas um labirinto de privilégios, dados pela história, identificados à natureza, e herdados do berço. O mundo social estava muito intrincadamente diferenciado para um mero censo expressar muito sobre o que realmente importava". O espírito dos tempos, culminando na Revolução, foi um poderoso estímulo para a evolução do pensamento estatístico.

É impossível falar sobre aquele período sem manifestar admiração pela constelação de cientistas franceses cujas contribuições marcaram época na estatística, imediatamente antes e depois da Revolução. No topo de tal lista deve estar o trabalho desbravador de *Laplace*, não apenas em probabilidade e no que hoje chamaríamos estatística matemática, mas também em demografia, técnicas de amostragem e astronomia. Seu contemporâneo, *Legendre*, descobriu e forneceu uma elegante prova do método de

*Discurso de abertura da 46ª sessão bianual do International Statistical Institute - ISI - em Paris.

**Presidente do ISI, Chief Statistician of Canada.

mínimos quadrados. O famoso físico *Fourier* foi o diretor do *Bureau de Statistique du Departament de la Seine* e deu importantes contribuições, tanto à estatística matemática, como à publicação e análise de estatísticas oficiais. E, finalmente, *Poisson*, cujo trabalho na teoria das probabilidades deveu-se à sua ampla gama de interesses em pesquisa, abrangendo desde a mecânica e teoria do calor até o direito penal na França.

Além da curiosidade e brilho, estes homens compartilhavam a característica de que nenhum se teria considerado como estatístico. Eles foram conduzidos à estatística através de seu interesse em ciência, dados e análise, o qual explica por que, em muitos países, e mesmo internacionalmente, associações estatísticas estiveram entre as primeiras sociedades científicas. Pessoas de várias disciplinas, que necessitavam compartilhar dados escassos e que queriam trocar conhecimentos sobre métodos comumente aplicáveis de análise descritiva ou inferencial, tinham uma mesma forte e inusitada motivação para estabelecer elos uns com os outros. Sua grande motivação para reunir dados e conhecimentos deu origem à nova disciplina de estatística. Um importante fator que contribuiu para isto foi o interesse do estado, – inicialmente por aspectos demográficos, mas logo por uma ampla gama de problemas relacionados ao comércio, produção, saúde, crime, etc.

Tendo-se originado das demandas coletivas de outras disciplinas e do estado, a estatística se tornou uma disciplina bem-sucedida e auto-sustentada. Contudo, este mesmo sucesso tem seus perigos – e esta é minha principal mensagem: o de que um pré-requisito para continuar nossa bem-sucedida evolução é o de que devemos alcançar nossos usuários e colegas cientistas com brevidade e perseverança. Pois nós somos desafiados por importantes pressões e oportunidades.

Eu me concentrarei em questões relacionadas com estatísticas governamentais, que melhor conheço, e enfatizarei coisas que podemos e devemos fazer para melhor responder aos desafios.

1. Informação Estatística é um produto com atributos peculiares. Um deles é que seus usuários estão raramente na posição de verificar diretamente sua qualidade. Mesmo assim, dados não confiáveis são de pequena utilidade, qualquer que seja sua qualidade intrínseca. Mas, na falta de controle direto de qualidade, o grau de confiança que os usuários atribuem ao produto é necessariamente uma função direta de sua confiança no produtor. Tal confiança do usuário só pode ser atingida através de alto profissionalismo, objetividade e consciência de qualidade da organização produtora. Mas isto pode não ser suficiente. Estatísticas governamentais devem perseguir uma imagem profissional: pela explicitação de metodologias e limitações de qualidade; pelo convite a revisões de nossos produtos por parte de grupos de consultores imparciais externos, altamente renomados; pela participação em atividades profissionais; pelo contínuo esclarecimento de comentaristas, que escrevem sobre aspectos ordinários de problemas econômicos e sociais; pela refutação de relatos enganadores da mídia.

2. Um outro atributo dos dados estatísticos é o de que a acumulação, ao longo do tempo, de dados relacionados, incrementa desproporcionalmente seu valor. Por exemplo, é útil conhecer que o desemprego cresceu. Mas é imensamente mais útil, se for possível, analisá-lo à luz do que diferentes indústrias produzem, de como a produtividade está mudando a longo prazo, das mudanças na importação e exportação, dos lucros, etc. Não obstante, organizamos nossos resultados da forma conveniente às nossas operações internas, não a partir da perspectiva de nossos usuários. A maioria de nós produz publicações demais baseadas em levantamentos e muito poucas sinopses sobre temas específicos; muitas tabelas e poucas análises; demasiados produtos impressos e poucas saídas na forma eletrônica, dificultando análises secundárias. É essencial compatibilizar melhor nossos produtos com as necessidades de nossos usuários.

3. Uma pressão diferente deriva da dificuldade de estabelecer as prioridades entre programas conduzidos por órgãos oficiais de estatística. A demanda excede de longe os recursos; não há ordenação única de programas estatísticos correntes e muito menos dos programas prospectivos. Ainda assim, a despeito de dificuldades, nós devemos fazer e fazemos julgamentos de prioridades. A qualidade e a defensibilidade de tais julgamentos dependem diretamente de nosso compromisso em detectar o leque de necessidades do usuário. Necessitamos de consultas bilaterais e multilaterais explícitas a departamentos do governo, em todos os seus níveis; de encontros regulares com associações de indústrias e representantes de sindicatos; de participação ativa nas diversas associações profissionais; do *feedback* de análise de mercado de vendas de nossos produtos estatísticos. Eu enfatizaria especialmente mais uma atividade-chave: a de propiciar análises substantivas. Análises exaustivas e de grande amplitude de seus próprios dados, são essenciais para órgãos de estatística desenvolverem em si próprios uma perspectiva de usuário – incluindo uma compreensão aperfeiçoada de problemas causados por hiatos nos dados, sua qualidade e o acesso a eles.

4. As relações com entrevistados criam uma outra pressão. Os benefícios gerados a partir de bons dados estatísticos são indiretos, para a maioria de entrevistados individuais. Em contrapartida, o transtorno de responder é diretamente sentido. Para tais entrevistados, é natural questionar o valor da coleta de dados. A rigorosa manutenção de confidencialidade e a apropriada resposta às preocupações de privacidade são condições necessárias para manter a cooperação do entrevistado. Mas, em muitos países, estes cuidados básicos deixaram de ser suficientes.

Um esforço tem que ser feito, pesquisa a pesquisa, para explicar aos entrevistados como a sociedade se beneficia dos dados que eles são solicitados a fornecer. Para tal explicação ser convincente, órgãos estatísticos devem desenvolver uma imagem de importância pública e de legitimidade. Uma tal imagem é adquirida ao longo do tempo, através do número e qualidade de referências feitas pela mídia e personalidades públicas às informações divulgadas pelos órgãos de estatística. Isto, por seu turno, é facilitado por um fluxo constante de produtos "amigáveis": manchetes de

efeito acompanhando publicações de dados, produtos analíticos objetivos tratando de problemas socialmente importantes, apresentações públicas, resumos tão profundos quanto possível aos representantes da mídia. Para esclarecer, é difícil achar a fronteira exata entre análises objetivas de problemas públicos (inteiramente apropriadas para órgãos estatísticos) e a advocacia de políticas (que não o é!). É igualmente difícil achar a fronteira exata entre o que deve ser citado como manchete estatística (inteiramente apropriado) e a sensacionalização de achados (que não é!). Estas dificuldades não devem, contudo, deter-nos na tentativa de nos estabelecer no juízo público como fonte da maioria das informações quantificadas, necessárias à compreensão daquelas questões que preocupam o público, com as quais se defrontam nossas respectivas sociedades. Sem isto, não poderemos legitimar o incômodo de informar que colocamos sobre nossos entrevistados.

5. Órgãos estatísticos têm uma estrutura especial de custos, caracterizada por uma alta proporção de custos fixos. Eles necessitam desenvolver uma substancial capacidade de computação, organização de campo, agentes de coleta, equipes de amostragem, pessoal necessário para manter e aplicar sistemas de classificação, bancos de dados e um amplo espectro de profissionais com especialização substantiva nos domínios onde aqueles órgãos atuam. Numa tal infra-estrutura básica, a variável custo, relativa aos produtos estatísticos específicos, é usualmente muito baixa. Esta é uma vantagem em época de recursos abundantes, mas o inverso também é verdadeiro: dados seus custos marginais baixos, grande número de programas têm de ser cortados para proporcionar economias expressivas. Permitam-me, apenas, apontar um corolário evidente: pequenos aumentos nas receitas da venda de produtos podem ajudar a preservar informações críticas, em tempo de cortes orçamentários.

Geração de receita, até certo ponto, pode ser benéfica, se aos órgãos estatísticos for permitido conservar a receita resultante e se ela for auferida com a devida consideração pelo interesse público. Não apenas ela pode contribuir para a preservação de informações estatísticas importantes, em face de cortes orçamentários, mas também no incremento, que imprime, de maior orientação para o usuário. Permitam-me asseverar, a partir de experiência pessoal, que nada motiva tanto a revisão crítica de produto publicado do que a necessidade de vendê-lo a preços realísticos. Isto fornece um incentivo ao engajamento no *marketing*, no melhor sentido da palavra: i.e., o de interagir com clientes, a fim de descobrir e responder a suas reais necessidades.

6. Obter recursos para o trabalho de órgãos estatísticos é uma outra fonte de pressão. De um lado, a uma interminável demanda de bens que sejam gratuitos. Por outro lado, um órgão estatístico governamental completamente orientado pelo mercado é, ao mesmo tempo, técnica e politicamente difícil de conceber. Tecnicamente, porque o custo de coletar e compilar informação estatística é geralmente alto. Se o custo inteiro devesse ser recuperado através da venda de produtos, estes seriam de altos preços. Mas, dado o baixo custo de reprodução, qualquer um poderia comprar uma única

cópia, duplicá-la e revendê-la a uma fração do preço cobrado pelo órgão estatístico. Alternativamente, a recuperação completa do custo exigiria que se operasse através da compra, por usuário, da totalidade de empreendimentos estatísticos. Mas a alocação de custos entre múltiplos clientes seria difícil – e.g., quem pagaria, e por qual porção do custo, para se desenvolverem estimativas do Produto Nacional Bruto? Se um só usuário pagasse suas necessidades receberia especial atenção, em detrimento da objetividade perseguida para a informação resultante.

Assim, não obstante os benefícios da geração de receitas, parece não haver nenhuma alternativa ao custeio direto pelo governo de uma grande proporção do custo das estatísticas oficiais.

Mas recursos governamentais, entre outros fatores, dependem de forte suporte político, razão adicional para enfatizar a importância de um clima no qual os usuários apóiem ativamente, e a população-alvo se abstenha de hostilidade. Mas isto é apenas condição necessária mínima. Decisões sobre a captação de recursos direcionados para programas estatísticos são tomadas, ordinariamente, por políticos, sob a inspiração de pequeno número de burocratas graduados. Por seu lado, estes funcionários podem depender agudamente de boa informação estatística, mas não terem senão apenas vaga consciência de sua própria dependência, pois recebem, ao invés de dados, análises elaboradas por subordinados. É, portanto, essencial, para os órgãos estatísticos, desenvolverem canais regulares de comunicação com funcionários graduados: seja através de cartas pessoais, sublinhando achados significantes, seja pela participação direta em comitês interdepartamentais, seja pela apresentação de análises estatísticas aos analistas que lidam com questões públicas relevantes, etc.

7. Em anos recentes, cortes nas despesas governamentais têm representado uma importante pressão sobre órgãos de estatísticas de muitos países. Dada a preocupação com déficits, muitos governos optam por reduzir desproporcionalmente despesas em pesquisas e desenvolvimento e em infra-estrutura de informação. Creio que isto é contrário ao interesse nacional. Sistemas estatísticos governamentais, nos anos pós-guerra, exigiram estatísticas para o gerenciamento de políticas macroeconômicas e para monitorar a evolução do bem-estar público. Estas exigências são ainda válidas, mas hoje há preocupações microeconômicas adicionais com indústrias particulares, regiões e desenvolvimento de mão-de-obra. Há crescentes preocupações com ajustamento ou reformas necessárias ao bem-estar público, implicando novas necessidades estatísticas para medir os impactos dos sistemas sociais, tais como cuidados dispensados à saúde, à educação e a programas de redistribuição de renda.

A carga de débito suportada pela maioria dos países do terceiro mundo, os déficits de muitos países industrializados e os enormes saldos de uns poucos outros implicam significativas pressões estruturais na economia mundial, que, mesmo quando forem solucionados, terão impactos sociais e econômicos de longo alcance. Políticas para fazer frente aos ajustamentos resultantes requerem não apenas informação adicional, mas

grandes avanços são necessários para integrar sistemas de dados sociais e econômicos. Ademais, o desenvolvimento de informações vitalmente necessárias sobre o ambiente exige não apenas grandes avanços conceituais, mas íntima colaboração internacional.

Como dever profissional, é nossa tarefa urgente comunicar as graves conseqüências futuras do apoio inadequado ao desenvolvimento da informação.

Venho falando sobre estatísticos governamentais, mas há necessidade de atingir, da mesma forma, os outros estatísticos. *Leslie Kish*, em seu discurso presidencial na *American Statistical Association*, põe o dedo na questão: "... a Estatística difere fundamentalmente das outras ciências. Os dados de outros cientistas vêm principalmente de suas próprias disciplinas ... Em acentuado contraste, estatísticos não têm campos próprios onde colher seus dados. Os estatísticos obtêm todos os seus dados a partir de outras áreas ... A nossa é uma forma simbiótica de vida ..." Como a maior parte da ciência é (ou deveria ser) baseada em dados empíricos, cujas medidas quase invariavelmente envolvem acaso e erro, estatísticos têm, potencialmente, um papel de excepcional importância a desempenhar. Mas como cursos elementares de estatística são largamente ensinados; *softwares* estatísticos "amigáveis" estão amplamente disponíveis; é fácil para nossos colegas de outras disciplinas acreditarem-se autosuficientes. Planejamento experimental adequado, análise apropriada e a obtenção de inferências bem fundamentadas freqüentemente farão a diferença entre o sucesso e a mediocridade. Mas um pré-requisito para o sucesso é o de que devemo-nos tornar parceiros na atividade científica, e não apenas consultores. Devemos, portanto, alcançar outras disciplinas, entender suas preocupações e, baseados em tal entendimento, impressionar nossos parceiros potenciais sobre os benefícios da parceria.

Por último, mas não em último, todos os engajados em estatística aplicada devemos não apenas cooptar estatísticos teóricos, a fim de melhorar nossa prática através do uso das teorias mais apropriadas, mas também chamar sua atenção para a multiplicidade de problemas surgidos de nossa prática que se poderiam beneficiar dos avanços teóricos.

Para resumir a minha mensagem principal, nós enfrentamos um desafio no sentido de "ameaça" e "oportunidade explorável". Estou convencido de que a ameaça é real e, se não enfrentada, pode acarretar dano grave e duradouro à nossa profissão – e, mais importante, à sociedade. A fim de transformar esta ameaça numa oportunidade, devemos incrementar o entendimento (atual e potencial) da contribuição da estatística à sociedade.

A maioria de nós, que somos estatísticos governamentais, vive em democracias. Se nós desejamos melhorar nosso ambiente geral, devemos convencer um grande segmento do público leigo da importância da estatística. Isto ocorrerá através de conferências direcionadas para colegas; através da mídia, na qual a maioria das pessoas obtém o grosso de sua informação geral. Devemos cultivar nossos usuários, tentando ir ao encontro de suas necessidades e reorientando nossos produtos para satisfazê-los mais efetivamente. Um grupo especialmente importante de usuários é o dos executivos-

chefes no governo, nos negócios e nos sindicatos. É essencial achar modos de acentuar, para estes grupos, sua dependência atual e potencial de informação estatística relevante. Temos também de captar o apoio explícito de aliados potenciais: acadêmicos e técnicos de demografia, sociologia, economia, especialistas em áreas tais como saúde, educação, comércio, consultores de negócios, e assim por diante. Devemos envolvê-los na prioridade da revisão de nosso trabalho profissional, fazê-los conscientes de nossos problemas, de modo a poder torná-los nossos porta-vozes autorizados. Também necessitamos deles como parceiros para desenvolver referenciais analíticos, que são pré-requisitos tanto para o entendimento dos fenômenos sócio-econômicos, como para o desenvolvimento de sistemas de dados necessários. Finalmente, estatísticos governamentais necessitam irmanar-se com estatísticos teóricos (e vice-versa!): sua ajuda é necessária para resolver muitos problemas crônicos.

Isto me conduz aos estatísticos fora dos governos. Da mesma forma que estatísticos governamentais necessitam projetar a imagem de operários de uma fábrica de números, assim também estatísticos consultores devem vir a ser vistos como parceiros em empreendimentos científicos – não apenas especialistas a serem chamados quando necessários. Seguramente, uma condição necessária para mudar a imagem é que a realidade deve mudar – pela nossa compreensão das áreas substantivas às quais a estatística é aplicada. Estatísticos teóricos também necessitam aumentar consideravelmente a comunicação de nossas realizações profissionais em áreas aplicadas; e, como professores, eles têm que motivar os alunos não apenas através de uma forma inspiradora de ensinar teoria, mas também articulando consistentemente como essas teorias conduzem a metodologias sólidas e boas aplicações.

Finalmente, creio que o pré-requisito mais importante para a mudança é nossa atitude.

É uma grande satisfação abrir esta conferência – uma demonstração concreta de quanto temos a oferecer.

DATA EDITING AND QUALITY*

Leopold Granquist**

0 — FOREWORD

This paper is an edited version of the presentation material of a minicourse on "data editing and quality" given at the IBGE Workshop on "data editing and imputation methods" in February, 1990. The course consists of two main parts, namely "quality aspects of data editing" and "macro-editing methods". The first part may be considered as a background, justifying the need for other more efficient methods for error detection. The second part gives a solution to the lack of efficiency problem of traditional editing methods. The solution is macro-editing methods. In experimental studies in data processing environment and in production they have proved to be superior to micro-editing methods in editing quantitative data. Savings of the manual verifying work from 35 up to 80 per cent are reported.

1 — OBJECTIVE

The over-all aim of the paper is to make the reader question traditional methods of data editing irrespective of how sophisticated the software support may be. The editing system should focus on important errors and not on possible checks.

*Apresentado no "Workshop sobre Métodos de Crítica e Imputação de Dados" promovido pelo IBGE em Fevereiro de 1990.

**Estatístico do Statistics Sweden.

Another objective is to introduce macro-editing methods as a means for rationalizing traditional micro-editing of quantitative data. Thus a great part of the paper is devoted to an overview of studies on using macro-editing methods. It may serve as a basis for considering macro-editing methods when designing an editing system for a survey with quantitative data.

2 — BACKGROUND - THE TRADITIONAL MICRO-EDITING PROCEDURE

In the seventies the statistical agencies developed computer assisted editing processes of the kind described below. Most surveys are in principle edited in the same way to-day. Thus the following description (which can be found in "BLAISE 2.0 AN INTRODUCTION", p 52) will serve as a background to various approaches to editing or developments of editing systems, software, methods undertaken by statistical agencies all over the world:

> *"After the forms have been collected, subject-matter specialists check them for completeness. If necessary and possible skipped questions are answered, and obvious errors are corrected on the forms. Sometimes, the forms are manually copied to a new form. Next, the forms are transferred to the data entry department, where data typists enter the data in the computer at high speed without error checking. After entry, the files are transferred to a mainframe computer, where an error detection program is run. Detected errors are printed on a list, which is sent to the subject-matter department. Specialists investigate the error messages, consult the corresponding forms, and "correct" the errors on the lists. Lists with corrections are sent to the data entry department, and data typists enter the corrections in the data entry computer. The corrected records are transferred to the mainframe computer, and merged with already present correct records. The cycle of batch-wise error detection and manual correction is repeated until the number of detected errors is considered to be sufficiently small. After the final step of the editing process the result is a "clean" data set, which can be used for tabulation and analysis."*

3 — COST-BENEFIT DISCUSSION ON APPLICATIONS OF THE BASIC EDITING PROCEDURE

3.1 – The cost of editing

The described editing procedure is very expensive and far from rational. Editing and imputation usually accounts for 20 - 40 per cent of a survey budget. That cost does not include the development and maintenance of software, which in most agencies is made in-house.

3.2 – Over-editing

A vast number of checks with too narrow bounds produce lots of error messages, many of which imply expensive contacts with the respondents. The checks are not focused on possible important errors rather the subject-matter experts invent all possible checks, viz. all that can be checked will be checked. This particular problem, nowadays over-editing, is faced by a growing number of statistical agencies.

3.3 – Quality is not improved

The resources spent on editing are justified by quality arguments. Editing is considered a guarantee that the survey will meet some specific quality requirements. However, it is hard to find any evaluation of an editing process, which definitely shows that the quality was improved or that the editing was worth the cost.

3.4 – Editing does not improve the knowledge of the survey data

Another serious objection against the application of a traditional editing procedure is that nothing is learnt about errors, error structures, the problem areas of the survey design and so on. According to Granquist (1984) this is the essential role of editing.

4 — APPROACHES/SYSTEMS/METHODS TO RATIONALIZE THE TRADITIONAL EDITING PROCEDURE

4.1 – References

An excellent, brief, introduction to editing practices and tools for editing is given in Ferguson (1989). A more detailed and comprehensive paper on the same subject is Pierzchala (1988).

4.2 – International co-operation on editing and imputation issues

Statistical agencies work to rationalize their editing systems or editing practices by developing tools, generalized systems and methods.

However, each of the agencies working in this field tends to concentrate on the part of the basic procedure, which seems most important to the agency. Fortunately, the agencies have different opinions or needs, which means that any part of the procedure is covered by at least one country. Some approaches are very expensive and there is no agency that can afford or has resources to work over the whole field. That is why international co-operation has become important. Besides bilateral co-operation, conferences dedicated to data editing and related problems are held. Data editing and imputation is a topic on many conferences on statistical methodology. Within the Statistical Computing Project (SCP-2) of Economic Commission of Europe (ECE) there is an international work group on data editing issues, called "Data Editing Joint Group".

4.3 – Review of international work on improving editing systems or practices

Ferguson (1989) and Pierzchala (1988) cover practically all important work completed or in progress and give references to detailed papers or reports on studies, methodological work, software and so on. Therefore, only very brief comments or examples will be given below for each of the different fields of data editing and imputation.

4.3.1 – Generalized editing and imputation systems (GS)

A Generalized System (GS) executes all checking and imputation. The user only

has to specify his edits according to some syntax rules. The Fellegi-Holt methodology is applied. Some examples are:

Generalized Editing and Imputation System (GEI), developed by Statistics Canada, for quantitative variables.

DIA, developed by the National Institute of Statistics of Spain, for qualitative variables only.

SPEER, developed by the US Bureau of the Census for quantitative variables.

Granquist (1988) is hesitant about GS. To a great extent his criticism is based on the results from an evaluation study of the machine editing of the World Fertility Survey by Pullum et al. (1986). GS are very complicated, cost to much to run, require much storage or machine power and do not solve any important quality problems. They can be justified if:

i) they can be used by subject matter-specialists without assistance by EDP-specialists

ii) they become cheaper to run than at present

iii) they are used with common sense, i.e., only necessary checks are included

iv) the user is aware of the limited improvement of the quality of the data

GS may contribute to the quality, when they provide the possibility to measure the impact of the automatic imputations on the variances of the estimates. The manual up-dating in error-detection systems is most often a completely uncontrolled operation. For instance we do not know anything about the respondent interviews carried out by the clerks, and very little about the imputation method used in specific cases.

4.3.2 – Generalized data editing software

Almost every statistical agency has at least one generalized software system for supporting the traditional micro-editing procedure. It is usually an in-house developed product or a product acquired from another statistical agency or organization. Lately generalized commercial software, like SAS, has begun to be used for editing purposes. Such software has to a great extent replaced the tailor-made programs for detecting errors and for up-dating used in the seventies. Besides an error detecting program almost all of these products contain an inter-active up-dating procedure, which includes checking of the up-dated record. End-users are EDP-specialists or subject matter specialists. Examples: GODAR (developed by the Serbian Statistical Office of Yugoslavia); NASS's system for editing and imputation written in SAS (National Agriculture Statistics Service, USA); UNEDIT (United Nations Statistical Service); CONCOR (U.S. Bureau of the Census); EDIT-78 (Statistics Sweden).

4.3.3 – Data entry/capture editing

Such systems do all the checking when the data are keyed-in, either at the office (Heads Up Data Entry according to Ferguson (1989)) or when the data are collected (by telephone or personal interview). They are eliminating the cycles of a traditional editing procedure and permit the survey staff to process the whole survey without any assistance from other departments or experts of any kind.

BLAISE, developed by the Netherlands Central Bureau of Statistics is the most renowned and certainly the most developed existing system of that kind to-day. See Bethlehem et al. (1989). Compared to the traditional editing procedure, the CBS saves up to 50% of the cost.

According to Granquist (1988) such systems can replace any much more sophisticated editing systems, achieving the same results to much lower costs and, in certain cases, in shorter time.

4.3.4 – On-line or interactive up-dating editing

Such programs or software are used by subject-matter specialists to up-date flagged data from an error detecting program run in batch. The same checking system which produced the error lists or messages is then used to check the up-dated records. Errors founds are displayed on the screen and can be handled immediately by the scrutinizer/verifier, which also means that a record is handled by one scrutinizer only. Thus further cycles are eliminated. Most of these programs are tailor-made and used in connection with tailor-made error-detecting programs for large surveys, where the agency's generalized editing software is not appropriate.

4.3.5 – Developing more efficient error detecting methods

One direction of the development work is to rationalize the editing by focusing on the possible serious errors and to try to find appropriate methods to fight just those errors, instead of designing check systems consisting of all possible checks.

Two important principles in this context are that edits and edit bounds should be based on statistical methodology, and that the manual part of an editing procedure should be carried out by statisticians instead of by clerks.

In this work, macro-editing methods have been found to be much more efficient than the traditional applied methods for editing quantitative variables.

Statistics Sweden is devoting much effort to this matter.

4.3.6 – Data editing as a quality operation

It should be noted that neither traditional micro-editing procedures nor any of the procedures mentioned in 4.3.1 - 4.3.5 will essentially improve the quality of the survey estimates. It is true that certain (kinds of) errors are removed, but it is not clear that the estimates are improved. Below a few references are given to methods which focus on misunderstanding errors and thus yield better estimates.

Werking et al. (1988) present the Response Analysis Survey approach, which implies that in an on-going survey, a special survey on how the respondents have answered certain questions is conducted by a self-response touch-tone technique. By this technique, errors which cannot be found by traditional editing methods are detected, which reduces the bias in the estimates.

Mazur (1990) presents a method to find so called in-liers in survey data. In repetitive surveys there are respondents which for every period report the same figures also to questions which certainly require the answers to change between periods. They are termed in-liers, because they always lie between the bounds of traditional edits. In her experiments, Mazur has found that these in-liers may cause considerable bias in the estimates.

5 — EVALUATIONS AND EXPERIENCES OF EDITING

5.1 – Background

There are only a few published evaluation studies of editing processes. However, all of them indicate that the editing is counter-productive or has not had any impact on the estimates. In some cases where the estimates were notably changed, a very low percentage of the errors was found to account for almost all of the total change.

Studies on distributions of errors and error types form a valuable basis for designing more efficient editing methods. They may also help to persuade subject-matter experts to accept macro-editing methods in their surveys.

Macro-editing methods of the type presented here are always better than micro-editing techniques when applied to checks of quantitative data. They can even be dramatically better in surveys where it is clear that the edits only have to detect the largest errors of those found by the micro-editing procedure.

5.2 – Data error analysis

A data error analysis which focuses on the impact on the estimate by the errors

classified by size may show the potential value of implementing macro-editing methods. Such an analysis is reported in Greenberg and Petkunas (1987).

At Statistics Sweden we have carried out the same type of analysis and arrived at similar results as Greenberg and Petkunas. We found the analytic method valuable, not only for convincing subject-matter experts about the advantages of using macro-edits instead of micro-edits, but also for designing the editing system in such a way as to focus on the error types with important impact on the estimates. Indeed, we are taking steps to get an extended version of the procedure adopted at Statistics Sweden as a compulsory stage in every survey process.

5.3 – An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in the Business Division (Greenberg, Petkunas (1987))

5.3.1 – Contents of the report

The report examines the performance of the edit and imputation programs, proposes a number of recommendations, and explains the analysis undertaken (the basis for the recommendations).

5.3.2 – The evaluation method

All records from six selected kind-of businesses (KB's= a sub-classification in the SIC) were analyzed after each run through the edit and imputation cycle.

To study more in detail the observation that relatively few records accounted for a very major part of the total change, a change file was established. It consisted of all the records for which the keyed-in reported value was different from the tabulation field value.

X_i = reported value for case "i"

Y_i = tabulation field value for case "i"

$i = 1, \ldots, N$ (N = number of cases (changes))

$d_i = |x_i - y_i|$

The cases were ordered by the absolute value of the difference, i.e.: if $i \geq j$ then $d_i \leq d_j$

$$D = \sum_{i=1}^{N} d_i$$

Then d_i/D = the proportion of the total change contributed by the ith case.

They defined $q_i = (\sum_{j=1}^{i} dj/D)100\%$ and $p_i = (i/N)100\%$ for $i = 1, \ldots, N$

q_i represents the percentage of total change contributed by cases for which the change was equal to or greater than the change for case i.

p_i represents the percentage of cases in the change file for which the change was greater than or equal to the change for case i.

Some results are shown in the graphs in the appendix. For all items studied and all KB's approximately 5 per cent of the cases contributed over 90 per cent of the total change. Many of these large changes were due to reporting in units rather than in thousands.

5.4 – The evaluation study on the editing process of one part of the Swedish 1987 Survey on Financial Accounts

5.4.1 – The study

The study covers the editing of all enterprises which have more than 49 employees. The population size is a little more than 4 000. The survey is edited by a basic micro-editing procedure with ratio, validity and consistency checks. The tailor-made error detecting program produces in batch error messages which are verified by very knowledged clerks. The up-dating procedure is inter-active.

5.4.2 – The evaluation method

For three selected variables all records were analyzed after the whole editing procedure was completed. This report covers the variable "Value added". We proceeded exactly as in the Greenberg-Petkunas study.

The results were in quite accordance with the findings of the Greenberg-Petkunas study.

5.4.3 – A modified analysis

We also related the changes to the estimated value of the variable by defining p_i as above and defined the following variables:

N_s = the number of unchanged records (approximately 3 000)

$d_i' = x_i - y_i$

$$D' \sum_{i=1}^{N} (x_i - y_i)$$

S = the total of all (N_s) unchanged values
$C(E)$ = the total of all changed values after editing (y_i)
$C(R)$ = the total of all changed values as originally reported (x_i).

$$r_i = (\sum_{j=1}^{i} d_i' + S + C(R))/(S + C(E))100\%$$

which means that r_i = the percentage of:
("the sum of all positive and negative changes contributed by cases for which the change was equal to or greater than the change for case i" + "the total of all cases as originally reported") to
("the total of all values after editing")

This means that
r_0 = the total relative change
$r_{999} = 100$
(999 was the number of changes for this item of the survey)

5.5 – Other experiences

In all of our studies on macro-editing methods we have found the same distributions of changes made in the editing process as the ones reported above. This can be found in the tables of the change files and it also follows indirectly from the results of the studies on the macro-editing methods.

6 — MACRO-EDITING METHODS FOR RATIONALIZING THE EDITING OF SURVEY DATA

6.0 – Background

As a conclusion of the previous five chapters, the essential problem of traditionally applied micro-editing procedures might be formulated as follows: too many checks with too narrow bounds cause many error-detecting systems to produce too many error messages to be verified manually by clerks. They are not able to assess the importance of a suspected error. Every flagged data has the same weight and claims the same amount of resources, but many errors have a negligible impact on the estimates as

they are small or cancel out. Generally, the bounds of the checks are subjectively set on the principle "safety first" which means that only those data are accepted for which there are no reasons to suspect any error. For example, a very generally used check in business surveys in Statistics Sweden is to flag every data which indicates that the relative change since the previous survey exceeds ±10 per cent. A considerable amount of over-editing is a general consequence of such micro-editing procedures.

6.1 – Objective and contents of the remaining part of this paper

The objective of the remaining part of this paper is to present macro-editing methods as a solution to the over-editing problem. It is a short review of some different macro-editing methods (Chapters 7-11) tested or used on real survey data.

An emphasis is given on the rational aspects of macro-editing as compared to micro-editing by presenting the results of some studies and by discussing the problems connected with micro-editing. The methods are described in a brief and schematic form to make them easy to understand.

Detailed descriptions of the methods and studies are found in the references given in the text and in the reference list. Stress is here laid on the specific features of every method in order to facilitate a choice.

The paper ends with a summing up discussion (Chapter 12) on macro-editing versus micro-editing methods.

6.2 – Definitions

This paper deals with checks on quantitative data which flag "suspicious" data for a manual review. This type of checks may be considered as opposite to validating checks, which indicate data that are erroneous.

Ferguson (1989) calls them "Statistical Edits". They use the distributions of the data to detect possible errors. Such procedures use current data from many or all questionnaires or historic data of the statistical unit to generate feasible limits for the current survey data.

In this paper macro-editing means a procedure for pointing out suspicious data by applying statistical checks/edits based on the weighted keyed-in data. It means that the upper and lower limits of a macro-editing check (macro-edit) should be based

i) only on the data to be edited

and

ii) on the importance of the data on the total level.

6.3 – The evaluation technique

The studies on the methods reported below have been simulation studies on real survey data. The results of the method studied have been compared with the results of the micro-editing methods applied when the survey was processed. The changes made as a result of the editing process of that survey were entered to a change file and the study consisted in investigating (by calculating a few measures) which data in the change file were flagged by the macro-editing method and which were not. The rationalizing effect was measured as the reduction of the number of flagged data and the quality loss as the impact of the remaining errors (the errors found in the processing of the survey, but not flagged by the macro-editing method under study).

6.4 – The experimental data

The studies carried out at Statistics Sweden used data from the Survey on Employment and Wages in Mining, Quarrying and Manufacturing (SEW) and the Survey of Delivery and Orderbook Situation (DOS).

6.4.1 – The Survey on Employment and Wages in Mining, Quarrying and Manufacturing (SEW)

The survey is a monthly, sample survey on employment and wages in the Swedish industry. The number of reporting units (establishments) is about 3000. The main variables are:**number of workers; number of working hours; the payroll; wages/hour.**

A traditional editing procedure is applied, but with inter-active up-dating, which includes checking against the edits. This editing procedure was recently revised, by extending the acceptance limits of the ratio checks, co-ordinating all the checks and tuning the checks. The verifying work was then reduced by 50 % without any drop in quality. It should be noted that the rationalizing effect of the macro-editing methods studied in this survey is compared to the revised survey.

6.4.2 – The Survey of Delivery and Orderbook Situation (DOS)

The survey is a monthly sample survey of enterprises. There are about 2000 reporting units (kind of activity units) in a sample drawn once a year from the Swedish Register of Enterprises and Establishments. DOS estimates changes in deliveries and the orderbook situation (changes and stocks) for both the domestic and the foreign

market (six variables) for the total Swedish manufacturing industry and 38 divisions (classified according to the Swedish standard industrial classification of all economic activities).

The questionnaire is computer printed. The entry of the questionnaire data is carried out in three batches every production cycle by professional data-entry staff. The Top-Down macro-editing method is applied.

6.5 – Macro-editing methods

Descriptions, studies and results are given on *"The Aggregate Method"*, *"The Hidiroglou-Berthelot Method" ("Statistical Edits")* and *"The Top-Down Method"*.

"The Box-Plot Method" and *"The Box-Method"* are also treated, but as modifications or developments of the Aggregate Method and the Top-Down Method respectively. The Box-method is under development and is here treated as an idea.

7 — THE AGGREGATE METHOD

7.1 – References

The original Aggregate Method is described in detail in Granquist (1988). It was developed in SAS as a prototype of a complete editing system for the SEW, to be run on the main-frame. A modified version of the method is reported in detail in Lindström (1990). It was developed in PC-SAS as a prototype for the SEW to be run on a personal computer.

7.2 – Description of the method

Editing on aggregate level followed by editing on micro-level of the flagged aggregates

The basic idea is to carry out checks first on aggregates and then on the individual records of the suspicious (flagged) aggregates. All records belonging to a flagged aggregate of any of the variables form the error file. The checks on the individual level are carried out on that error file.

The acceptance bounds are set manually on the basis of the distributions of the check functions

The most essential feature of the aggregate method is that the acceptance limits are set manually by reviewing lists of sorted observations of the check functions. Only

the "s" largest observations and the "m" smallest observations of the check functions are printed on the lists.

Both the check function A on the aggregate level and the check function F on the individual level have to be functions of the weighted (according to the sample design) value(s) of the keyed-in data of the variable to be edited. By using the weighted values in function A there are no problems to calculate the checks on the aggregate level in the same way as is done in traditional micro-editing. It means that the macro-editing process can be run as smoothly as a micro-editing process.

The lists of the sorted observations can be used either directly as a basis for reviewing observations manually (when identifiers are printed out together with the observations) or indirectly as a basis for setting acceptance limits for an error detecting program, which then produce error messages of suspected data for manual reviewing. The advantage of using an error detecting program is that the process can be made to look identical to the old one. To implement the Aggregate Method, only programs for printing the lists of the sorted observations are needed. Such programs can easily be added to the old system.

Improvements by providing the lists with statistics or graphs

Obvious improvements are to provide the lists of the tails of the distributions with statistics as the median, the quartiles, the range, interquartile range and so on or graphs. Here, Box-plots (see Tukey (1975)) are recommended.

7.3 – The main-frame application on the SEW

7.3.1 – The editing process

The checks used in this application of the Aggregate Method consisted of a ratio and a difference check. Both checks had to be rejected to indicate a value as suspicious.

The aggregates used were the 89 four-digit SIC-groups (SIC = The Swedish Standard Industrial Classification). The questionnaires were processed in lots on arrival at the office. There were four or more editing runs every month.

The macro-editing process was preceded by a micro-editing process in order to fight validation and consistency errors.

7.3.2 – The studies

A AN EXPERIMENTAL STUDY

The first study was a pure experiment. The aim was to find out how methods, computer programs and so on should be constructed. The editing procedure was

applied to the whole body of the edited data for a selected month. We then detected two serious errors, which had not been detected when the survey was carried out. This was an indication that *micro-editing may not always detect even serious errors.*

B STUDY No. 1

The first study was done on a survey a few months after it was processed. However, the records of the data file were divided into lots exactly as when that SEW was processed.

RESULTS:

Number of records: 2 951
Number of flagged records: 274
Number of errors found: 76

When that SEW was processed the number of flagged records was 435 and the number of errors found was 205. Thus, we got a reduction by 161 flagged records = 34 per cent.

The impact of the remaining errors was calculated for each variable, for all the 89 groups for which SEW data are published.

TABLE 1 shows the number of aggregates by the total relative difference in % of the estimates:

DIFFERENCE	WORKERS	HOURS	PAY-ROLL	WAGES/HOUR
0.001-0.05	4	1	6	6
0.1 - 0.4	4	5	5	8
0.5 - 0.9	1	2	4	2
2.5	1	0	0	0

STUDY No. 2

The simulation was run in parallel with the regular processing in order to eliminate the possibility of the results of the ordinary editing influencing the bounds for the checks of the aggregate method.

RESULTS:

Number of records: 2 996
Number of flagged records: 225
Number of errors found: 50

When that SEW was processed, the number of flagged records was 389. The number of errors found was 110. Thus, we got a reduction by 164 flagged records = 42 per cent.

The impact of the remaining errors was calculated for each variable, for all the 89 groups for which the SEW data are published.

TABLE 2 shows the number of aggregates by the total relative difference in % of the estimates:

DIFFERENCE	WORKERS	HOURS	PAY-ROLL	WAGES/HOUR
0.001-0.05	4	2	12	8
0.1 - 0.4	3	6	4	8
0.5 - 0.9	1	1	5	3
1.0 - 1.9	1	1	0	1
2.0 - 2.9	0	0	0	0
3.0 - 3.9	1	1	0	0
4.0 - 4.9	0	0	1	0

This study was carefully analyzed. The most important findings were that: *the acceptance intervals of the checks should be wider and not be symmetric around 1 for the ratios and around zero for the differences.*

The best strategy is to set the limits as close as possible to the first outlier on both sides.

If the limits had been changed according to the findings, the outcome would have been that the number of flagged data would have been 134 which means a reduction of 66 % of the verifying work.

The corresponding quality table on the impact of the remaining errors on the estimates would have become:

TABLE 3 shows the number of aggregates by the total relative difference in % of the estimates (the figures of table 2 within parenthesis):

DIFFERENCE	WORKERS	HOURS	PAY-ROLL	WAGES/HOUR
0.001-0.05	4 (4)	3 (2)	13 (12)	10 (8)
0.1 - 0.4	3 (3)	6 (6)	4 (4)	8 (8)
0.5 - 0.9	1 (1)	1 (1)	6 (5)	3 (3)
1.0 - 1.9	1 (1)	2 (1)	0 (0)	1 (1)
2.0 - 2.9	0 (0)	0 (0)	0 (0)	0 (0)
3.0 - 3.9	0 (1)	0 (1)	0 (0)	0 (0)
4.0 - 4.9	0 (0)	0 (0)	0 (1)	0 (0)

7.4 – The PC-application on the SEW

A prototype of the same editing process was developed in PC-SAS for a micro computer. It was evaluated on the SEW data from August, 1989. However, the realization of the Aggregate Method was somewhat modified. Instead of forming an error file from the aggregate check consisting of all questionnaires of all aggregates

which had failed at least one edit, the records of a flagged aggregate were given a specific signal, telling which of the four variables that did not pass the edit. The check on the micro level was then applied only to those questionnaires which belonged to the aggregates which had failed the check for that variable. This version of the method may imply a small reduction in the number of flagged data, because fewer records are checked on the micro level than in the realization described above.

The result of this study was a reduction in the number of flagged data by nearly 80 per cent.

However, the loss in quality was slightly higher than in the preceding studies due to the modification, the unusually large number of questionnaires in the second processing round, or to the wider acceptance interval. It was found that this loss in quality was caused by a few large errors, which did not cause the aggregates to be flagged. We found that these errors were very easy to detect. One interesting solution was to let all data pass a ratio check with very wide acceptance intervals. This could be done in the micro-editing part of the procedure or as a final check.

The aggregate checks method as such was then questioned. The only advantage of the aggregate checks is that they can save storage or computer time. However, when there are no problems with either the storage capacity or the computer cost, the aggregate checks can be skipped. The method is still a macro-editing method but the term Aggregate Method may not be adequate. When the tails of the distribution of the check function is provided with Box-Plots (which is recommended), this method is called the Box-Plot Method.

7.5 – Conclusion

The simulations show that the Aggregate Method can supply the main part of the *editing* process in the *production processing* of a survey. The method *reduces* the verifying work by 35 - 80 per cent without losses in *quality or timeliness*. The macro-editing concept is a realistic alternative or complement to micro-editing methods, and can be applied during the processing of the data under the same conditions as computer-assisted micro-editing methods, which reduces the manual verifying work to a considerable extent.

For small surveys the Box-Plot Method should be considered.

8 — THE TOP-DOWN METHOD

8.1 – References

The Top-Down Method is described in Granquist (1987) and in Lindblom (1990). The method has been implemented in the main-frame production system of "The Survey of Delivery and Orderbook Situation" (DOS) (see 6.4.2). The production system is written in APL. A prototype written in PC-SAS for running on a micro computer is developed and reported in Lindblom (1990).

8.2 – Description of the method

The idea behind the method is to sort the values of the check functions (which are functions of the weighted keyed-in values) and start the manual review from the top or from the bottom of the list and continue until there is no noticeable effect on the estimates.

The method is described as it is applied in the DOS production system. The generalization is obvious.

The procedure is governed by an inter-active menu program, written in APL. The in-data file is successively built up by the records of the three batches from the data-entry stage. There are three functions to select the records to be studied, i.e.

i) the 15 highest positive changes
ii) the 15 highest negative changes
iii) the 15 greatest contributions

which for every variable can be applied to the total, and to every one of the 38 branches. For a selected function and domain of study, the screen shows the following list for the 15 records of the in-data file sorted top-down:

IDENTITY	DATA	WEIGHT	WEIGHTED VALUE	TOTAL
...	...	...	...	...
...	...	...	...	...

The operator selects a record and immediately gets the entire content of the record on the screen. If an error is identified he can up-date the record on the screen and immediately see the effects. The record usually looses its position on the top 15 list and the total is changed. The operator goes on until further editing does not change the total.

8.3 – Experiences

A direct study to compare the Top-Down Method with a corresponding micro-editing method has not yet been carried out. However, the implementation of the Top-Down Method into the DOS processing system made such an evaluation study unnecessary.

At the same time a micro-editing procedure was developed with the intention that it should be the editing procedure for DOS. When that system was run for the first time it produced so many error messages that the subject matter specialists realized that they had neither resources nor energy to handle them. Especially as they knew by experience that only a small percentage of the flagged data indicated detectable errors. They had constructed the checks on basis of all the experience and subject matter knowledge they had gained during years of work with the survey. The procedure was flexible, user-friendly and easy to fit to the data to be edited. In spite of all that, the procedure did not work.

The Top-Down procedure, which had been developed to be a complementary or reserve procedure, had to be taken in use at once. The staff is very satisfied with it. It is continuously educating the staff on the subject matter and problems of the survey.

Since the first production with the macro-editing process, the number of records for manual review has decreased slowly. The subject-matter statisticians have become convinced that there is no need for editing on the branch level. The Top-Down lists are now only produced at the whole manufacturing industry level. Still, there seems to be a certain amount of over-editing. Anyway it is, doubtless, the most rational editing procedure at Statistics Sweden.

According to Anderson (1989a) the method is also used as an out-put editing method and considered as the most efficient out-put editing method in use at the Australian Bureau of Statistics.

8.4 – Conclusion

We have shown that the Top-Down Method can be used as an editing method during the processing of a survey without losses in quality and timeliness. The method can reduce the verifying work by 50-75 per cent. The subject-matter clerks are very satisfied because they dominate the editing task and can see the effects of their work.

The method should not be applied to more than ten variables of the survey.

9 — THE HIDIROGLOU-BERTHELOT METHOD (STATISTICAL EDITS)

9.1 – References

The Hidiroglou-Berthelot Method (the HB-Method) is described as a micro- editing method in Hidiroglou-Berthelot (1986). The method is a ratio check inspired by Tukey's Explorative Data Analysis (EDA) methods (see Tukey (1977)), and in the paper considered as a solution to some problems connected with the traditional ratio-check method. It is in use at Statistics Canada and known as Statistical Edits.

As a macro-editing method it is reported in Höglund (1989). At Statistics Sweden the HB-Method has been studied on both DOS and SEW data (see 6.4). Only the DOS study has been reported in English (Höglund (1989)).

9.2 – Description of the method

The HB-Method is a ratio method, for which the bounds are automatically calculated from the data to be edited. The method uses the robust parameters median, quartiles (Q_i) and interquartile ranges (D_{rQi}) instead of the mean and standard deviation, because the bounds should not be influenced by single outliers. Then the lower (l) and the upper (u) bounds should be:

$l = R_{\text{MEDIAN}} - k * D_{rQ1}$

$u = R_{\text{MEDIAN}} - k * D_{rQ3}$

However, such a straightforward application of the ratio method has two drawbacks,

i) the outliers on the left tail may be difficult to detect

ii) the method does not take into account that the variability of ratios for small businesses is larger than the variability for large businesses

The HB-Method solves these drawbacks by a symmetric transformation followed by a size transformation.

The symmetric transformation

$$S_i = \begin{cases} 1 - R_{MEDIAN}/Ri, & 0 < R_i < R_{MEDIAN} \\ R_i/R_{MEDIAN} - 1, & R_i \geq R_{MEDIAN} \end{cases}$$

The size transformation

$E_i = S_i * (\underset{0 \leq U \leq 1}{\text{MAX}}(X_i(t), X_i(t+1)))^U$

E_{Q1}, E_{Q3} are the first and third quartiles of the transformation E.

$D_{Q1} = \text{MAX}(E_{MEDIAN} - E_{Q1}, |A * E_{MEDIAN}|)$

$D_{Q3} = \text{MAX}(E_{Q3} - E_{MEDIAN}, |A * E_{MEDIAN}|)$

which gives the lower and upper limits of the checks:

$l = E_{MEDIAN} - C * D_{Q1}$

$u = E_{MEDIAN} + C * D_{Q3}$

A is considered as a constant, equal to 0.05, which means that there are only two parameters, U and C, which have to be set in advance to get the method to run. The real reason behind the symmetric transformation is to get rid of one parameter. We have found that the parameters are not very sensitive. The same values can be used for many variables of a survey.

The figure in the appendix may explain the transformations and the method. It is produced by the prototype of the Box Method for the SEW data. The acceptance limits have been calculated for a few values of the parameters U and C. This method can be applied when implementing the HB-Method forming an operation in the production process. That way the acceptance bounds would be completely determined by the data to be edited.

To make the method a macro-editing method we only had to inflate the keyed-in values of the variable X, in principle by the inverted sample fraction.

9.3 – The study on the Survey of Delivery and Orderbook Situation (DOS)

In the study the keyed-in data of January-88 were used together with the final data of December 1987. A change file was formed by the January-88 data edited with the Top-Down Method and the original keyed-in data of January-88. The H-B Method was tuned and evaluated against the errors (changes) found (the data on the change file) by the current editing procedure, the Top-Down procedure.

Table 1. The change-file of the variable "domestic deliveries" (SEK 1000's)

OBS	DECEMBER	JANUARY		
		Old value	New value	Change
1	14648285	7403131	7403	-7395728
2	1560	1605000	1605	-1603395
3	1032	973693	974	-972719
4	1500	925000	925	-924075
5	1302	902805	903	-901902
6	972	593636	594	-592942
7	1926	501675	502	-501173
8	8473	38010	25328	-12682
9	110893	107891	101520	-6371
10	31745	42201	37201	-5000
11	5000	8300	5200	-3100
12	21930	19013	16730	-2283
13	3465	3084	1007	-2077
14	2879	5495	3650	-1845
15	22839	22400	20977	-918
16	7344	11758	10840	-525
17	1651	1601	1076	-370
18	4072	4061	3691	-370
19	27	262	114	-148
20	8839	1200	1130	-70
21	434	480	462	-18
22	831	664	649	-15
23	11990	9790	9780	-10
24	4834	4404	4398	-6
25	301	116	111	-5
26	1882	4995	4994	-1
27	6245	5658	5659	1
28	274190	271455	271456	1
29	47500	38509	38539	30
30	148	35	80	45
31	1046	1860	1936	76
32	14551	14000	14169	169
33	5383	463	1032	569
34	20400	2500	3600	1100
35	173860	16500	167261	2261
36	6180	1702	4131	2429
37	6000	6000	14300	8300
38	294600	54	54000	53946

The sum of all 38 changes: 12 997 728 (SEK 1000's)

Table 2. The impact of changes found by the HB-Method

Number of changes found	Accumulated sum of changes found	Acc. sum relative to the sum of all changes
7	5 550 152	42,7%
8	5 562 834	42,8%
9	12 958 565	99,7%
10	12 959 665	99,71 %
11	12 960 234	99,71 %
12	12 960 304	99,71 %
13	12 960 305	99,71 %
14	12 962 566	99,73 %
15	12 964 995	99,75 %
16	12 966 840	99,76 %

Table 3. The impact on the other variables edited by the HB-Method with the values on the parameters C and U which were used for the variable "deliveries to the domestic market".

Variable	Deliveries		Orderbook		Stocks	
	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
Number of changes	38	19	109	72	31	30
Number of changes found by HB	9	4	6	6	4	4
Number of flags	28	25	30	35	12	10
Sum of all changes	12.997728	1.157332	9.080144	825626	4.316568	6.096551
Sum of changes found by HB	12.958562	1.029825	8.701388	485480	4.209380	5.856540
Relative to the sum of all changes, %	99,7	88,98	95,82	58,80	97,52	96,06

REMARK

Later, another study designed almost exactly as the one reported has been carried out by Statistics Sweden on data of the SEW. Then the method was compared with the current editing procedure (a traditional one) and with the Aggregate Method which not yet is in use for that survey. The HB-Method was found to be superior to the current procedure and equal to or better than the Aggregate Method.

10 — THE BOX-PLOT METHOD

10.1 – References

The concept of Box-Plots was introduced by Tukey (1977). The Box-Plot Method as a micro-editing method is reported in Anderson (1989 b). It is suggested as a macro-editing method in the discussion on the Aggregate method.

Anderson (1989 a) reports an experiment carried out on data from the Australian Bureau of Statistics (ABS) survey Average Weekly Earnings (AWE). ABS extended the bounds of every ratio check used in that survey to (Q1-3*IQR, Q3+3*IQR), where Q1, Q3 and IQR are respectively the first and the third quartile and the interquartile range. The study indicates that 75 per cent of the resources for the manual reviewing of flagged data from the whole editing process could be saved by limiting the manual review to "extreme outliers". Anderson used the same method for evaluation as in the studies related above. The remaining errors in data had no significance at all on the estimates.

In the latter version of the report, Anderson suggests that the lower and upper bounds of the checks should be set on the basis of a manual analysis of box-plots of the check functions. Then the bounds can be modified by taking into account outliers near the bounds for extreme outliers. Above all, the survey staff will get full control of responsibility for the data editing.

10.2 – Description of the method

The distributions of the check functions of the weighted keyed-in values are displayed as box-plots. Then the acceptance intervals for the checks are set by the staff on the basis of these graphs and put into the regular error-detecting program. By Anderson (1989 a) and the studies on the Aggregate Method reported in this paper, the definitions of extreme outliers may serve as guide-lines for efficient limits. The only difference to the method reported in Anderson (1989 b) is that the values should be weighted by (approximately) the inflation factor (according to the sample design).

11 — THE BOX METHOD

11.1 – Introduction

The Box Method is a graphical macro-editing method under development at Statistics Sweden. A first version of a prototype for the Survey of Employment and Wages is expected by November 1991.

The basic principles are to utilize computer graphics to visualize the distribution of the check function of the weighted data and the inter-activity of a computer to get indications when to stop the manual verifying work. The method may be considered as a combination of a generalized Box-Plot method and the Top-Down Method.

11.2 – Description of the method

The keyed-in data are weighted and then put into the check function. Any mathematical expression may be used as a check function. The values of the function are plotted on the screen and acceptance regions of any shape can also be provided. The reviewer draws a box around the observations he wants to review. On the screen, the data then appear on in advance selected items of the records belonging to the data points inside the box. For every check function the user can select the items of the records to be displayed. A change is entered inter-actively and some data (statistics) of the impact of the change will be displayed.

The method may also be used as a tool to find appropriate values of acceptance regions for other editing methods (e.g. the HB-Method).

12 — SUMMARY

12.1 – Description of the macro-editing methods

The basic principle of all the reported macro-editing methods is that the acceptance regions are determined solely by the distributions of the received observations of the check functions. The keyed-in values of the variable to be checked are weighted by the inflation factor before it is put into the check function.

In the Aggregate Box-Plot, HB and Top-Down methods all values of the check function are sorted by size.

The tails of the distributions are displayed and analyzed in the Aggregate and

the Box-Plot methods in order to set the acceptance bounds for the checks. In the HB-Method this setting of the acceptance limits is done automatically.

In the Top-Down and the Box methods the effects of the detected errors on the estimates "determine" how far the manual review should go. In the Top-Down Method the manual review work starts with the extreme values and goes towards the median value, while in the Box Method the records for manual reviewing are selected by the reviewer from a graphical display of the values of the check functions. This selection may be supported by guide-lines (acceptance regions) displayed in the same graph.

The choice of method should be based on the number of variables to be edited by the macro-editing method and on how the staff wishes to work.

12.2 – Macro-editing versus micro-editing methods

Macro-editing is not a new concept. It has always been used in data editing, but only as a final checking. Another term is out-put checking.

What is new is that such out-put check methods can be used in data editing procedures in the same way as micro-editing methods and that they have proved to be much more efficient than traditionally applied micro-editing procedures.

Macro-editing methods of the type described in this paper may be considered as a statistical way of providing micro-editing checks with efficient acceptance limits. The limits are based only on the data to be edited. The methods bring a kind of priority thinking to the verifying work. Data are edited according to their impact on the estimates. The macro-editing methods solve the general problem inherent in micro-editing methods, i.e. that they produce too many error signals without giving any guidance as to how the resources of the verifying work are to be allocated. We have seen that with micro-editing procedures even very large errors are not always detected, due to the large number of flagged data.

All published studies on the impact of editing show that only a few of the detected errors influence the estimates. If this is accepted the macro-editing methods offer a possibility to reduce the reviewing resources to a considerable extent. Here reported studies reduce the work by 35 - 80 per cent. However, there are no limits concerning the number of cases to be selected for a manual review. The reviewer can select all the cases he deems necessary. The difference to micro-editing methods is that the cases are selected in a priority order, i.e. according to the impact they may have on the estimates. The selection is mainly done by the reviewer, which means that he governs and has the full responsibility for the whole scrutinizing work. In micro-editing procedures, the user is dominated by the computer and cannot see the effects of his work.

Both procedures focus on randomly appearing negligence errors and utilize the

same principle to point out outliers or extreme observations. Micro-editing procedures flag data by criteria fixed in advance, based on historical data, while macro-editing procedures focus on data, which at that very moment and relative to the estimates are the most extreme ones.

Systematic errors, e.g. that many respondents misunderstand a question in the same way or deliberately give wrong answers cannot (in principle) be detected by either macro-editing or micro-editing methods.

We do not know whether any essential improvement is gained by either of the two procedures!

But macro-editing certainly is a more efficient way of reaching the same "quality" standard and may release resources for editing the misunderstanding errors.

BIBLIOGRAPHY

ANDERSON, K. (1989): Draft, Output Edit Study, Average Weekly Earnings. c/- Statistical Services Branch, Australian Bureau of Statistics, September 1989.

ANDERSON, K. (1989 b): "Enhancing Clerical Cost-Effectiveness in the Average Weekly Earnings", Draft, Australian Bureau of Statistics, Statistical Services Branch, 9 November 1989.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., VERMEULEN, L.F.M. (1989): BLAISE 2.0 An introduction. Central Bureau of Statistics, The Netherlands, February 1989.

BETHLEHEM, J.G., DENTENEER, D., HUNDEPOOL, A.J., and SCHUERHOFF, M.H. (1988): Automating the Data Editing Process with the BLAISE System. Seminar on Statistical Methodology, Geneve 1-4 February 1988.

COCHRAN, W. G. (1963): "Sampling Techniques", Second Edition 1963.

FERGUSON, DANIA P. (1989): Review of Methods and Software Used in Data Editing. SCP2/DE/WP.33 (U.S. Department of Agriculture, National Agricultural Statistics Service), October 1989.

GARCIA-RUBIO, E., PEIRATS, V.(1989): Evaluation of Data Editing Procedures, SCP II/DE/WP.28.

GRANQUIST, L. (1984:a): On the Role of Editing, Statistisk Tidskrift 1984:2.

GRANQUIST, L. (1984:b): Data Editing and Its Impact on the Further Processing of Statistical Data. Workshop on the Statistical Computing Project, Budapest, 12-17 November 1984, Invited paper.

GRANQUIST, L. (1987): Macro-Editing - The Top-Down Method, Statistics Sweden, Report 1987-04-09.

GRANQUIST, L.(1988:a): On the Need for Generalized Numeric and Imputation Systems. The Seminar on Statistical Methodology, Geneva, February 1-4, 1988.

GRANQUIST, L. (1988:b): Macro-Editing - The Aggregate Method, Statistics Sweden, Report 1988-08-18.

GREENBERG, B., PETKUNAS, T. (1987): An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division, 1982 Economic Censuses and Census of Governments, Evaluation Studies.

HIDIROGLOU, M.A. and BERTHELOT, J.-M. (1986): Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, June 1986, Vol12. No 1. pp 73-83

HILL, Ch.J., (1978): A Report on the Application of a Systematic Method of Automatic Edit and Imputations to the 1976 Canadian Census. ASA meeting, August 1978, San Diego.

HÖGLUND DAVILA, E. (1989): Macro-editing - The Hidiroglou-Berthelot Method (Statistical Edits), Statistics Sweden, Report 1989-03-28.

LINACRE, S.J. and TREWIN, D.J.: Evaluation of Errors and Appropriate Resource Allocation in Economic Collections. Undated. Internal paper. Australian Bureau of Statistics.

LINDBLOM, A.: A review of the macro-editing procedure Top-Down, Data Editing Joint Group Product No SCP2/D.12/f, June 1990.

LINDSTRÖM, K.: A macro-editing method application developed in PC-SAS, Data Editing Joint Group Product No SCP2/D.11/f, May 1990.

MAZUR, C., (1990): "A Statistical Edit for Livestock Slaughter Data", SRB Research Report Number SRB-90-01 USDA-NASS, Washington DC 202 50, October 1990.

MEGILL, DAVID and ROWLAND, SANDRA (1987): Data Edit Error Analysis Proceedings of the American Statistical Association Section on Survey Research Methods, 1987 pp 154-159.

PIERZCHALA, M. (1988): A Review of the State of the Art in Automated Data Editing and Imputation. Staff report for National Agricultural Statistics Service, U.S.D.A. No. SRB-88-10.

PONS ORDINAS, JUAN (1988): Proceso de Macroedición, Analisis y Transferencias, Macro-Micro en la Encuesta Nacional. Desagregación en Cascada de Tablas de Series, Instituto Nacional de Estadistica, España, Documento de Trabajo, Diciembre 1988.

PRITZKER, L., OGUS, J. and HANSEN, M.H., (1965): Computer Editing Methods - Some Applications and Results, Bullentin of the International Statistical Institute, Belgrade.

PULLUM, T.W., HARPHAM, T. & OZSEVER, N., (1986): The Machine Editing of Large Sample Surveys: The Experience of the World Fertility Survey, International Statistical Review, Volume 54, Number 3, December 1986.

SUBCOMMITTEE on Data Editing in Federal Statistical Agencies, (1990): "Data Editing in Federal Statistical Agencies" Statistical Policy Office, Working Paper 18, May 1990.

TUKEY, J.W. (1977): "Exploratory Data Analysis", Addison-Wesley Publishing Company.

UNDP-ECE (1982): Glossary of Terms on the Statistical Computing Project, Statistical Computing Project.

WARWICK, D.P. and LININGER, C.A. (1975): The Sample Survey: The Theory and Practice, Mc Graw-Hill Book Company, New York.

WERKING G., TUPEK A. and CLAYTON R., (1988): CATI and Touchtone Self-Response Applications for Establishment Surveys, Journal of Official Statistics, 1988-4.

ABSTRACT

This paper deals with data editing and quality. First, it presents the background that justifies the need for more efficient error detection methods. Then, "macro-editing methods" are introduced as a solution for the problem of lack of efficiency of traditional "micro-editing methods". Some experimental studies described here have shown that macro-editing methods are superior to traditional micro-editing methods in editing quantitative data.

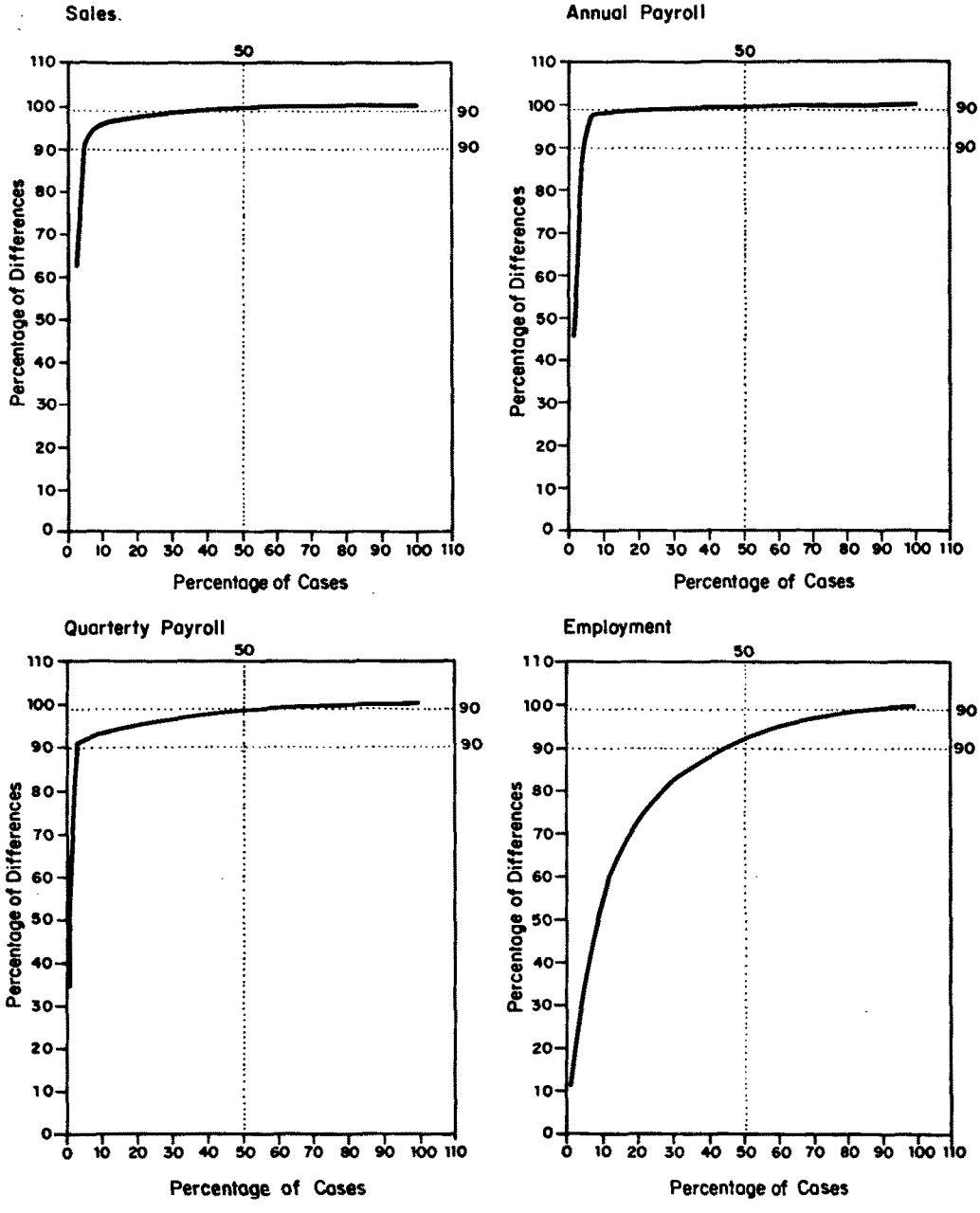
RESUMO

Este artigo trata da depuração e qualidade de dados. Primeiro, apresenta os fundamentos que justificam a necessidade de métodos mais eficientes para a detecção de erros. Em seguida, métodos denominados *macro-editing* são introduzidos como solução para o problema da falta de eficiência dos métodos tradicionais de detecção de erros (*micro-editing*). Alguns resultados de estudos experimentais, descritos aqui, mostraram que os métodos *macro-editing* são melhores que os métodos tradicionais de detecção (*micro-editing*) para dados quantitativos.

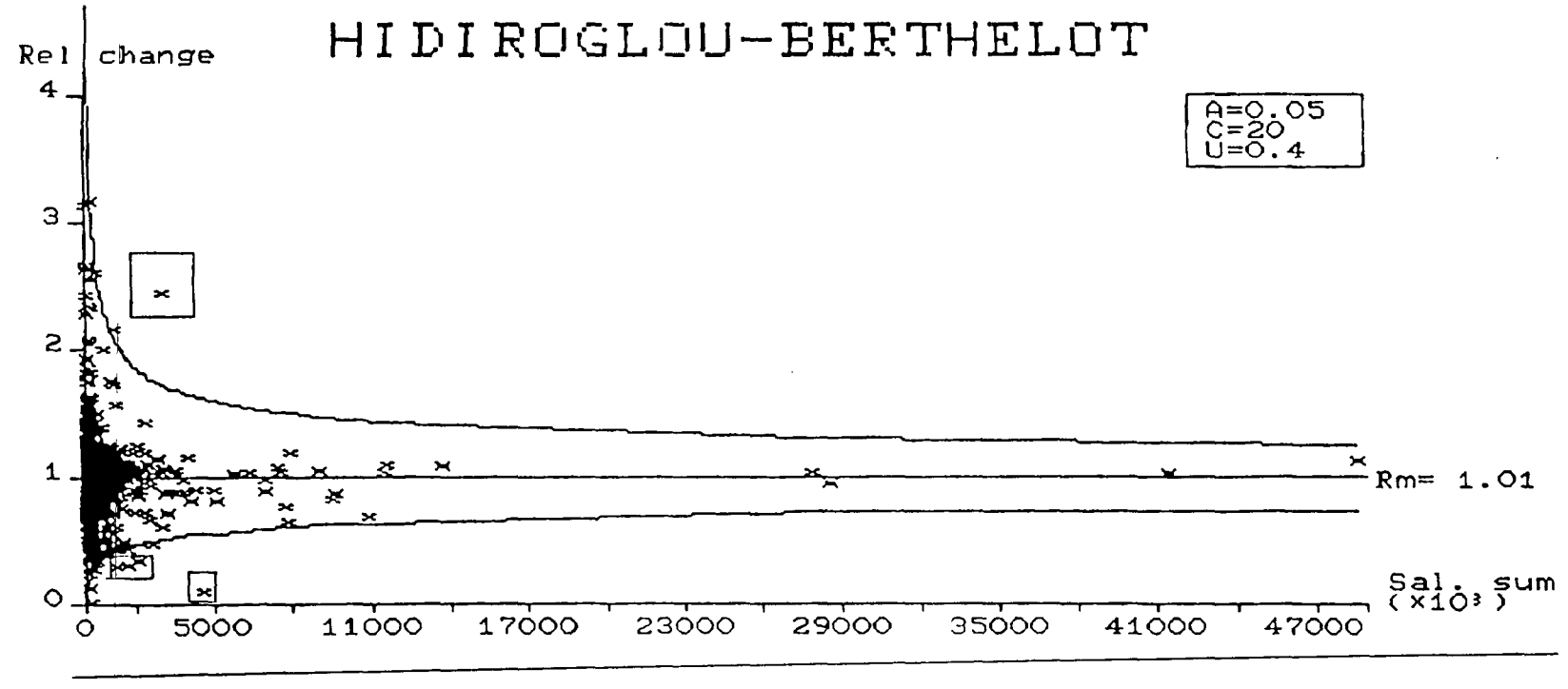
APPENDIX

Reprinted from Greenberg and Petkunas (1987)

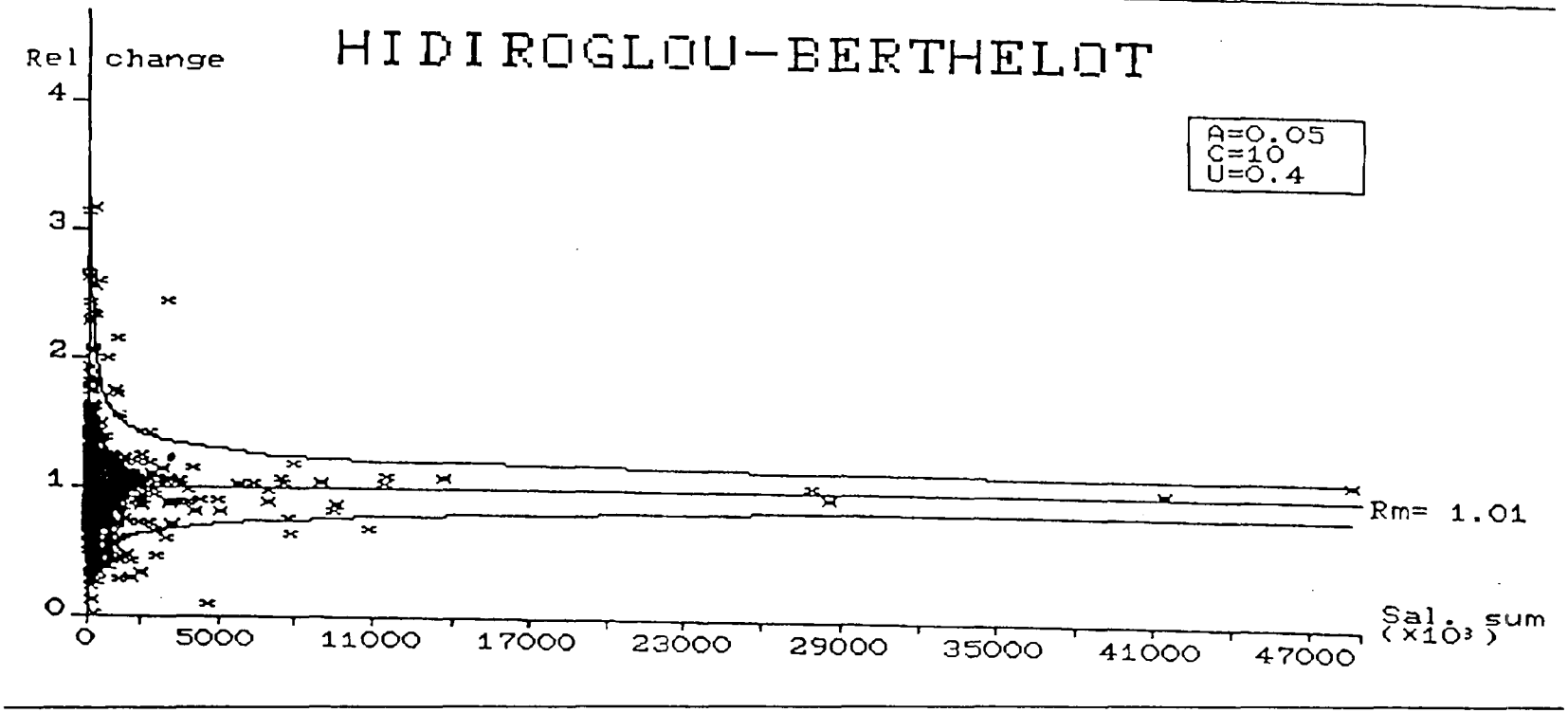
Figure 1. Graphs of Percentage of Total Change by Percentage of Cases SERVICE SIC 783300



STATISTICS SWEDEN
GRAPHICAL EDITING TEST
1989



STATISTICS SWEDEN
GRAPHICAL EDITING TEST
1989



AVALIAÇÃO DOS EFEITOS DE REDUÇÃO DA FRAÇÃO DE AMOSTRAGEM NO CENSO DEMOGRÁFICO

José Carlos da Rocha C. Pinheiro*
e
José Matias de Lima**

1 — INTRODUÇÃO

Este documento pretende fornecer subsídios para o enriquecimento das discussões referentes aos estudos de redução da amostra do Censo Demográfico de 1990.

O principal objetivo deste estudo consiste em avaliar e analisar a perda de eficiência das estimativas de total de características de domicílios obtidas com a utilização de uma fração de amostragem inferior à adotada no Censo de 1980 [1] e, também, avaliar as frações amostrais necessárias para atender um determinado nível de precisão fixado *a priori.*

Costa, L.N. & Lima, J.M. [2] constataram, através de amostras com fração de 1/10, simuladas a partir da amostra de 25% do Censo Demográfico de 1980 e considerando três municípios cuja população não excedia 20 000 habitantes, que a amostra de 10% fornece estimativas de características de domicílios, a nível de município, sem grande perda de precisão, comparado com aquelas obtidas na amostra de 25% do Censo Demográfico de 1980.

*Professor da ENCE/IBGE.
**Analista Especializado da ENCE/IBGE.

Neste trabalho, foram considerados 48 municípios abrangendo todas as regiões brasileiras com contingentes populacionais variando de 1 966 a 1 307 608 habitantes. A exemplo de Costa, L.N. & Lima, J.M. [2], foram utilizadas no estudo as variáveis referentes às características de domicílios do plano tabular do Censo Demográfico de 1980 que tiveram suas estimativas divulgadas a nível de município. A relação dos municípios, bem como das variáveis consideradas, encontra-se nos Anexos 1 e 2, respectivamente.

É desenvolvida neste trabalho uma medida de perda de eficiência da precisão das estimativas associada à redução da fração de amostragem. Esta medida é definida a partir das perdas de eficiência individuais do elenco de variáveis considerado, calculadas através da precisão das estimativas.

São esperadas maiores perdas de eficiência e também frações amostrais elevadas para atender determinados níveis de precisão previamente estipulados, quando considerados municípios com baixo contingente populacional.

Na segunda seção, apresenta-se a metodologia utilizada no trabalho. Os métodos ali apresentados são descritos, propositalmente, de forma teórica, sem restrições a este estudo particular, de modo a permitir sua utilização e/ou adaptação a outros estudos de natureza semelhante.

Na terceira seção, são apresentados e analisados os resultados da aplicação da metodologia ao conjunto de municípios considerados. A análise se desenvolve através de gráficos com o intuito de facilitar o seu entendimento.

Finalmente, são apresentadas algumas conclusões sobre os resultados obtidos (bem como sugestões).

2 — METODOLOGIA

O objetivo principal deste trabalho, como já mencionado anteriormente, é o de avaliar os efeitos da redução da fração amostral utilizada no Censo Demográfico do IBGE sobre a precisão dos estimadores das características investigadas a partir da amostra. Ou seja, pretende-se analisar até que ponto a perda de eficiência representada pela redução da fração amostral poderá comprometer a qualidade das estimativas a serem divulgadas.

a) Medidas de Eficiência

A medida clássica de eficiência relativa entre estimadores é definida a partir do quociente entre os erros quadrados médios dos mesmos. Assim, se θ é um parâmetro de interesse e $\hat{\theta}_1$ e $\hat{\theta}_2$ dois estimadores alternativos, a eficiência de $\hat{\theta}_1$ com relação a $\hat{\theta}_2$

é definida como segue (ver [3], por exemplo):

$$e_\theta(\hat{\theta}_1, \hat{\theta}_2) = \frac{EQM_\theta(\hat{\theta}_2)}{EQM_\theta(\hat{\theta}_1)} \tag{2.1}$$

Neste trabalho, os estimadores considerados diferem entre si apenas pela fração amostral utilizada no seu cálculo. Pode-se, então, usar a seguinte notação, para caracterizar a dependência com relação à fração amostral (f):

$$\hat{\theta} = \hat{\theta}(f) \tag{2.2}$$

De maneira análoga à anterior, definimos a eficiência entre duas frações amostrais f_1 e f_2, na estimação de um parâmetro de interesse θ, da seguinte forma:

$$e_\theta(f_1, f_2) = \frac{EQM_\theta(\hat{\theta}(f_2))}{EQM_\theta(\hat{\theta}(f_1))} \tag{2.3}$$

Como o parâmetro de interesse θ é o mesmo nas duas frações consideradas, pode-se, na ausência de vício dos estimadores, e supondo $\theta \neq 0$, reescrever a fórmula anteriormente descrita em termos dos coeficientes de variação (CV) dos estimadores envolvidos:

$$e_\theta(f_1, f_2) = \left(\frac{CV(\hat{\theta}(f_2))}{CV(\hat{\theta}(f_1))}\right)^2 \tag{2.4}$$

Ocorre que esta medida clássica não é interessante para comparar frações amostrais, já que:

(i) ela mede a eficiência com relação a uma única característica (quando o que se deseja é avaliar o impacto de redução para o conjunto de variáveis investigadas); e
(ii) ela não leva em conta a precisão absoluta dos estimadores (um aumento do CV de 0,5% para 1% é, por esta medida, considerado tão importante quanto o acréscimo de 5% para 10%, ou 10% para 20%, enquanto que os impactos sobre a qualidade das estimativas são certamente distintos, nos três casos mencionados).

A fim de ilustrar os pontos mencionados no parágrafo anterior, considere-se, por exemplo, um esquema de Amostragem Aleatória Simples sem Reposição–AASSR. Na estimação de médias ou totais populacionais, tem-se (ver [4], por exemplo):

$$e_\theta(f_1, f_2) = \frac{N^2 S^2 (1 - f_2)/n_2}{N^2 S^2 (1 - f_1)/n_1} = \frac{f_1(1 - f_2)}{f_2(1 - f_1)} \tag{2.5}$$

onde:

N = tamanho da população;
S^2 = variância da característica investigada, na população; e

n_i = tamanho da amostra associado à fração f_i.

Observe-se, assim, que a eficiência independe de quaisquer parâmetros da população, sendo a mesma para qualquer característica investigada e para qualquer tamanho de população.

Considere-se, por exemplo, as frações de 25% e 10%. Obtém-se, neste caso:

$$e_\theta(0,1;0,25) = \frac{0,1(0,75)}{0,9(0,25)} = 1/3$$

ou seja, há uma perda de eficiência da ordem de 67% associada à redução da fração amostral de 25% para 10%. Observe-se, ainda, que substituindo o valor encontrado na fórmula (2.4), obtém-se:

$$CV(\hat{\theta}(0,1)) = \sqrt{3} \quad CV(\hat{\theta}(0,25))$$

Note-se que esta relação é válida apenas na suposição de AASSR e na estimação de médias ou totais populacionais.

As deficiências da medida clássica de eficiência, abordadas nos parágrafos precedentes, motivaram o desenvolvimento de outra medida mais apropriada aos propósitos deste trabalho. O objetivo principal foi o de obter uma medida que:

(i) levasse em conta o conjunto de variáveis investigadas; e

(ii) dependesse, em certo sentido, do valor absoluto das precisões envolvidas.

Com relação a **(ii)**, o que se pretende é que a medida trate de forma diferenciada aumentos ou decréscimos de variabilidade, de acordo com a magnitude dos valores envolvidos. Assim, na mesma linha de raciocínio do exemplo anteriormente citado, a passagem de um CV de 0,5% para 1% não deveria ser considerada como uma perda de eficiência significativa, já que, em ambos os casos, a precisão do estimadores é mais do que razoável para a maioria das aplicações práticas. De forma equivalente, a mudança do CV de 80% para 160% também não deveria ser tratada como uma perda substancial de eficiência, já que a precisão é bastante ruim nos dois casos. Restam, portanto, as situações intermediárias. Por exemplo, a passagem do CV de um estimador de 7% para 14%, com a mudança da fração amostral, poderia ser considerada uma perda significativa de eficiência, caso o limite de 10% fosse estabelecido como o máximo tolerável para a precisão dos estimadores.

Com base nas características mencionadas anteriormente, passamos, agora, à definição da medida de perda de eficiência utilizada no trabalho. Suponha que existam r características sendo investigadas e que foi estabelecido um limite máximo c, a partir do qual o CV associado ao estimador de uma característica de interesse é considerado inaceitável, em termos da precisão requerida para os resultados a serem divulgados. Seja

$$I_c(\hat{\theta}_i) = \begin{cases} 1, & \text{se } CV(\hat{\theta}_i) > c \\ 0, & \text{se } CV(\hat{\theta}_i) \leq c \end{cases}, \qquad i = 1, \ldots, r \tag{2.6}$$

Define-se a medida de *perda de eficiência relativa* entre as frações f_1 e f_2 como se segue:

$$P_c(f_1, f_2) = \sum_{i=1}^{r} |I_c(\hat{\theta}_i(f_1)) - I_c(\hat{\theta}_i(f_2))|/r \tag{2.7}$$

Note-se que, se com ambas as frações o estimador apresentar CV acima ou abaixo do limite c, a sua contribuição para a medida de perda de eficiência será nula. Caso haja a transposição do limite c, quando da mudança da maior para a menor fração, o numerador da expressão anterior será acrescido de uma unidade. Observe-se, ainda, que a medida é simétrica com relação às frações f_1 e f_2, ou seja, a eficiência de f_1 com relação à f_2 é idêntica à de f_2 com relação à f_1. A fim de evitar confusão, a perda de eficiência deve ser sempre considerada como da fração menor com relação à fração maior. O problema poderia ser contornado eliminando-se o módulo no numerador da expressão (2.7) e trabalhando-se com medidas negativas de eficiência. De modo a manter intacta a intuição por trás da interpretação dos valores da medida considerada, optou-se pela definição apresentada.

O limite de tolerância c, cuja definição é central para o cálculo de $P_c(f_1, f_2)$, é estabelecido de forma até certo ponto subjetiva, sendo difícil chegar a um consenso sobre o valor mais adequado. Este problema pode ser contornado considerando-se a medida como uma função do limite de tolerância c:

$$P(c, f_1, f_2) = P_c(f_1, f_2)$$

e obtendo-se o valor dessa função para diversos valores de c, dentro de uma faixa de variação que cubra os valores considerados razoáveis, em termos de aplicações práticas.

Observa-se que, pela definição da função $P(c, f_1, f_2)$, tem-se:

$$\begin{aligned} &\text{(i)} && 0 \leq P(c, f_1, f_2) \leq 1, && c \in [0, +\infty) \\ &\text{(ii)} && P(0, f_1, f_2) = 0, && f_1, f_2 \in [0, 1) \\ &\text{(iii)} && \lim_{c \to \infty} P(c, f_1, f_2) = 0 \end{aligned} \tag{2.8}$$

Torna-se interessante, neste ponto, introduzir o conceito de *perda de eficiência global*, que é associada a uma única fração amostral

$$P_c(f) = P(c, f, 1) = \sum_{i=1}^{r} I_c(\hat{\theta}_i(f))/r \tag{2.9}$$

Note-se que a eficiência relativa entre duas frações pode ser descrita a partir das eficiências globais, como na expressão seguinte:

$$P(c, f_1, f_2) = [P_c(f_1) - P_c(f_2)]$$

O desconhecimento dos verdadeiros coeficientes de variação faz com que, na prática, sejam utilizadas estimativas em substituição aos CVs teóricos nas fórmulas (2.6) e (2.7). Ocorre que, eventualmente, não é possível estimar o CV de um determinado estimador. Isso ocorre, por exemplo, quando nenhum indivíduo selecionado para a amostra apresenta uma determinada característica sob investigação. Duas variantes da medida anteriormente apresentada são consideradas, levando em conta este aspecto da estimação:

(i) Perda Corrigida – obtida da mesma maneira que a descrita em (2.7), com r substituído por

r' =número de características para as quais é possível obter estimativas do erro amostral; e

(ii) Perda Absoluta – obtida de forma análoga à descrita em (2.7), fazendo-se

$$[I_c(\hat{\theta}_i(f_1)) - I_c(\hat{\theta}_i(f_2))] = 0 \qquad \forall c \in [0, +\infty)$$

para aquelas características para as quais não é possível estimar o erro amostral, ou seja, estas características não contribuem para a perda de eficiência. A lógica por trás desta variante de $P_c(f_1, f_2)$ é que características para as quais não se pode estimar o CV do estimador são, usualmente, muito rarefeitas na população (e neste caso o CV é muito elevado).

As duas medidas são utilizadas na apresentação dos resultados práticos obtidos com a aplicação da metodologia, incluídos na seção seguinte.

b) Frações Amostrais Necessárias

Com o objetivo de enriquecer a análise sobre a questão da mudança da fração amostral, incluiu-se, neste trabalho, uma parte dedicada ao estudo da fração necessária para garantir um nível prefixado de precisão na estimação das características de interesse.

A definição da fração necessária pressupõe a especificação de um limite máximo de tolerância para o CV de um estimador (de forma análoga à da medida $P_c(f_1, f_2)$) e de um percentual mínimo de características que devem ser estimadas obedecendo o limite de tolerância prefixado para a precisão.

Define-se a fração amostral necessária para um limite de tolerância c e um percentual de variáveis de interesse α da seguinte forma:

$$f(c, \alpha) = \inf\{f | P_c(f) \leq 1 - \alpha\} \tag{2.10}$$

Para o cálculo da fração necessária, foi preciso restringir-se a um método de seleção (AASSR) e fixar uma fração de referência, f_o. Observe-se que, quando o método utilizado é AASSR, a fórmula (2.4) é válida e obtém-se:

$$CV(\hat{\theta}(f)) = \left(\frac{f_o(1-f)}{f(1-f_o)}\right)^{1/2} CV(\hat{\theta}(f_o)) \tag{2.11}$$

Sejam

$$CV(\hat{\theta}(f))_{(1)}, CV(\hat{\theta}(f))_{(2)}, CV(\hat{\theta}(f))_{(3)}, CV(\hat{\theta}(f))_{(4)}, \ldots, CV(\hat{\theta}(f))_{(r)}$$

os coeficientes de variação dos estimadores das r características de interesse, assumindo-se a fração amostral f, ordenados de forma crescente. Sob AASSR, a ordenação dos CVs não é alterada pela mudança da fração amostral e obtém-se:

$$CV(\hat{\theta}(f))_{(k)} = \left(\frac{f_o(1-f)}{f(1-f_o)}\right)^{1/2} CV(\hat{\theta}(f_o))_{(k)}, \quad k = 1, \ldots, r$$

Desta maneira, para assegurar que um percentual mínimo a $\alpha\times$ 100% das características de interesse serão estimadas com CV abaixo do limite c é suficiente garantir que:

$$CV(\hat{\theta}(f_o))_{(t)} \leq \left(\frac{f(1-f_o)}{f_o(1-f)}\right)^{1/2} c \tag{2.12}$$

onde

$$t = \begin{cases} r.\alpha, & \text{se} \quad r.\alpha \in \mathbf{N} \\ [r.\alpha] + 1, & \text{se} \quad r.\alpha \notin \mathbf{N} \end{cases}$$

com $[r.\alpha]$ representando a parte inteira de $r.\alpha$.
Fazendo $C_o(\alpha) = CV(\hat{\theta}(f_o))_{(t)}$, obtém-se, sob AASSR:

$$\{f \mid P_c(f) \leq 1-\alpha\} = \left\{f \mid f \geq \frac{f_o C_o^2(\alpha)}{f_o C_o^2(\alpha) + c^2(1-f_o)}\right\} \tag{2.13}$$

Como a fração amostral necessária $f(c,\alpha)$ é o ínfimo deste conjunto, temos:

$$f(c,\alpha) = \frac{f_o C_o^2(\alpha)}{f_o C_o^2(\alpha) + c^2(1-f_o)} = \left(\frac{c^2}{C_o^2(\alpha)} \frac{(1-f_o)}{f_o} + 1\right)^{-1} \tag{2.14}$$

Observe-se que:

$$\begin{aligned} &\text{(i)} && f(0,\alpha) = 1, \quad \forall \alpha \in [0,1) \\ &\text{(ii)} && \lim_{c \to +\infty} f(c,\alpha) = 0, \quad \forall \alpha \in [0,1) \\ &\text{(iii)} && \frac{\partial f}{\partial c}(c,\alpha) = -\frac{(1-f_o)}{C_o^2(\alpha) f_o} \left(\frac{c}{C_o^2(\alpha)} \frac{(1-f_o)}{f_o} + 1\right)^{-2} < 0 \end{aligned} \tag{2.15}$$

Tomando-se, por exemplo, $f_o = 0,25$, obtém-se:

$$f(c,\alpha) = \left(\frac{3.c^2}{C_o^2(\alpha)} + 1\right)^{-1}$$

Na seção seguinte, são apresentados diversos gráficos resultantes da aplicação da metodologia aqui descrita a municípios de diferentes regiões brasileiras, levando-se em conta características de domicílios investigadas no Censo Demográfico de 1980, a nível de município.

3 — ANÁLISE E APRESENTAÇÃO DOS RESULTADOS

Nesta seção é apresentada uma aplicação da metodologia descrita anteriormente, tomando-se por base duas frações amostrais: 10% e 25%.

Os dados para os cálculos foram obtidos a partir das características de domicílios investigadas no Censo Demográfico de 1980 que tiveram suas estimativas divulgadas a nível de município (cerca de 65 variáveis, ao todo). Foram utilizados no estudo 48 municípios com populações variando entre 1 966 e 1 307 608 habitantes, escolhidos de forma a abranger todas as regiões do País.

De modo a facilitar a apresentação dos resultados, os municípios foram agrupados segundo faixas de população apresentadas a seguir.

Tabela 1
Número de Municípios segundo Faixas de População

Faixa de População	Número de Municípios
< 4 000	10
4 000–14 000	16
15 000–30 000	10
60 000–100 000	4
100 000–140 000	4
> 400 000	4
Total	48

Conforme se observa no quadro acima, há uma predominância de municípios pequenos e médios. Esta escolha foi devida à crença, posteriormente confirmada, de que os efeitos da redução da fração amostral deveriam ser mais sentidos nos municípios com menores contingentes populacionais. Buscou-se, desta forma, analisar de maneira mais cuidadosa os impactos da redução da fração amostral nestes municípios.

Os resultados apresentados nesta seção foram inicialmente obtidos para cada município e depois agregados por faixa de população. Em nenhum momento são consideradas agregações de faixas de população, já que, como ficará claro no transcorrer do texto, os valores obtidos guardam uma estreita relação com o tamanho da população subjacente aos cálculos.

Conforme mencionado na descrição da metodologia, as perdas de eficiência e as frações amostrais necessárias são calculadas a partir dos coeficientes de variação amostral (CVs) dos estimadores das características de interesse. Na verdade, como explicitado ao final da seção anterior, são utilizadas estimativas dos verdadeiros CVs.

Estimativas para os CVs referentes à fração de 25% encontram-se disponíveis nas publicações do Censo Demográfico de 1980, já que esta foi a fração amostral adotada

naquela ocasião. A metodologia de cálculo das estimativas dos CVs dos estimadores das características investigadas no CD-80 é apresentada em [1].

Os coeficientes de variação amostral correspondentes à fração de 10% são obtidos a partir dos CVs referentes à fração de 25%, através da fórmula (2.11), da seção anterior. Assumiu-se, neste cálculo, o esquema de AASSR^2. Na verdade, o esquema adotado no CD-80 foi de seleção por amostragem sistemática, com expansão através de um método de pós-estratificação, denominado PIETOM (ver [1]). Os resultados obtidos, contudo, não diferem significativamente daqueles que teriam sido encontrados, caso fosse assumido o esquema de AASSR.

No cálculo dos valores das perdas de eficiência, bem como das frações de amostragem necessárias, foram utilizados os seguintes limites de tolerância c para o erro amostral (em %):

$$1,2,3,4,5,6,7,8,9,10,12,14,\ldots,30,35,40,\ldots,130,140,150,160,170 \text{ e } 180.$$

São apresentados a seguir, sob forma de gráficos, os principais resultados da aplicação, mantendo-se a subdivisão por faixas de população mencionadas anteriormente. Breves comentários dos gráficos apresentados no Anexo 3 são feitos, de modo a auxiliar sua compreensão. Os comentários gerais e as análises comparativas são guardados para a última seção.

– Municípios com população inferior a 4 000 habitantes

O Gráfico 1 do Anexo 3 mostra o comportamento da perda média de eficiência para as estimativas provenientes de uma amostra simulada de 10%, quando comparada com os resultados obtidos com a amostra de 25% do CD-80.

Para os municípios com contingente populacional nesta faixa, a utilização de uma amostra de 10%, em detrimento da de 25%, acarreta uma perda média com valor máximo em torno de 25% (eficiência corrigida) correspondendo às estimativas com erro amostral estimado em torno de 10%. Observa-se que para as estimativas com erro amostral estimado entre 10% e 40% correspondem as maiores perdas de eficiência, sendo que a perda de precisão para as estimativas com erro amostral superior a 40% oscila em torno de 10%, decrescendo exponencialmente para as estimativas com erro amostral correspondente superior a 110%. É interessante ressaltar que, mesmo para as estimativas com erro amostral elevado, por exemplo, CVs superiores a 40%, as maiores perdas de eficiência observadas correspondem, provavelmente, às estimativas de características que são bastante rarefeitas na população considerada. Para esses municípios existe uma diferença da ordem de, aproximadamente, 8% entre a perda média de eficiência corrigida e absoluta.

Observe-se que a perda média máxima de eficiência corrigida é da ordem de 25% aproximadamente enquanto que a absoluta não ultrapassa 17%. Esta diferença está relacionada com o fato de se estar trabalhando com características muito rarefeitas na população considerada.

O Gráfico 2 do Anexo 3 indica a fração de amostragem necessária para se obter com um nível de precisão prefixado estimativas para 60%, 75%, 90% e 100% das características de domicílios que tiveram suas estimativas divulgadas a nível de município no Censo Demográfico de 1980. O coeficiente de variação é utilizado para fixar o nível de precisão desejado *a priori*.

Observe-se que, fixado *a priori* um erro amostral igual ou inferior a 10%, seria necessária, em média, uma amostra de aproximadamente 60% dos domicílios do universo considerado para fornecer estimativas dentro da precisão requerida, para 60% das variáveis consideradas no estudo. Uma amostra de 30% forneceria, para 75% das variáveis, estimativas com erro amostral estimado menor ou igual a 35%.

– Municípios com população entre 4 000 e 14 000 habitantes

A perda média máxima de eficiência das estimativas obtidas para os municípios com contingente populacional nesta faixa é de aproximadamente 25% (eficiência corrigida) e 20% (eficiência absoluta), correspondente às estimativas com erro amostral estimado entre 12 e 18% aproximadamente. Observe-se que, para as estimativas com erro amostral estimado a partir de 18%, a perda média de eficiência (absoluta e corrigida) decresce rapidamente.

O Gráfico 3 do Anexo 3 permite visualisar o comportamento entre as eficiências (corrigida e absoluta). A diferença entre as perdas de eficiência é da ordem de 5% e ambas apresentam comportamento decrescente a partir de um erro amostral estimado em torno de 18%, sendo que a perda da eficiência corrigida se estabiliza em torno de 12% para estimativas com CVs da ordem de 30%, enquanto que a absoluta se estabiliza em torno de 10%. A partir de CVs da ordem de 100% ambas as curvas decaem.

Quanto à fração amostral necessária para se obter, com um nível de precisão prefixado, estimativas para 60%, 75%, 90% e 100% das variáveis investigadas, observa-se no Gráfico 4 do Anexo 3 que, se fixado um erro amostral em 10%, seria necessária, em média, uma amostra de 50% (fração amostral igual a 1/2) para fornecer estimativas com precisão dentro da requerida para 60% das variáveis consideradas. Uma amostra de 30% forneceria em média, para 75% das variáveis, estimativas com erro amostral estimado menor ou igual a 25%.

– Municípios com população entre 15 000 e 30 000 habitantes

Para os municípios com contingentes populacionais nesta faixa, a perda média máxima é da origem de 22% (corrigida) e 20% (absoluta), correspondente às estimativas com erro amostral estimado em torno de 80%. Observe-se que, para estimativas com erro amostral superior a 35% ambas as perdas médias de eficiência são inferiores a 10%, decrescendo rapidamente.

Observa-se no Gráfico 5 do Anexo 3 que a diferença entre a perda média máxima de eficiência absoluta e corrigida é da ordem de 2%, sendo que para estimativas com CVs superiores a 25% as perdas de eficiência apresentam comportamento decrescente,

permanecendo estáveis em torno de 8% para aquelas estimativas com erro amostral estimado entre 45% e 75%. Note que, para municípios com população na faixa considerada, é pequena a diferença existente entre as perdas médias de eficiência.

Fixado um nível de precisão previamente (CV de aproximadamente 10%), observa-se no Gráfico 6 do Anexo 3 que uma amostra com fração 1/3 forneceria estimativas para 60% das variáveis investigadas com a precisão desejada. Uma amostra de 30% forneceria estimativas para 75% das variáveis com erro amostral estimado não superior a 18%.

– Municípios com população entre 60 000 e 100 000 habitantes

A perda média máxima de eficiência observada para os municípios com população nesta faixa não ultrapassa 21% e corresponde às estimativas com erro amostral estimado em torno de 5%. Este valor decresce à medida que se relaxa o nível de precisão, sendo que, para um erro amostral fixado previamente em torno de 20%, a perda média de eficiência já é inferior a 10%, conforme pode ser observado na Figura 7.

Quanto ao comportamento das perdas de eficiência corrigida e absoluta, observa-se, no Gráfico 7 do Anexo 3, comportamento idêntico, sendo que ambas as curvas já se justapõem para os municípios com contingentes populacionais nesta faixa. A partir de um erro amostral da ordem de 15%, a perda de eficiência decresce, já sendo inferior a 10% para estimativas com erro amostral da ordem de 20%.

No Gráfico 8 do Anexo 3 pode-se visualisar o comportamento da fração de amostragem necessária para atender a um nível de precisão predeterminado. Observe que as curvas correspondentes aos percentuais de variáveis consideradas no estudo apresentam comportamento acentuadamente decrescente. As curvas decrescem exponencialmente à medida que o nível de precisão é relaxado. Observe que se fixado um erro amostral em 10% seria necessária, em média, uma amostra de aproximadamente 20% ($f = 1/5$) para fornecer estimativas com nível de precisão requerida para 60% das variáveis consideradas, uma amostra de 40% para estimar 75% das variáveis, e amostras de 75% e 95% para estimar, respectivamente, 90 e 100% das variáveis consideradas.

Se fixada uma fração de 10%, a amostra forneceria, em média, estimativas de 60% das variáveis com erro amostral estimado não superior a 12% e forneceria estimativas para 75% das variáveis consideradas com erro amostral estimado menor ou igual a 20%, aproximadamente.

– Municípios com população entre 100 000 e 140 000 habitantes

Para estes municípios a perda média máxima de eficiência é de aproximadamente 23% e corresponde às estimativas com erro amostral estimado inferior a 5%. A perda média apresenta comportamento decrescente a partir de CVs da ordem de 5%, permanecendo estável em torno de 10% para estimativas com erro amostral entre 10 e 20%, conforme mostra o Gráfico 9 do Anexo 3.

Neste gráfico pode-se observar a inexistência de diferença entre as perdas média de eficiências consideradas. A partir de um erro amostral estimado em torno de 20%, a

perda média máxima de eficiência observada já é inferior a 10% e apresenta comportamento decrescente vindo a permanecer estável em torno de 2% para estimativas com erro amostral estimado entre 85% e 140%.

Quanto à fração de amostragem necessária para atender um nível de precisão previamente determinado, observa-se no Gráfico 10 do Anexo 3 um decrescimento exponencial indicando que, se fixado *a priori* um erro amostral em 10%, seria necessário, em média, uma amostra de aproximadamente 10% ($f = 1/10$) para fornecer estimativas dentro da precisão requerida para 60% das variáveis.

Observe-se que, se fixada uma fração amostral de 10%, a amostra forneceria em média, estimativas para 60% e 75% das variáveis consideradas, com erro amostral estimado menor ou igual a 10% e 20%, respectivamente.

– Municípios com população superior a 400 000 habitantes

A perda média máxima de eficiência observada é de aproximadamente 24% e corresponde às estimativas com erro amostral estimado da ordem de 20%. Observa-se no Gráfico 11 do Anexo 3 que, para *CV*s a partir de 10%, a perda já é inferior a aproximadamente 5% e decai rapidamente à medida que o nível de precisão é relaxado. Não existe diferença entre as perdas de eficiência absoluta e corrigida (as curvas se justapõem). Observe-se ainda que, para estimativas com erro amostral estimado a partir de 30%, a perda de eficiência já é, de modo geral, inferior a 2%.

Quanto à fração de amostragem necessária para atender a um determinado nível de precisão fixado *a priori*, observa-se que, com exceção da curva referente à fração de amostragem necessária para estimar todo o elenco de variáveis consideradas (100%) no estudo, as demais curvas convergem para zero rapidamente.

Conforme pode-se observar no Gráfico 12 do Anexo 3, fixado um erro amostral máximo permitido em 10%, uma amostra de aproximadamente 5% forneceria estimativas para 75% das variáveis consideradas, enquanto que para obter estimativas de 90% e 100% das variáveis seriam necessárias amostras de 20% ($f = 1/5$) e 80%, respectivamente.

4 — CONCLUSÕES

A análise das perdas de eficiência através das diferentes faixas de tamanho do município consideradas no estudo indica que a perda máxima é quase independente do tamanho da população, situando-se sempre em torno de 25%. Tal perda não pode ser considerada como muito expressiva, significando que, na pior situação possível, um quarto das variáveis que eram bem estimadas com a fração de 25% deixaram de sê-lo com a fração de 10%.

Talvez mais importante do que a perda máxima seja o valor do *CV* para o qual

ela ocorre. Analisando-se esta última medida, observa-se uma forte dependência da mesma com relação ao tamanho da população. De um modo geral à medida que aumenta a população, decresce o valor do CV associado à máxima perda de eficiência. Assim, para municípios com população até 4 000 habitantes, a perda máxima ocorre para um CV de 11%, enquanto que para municípios com população acima de 400 000 habitantes a perda máxima se dá para CVs da ordem de 2%. Sem dúvida, isso oferece evidência de que, de um modo geral, a perda de eficiência é mais grave nos municípios pequenos do que nos grandes, o que, aliás, já era esperado.

À medida que aumenta o tamanho da população do município, deixam de existir diferenças entre as perdas de eficiência corrigida e absoluta. Tal fato se explica em razão da menor rarefação de determinadas características em populações maiores. Esse aspecto do comportamento das características investigadas na amostra aponta para a inadequação de um plano tabular único para todos os municípios. Aparentemente seria mais recomendável trabalhar-se com planos tabulares diferenciados, levando-se em conta a precisão com que determinadas variáveis podem ser estimadas, tomando-se por base o tamanho da população do município. Apenas variáveis que pudessem ser estimadas com um mínimo de precisão deveriam ser divulgadas.

Com relação às frações amostrais, o que se observa é um aumento considerável da velocidade de decaimento das curvas de fração amostral necessária com o aumento da população. A análise comparativa destas curvas leva às seguintes conclusões:

(i) – Para municípios pequenos (população menor que 14 000), são necessárias frações amostrais extremamente elevadas (50% ou mais) para assegurar precisão razoável CV (20%) para estimativas de um percentual de 75% das variáveis;

(ii) – Nos municípios grandes (população maior que 400 000), as frações necessárias para níveis razoáveis de precisão (CV de 10%) para a maioria das características investigadas (75%) são bem pequenas (4%); e

(iii) – Para os municípios intermediários (15 000 a 400 000 habitantes), as frações que parecem mais apropriadas oscilam entre 10% e o valor usualmente considerado (25%).

Com base nestas observações, pode-se concluir que:

(i) – Para os municípios pequenos, tanto a fração de 10% como a de 25% são inadequadas;

(ii) – Para os municípios grandes, as frações de 10% e 25% são excessivas; e

(iii) – Para os municípios intermediários, as frações de 10% e 25% parecem apropriadas.

Observa-se então que a adoção de uma fração de amostragem única para todos os municípios não parece a decisão mais acertada (a exemplo do que já havia sido comentado com relação ao plano tabular). A utilização de frações variadas, adequadas aos diferentes tamanhos populacionais, surge como uma alternativa mais atraente.

Inclui-se, a seguir, uma proposta de dimensionamento da amostra com emprego de frações de amostragem mistas, feita a partir das faixas populacionais consideradas no estudo. Tal proposta deve ser encarada mais como um exemplo da solução que poderia vir a ser adotada no censo, do que uma tentativa de estabelecer a composição ótima de frações e tamanhos populacionais.

Os objetivos principais da proposta, que doravante será denominada por *mista*, são o de garantir maior grau de homogeneidade na precisão das estimativas, por faixa de tamanho populacional, e obter redução significativa do tamanho total da amostra, com relação àquele que seria obtido com a fração única de 25%.

A proposta *mista* é descrita na Tabela 2, a seguir:

Tabela 2
Fração de Amostragem segundo Faixas de População

Faixa de População do Município (hab.)	Fração de Amostragem
até 14 000	1/2
14 001 a 30 000	1/5
30 001 a 400 000	1/10
mais de 400 000	1/25

Em termos de precisão, a comparação entre as três propostas de dimensionamento da fração de amostragem pode ser feita com base na Tabela 3 onde são indicados os CVs máximos esperados, por faixa de população, para 75% das características investigadas.

Tabela 3
Coeficientes de Variação Máximos Esperados, por Fração de Amostragem, segundo Faixas de Tamanho dos Municípios

Faixa de População do Município (hab.)	Fração de Amostragem Proposta		
	Mista	10%	25%
até 4 000	22	65	40
4 001 a 14 000	16	55	30
14 001 a 30 000	23	35	20
30 001 a 100 000	21	21	12
100 001 a 400 000	20	20	11
mais de 400 000	10	7	4

Pode-se observar a superioridade da proposta mista nos municípios pequenos, a equivalência entre as propostas mistas e 10% nos municípios intermediários e a primazia da fração de 25% nos municípios grandes.

Verifica-se que a proposta mista mantém a precisão dentro de limites aceitáveis para todas as faixas populacionais consideradas, o que não ocorre para as duas outras propostas.

Nos municípios grandes a proposta mista perde, em termos de precisão, para as frações de 10% e 25%. Entretanto, a precisão obtida com a proposta mista nestes municípios parece ser mais do que razoável para a maioria das aplicações.

Outra comparação interessante entre as três propostas diz respeito ao custo envolvido na sua utilização. Uma análise detalhada dos custos envolvidos certamente foge aos propósitos deste estudo. Pode-se, contudo, efetuar uma análise aproximada, tomando-se por base o número de domicílios selecionados em cada uma das propostas. Neste ponto, abrem-se duas possibilidades: considerar-se todos os domicílios particulares ou apenas os domicílios particulares ocupados. Os valores referentes aos domicílios particulares ocupados são apresentados entre parênteses.

Na Tabela 4 são apresentados os diferentes tamanhos de amostra associados às três propostas, considerando-se o Censo Demográfico de 1980, e utilizando-se tanto os domicílios particulares em geral como apenas os ocupados.

A partir desta tabela pode-se observar que a proposta mista acarreta uma redução de 62% no tamanho da amostra, com relação à proposta de fração de 25%, considerando-se os domicílios particulares em geral (66% com relação aos domicílios particulares ocupados), sendo, por outro lado, 55% maior que a amostra de 10%, relativamente aos domicílios particulares (51% com relação a domicílios particulares ocupados).

A proposta mista é uma solução intermediária entre as propostas de 10% e 25%, em termos de custo, tendo, por outro lado, **um grau de homogeneidade** na precisão das estimativas das características de interesse, superior ao das duas outras propostas.

Uma possível desvantagem associada ao uso da proposta mista seria a coexistência de diferentes frações de amostragem em uma mesma pesquisa. Isso só poderia vir a constituir um problema quando considerada a operacionalização da coleta no campo, já que em termos de estimação a proposta não encerra qualquer dificuldade. Contudo, tendo em vista que se supõe que a fração de amostragem é única para cada município, não se acredita que a proposta mista venha a trazer maiores dificuldades para coleta.

Assim sendo, a proposta mista tem como forte atrativo a garantia de precisão uniforme (dentro de limites aceitáveis) para todas as faixas de população consideradas, além de representar um ganho substancial em termos de recursos, com relação à amostra de 25%. As possíveis dificuldades operacionais que adviriam da sua utilização, que, a um primeiro exame, não parecem muito significativas, não são suficientes para contra-indicar sua adoção. Obviamente, a proposta mista mais adequada aos objetivos do Censo Demográfico de 1990 deve ser estudada em maior detalhe, bem como as implicações dela decorrentes. Este texto visa apenas a apresentar uma proposta, junto com uma análise superficial de suas vantagens, e não esgotar a discussão sobre o tema.

Tabela 4

Faixa de População do Município (hab.)	Tamanho da Amostra Proposta		
	Mista	10%	25%
até 14 000	1 989 545 (1 625 519)	397 909 (325 104)	994 773 (812 760)
14 001 a 30 000	1 025 489 (850 217)	512 745 (425 109)	1 281 861 (1 062 771)
30 001 a 400 000	1 249 406 (1 065 205)	1 249 406 (1 065 205)	3 123 514 (2 663 012)
mais de 400 000	323 167 (288 899)	807 916 (722 246)	2 019 788 (1 805 613)
Total	4 587 607 (3 829 840)	2 967 976 (2 537 664)	7 419 936 (6 344 156)

5 – ANEXOS

ANEXO 1

MUNICÍPIOS CONSIDERADOS NO ESTUDO

MUNICÍPIOS CONSIDERADOS NO ESTUDO, UNIDADE DA FEDERAÇÃO E POPULAÇÃO SEGUNDO O CENSO DEMOGRÁFICO – 1980

Município Considerado	UF	População
Afonso Cunha	MA	3 422
Água Comprida	MG	1 966
Barra de São Miguel	AL	2 328
Bento de Abreu	SP	2 037
Caracol	MS	3 819
Ilha de Paquetá*	RJ	2 252
Mairipotaba	GO	2 670
Nova Araçá	RS	2 252
Vianópolis	GO	3 728
Victor Graeff	RS	4 125
Abadia dos Dourados	MG	8 004
Anaurilandia	MS	7 222
Aparecida do Taboado	MS	14 027
Bernardino de Campos	SP	8 994
Cachoeira do Arari	PA	11 402
Cedral	MA	12 802
Dois Riachos	AL	7 952
Duas Barras	RJ	7 992
Monte Mor	SP	14 020
Petrolina de Goiás	GO	12 153
Pinheiro Machado	RS	14 359
Salinópolis	PA	14 349
Santa Cruz do Arari	PA	4 706

*Região Administrativa do Município do Rio de Janeiro.

Município Considerado	UF	População
Santana do Mundaú	AL	13581
Sumidouro	RJ	11386
Tocantins	MA	7684
Altamira do Maranhão	MA	24722
Bujaru	PA	25992
Delmiro Gouveia	AL	26768
Guaporé	RS	24630
Mariana	MG	29401
Mirandópolis	SP	21522
Naviraí	MS	28567
Piracanjuba	GO	24095
Piraí	RJ	28789
São Mateus do Maranhão	MA	25004
Araras	SP	65010
Bacabal	MA	81361
Cachoeirinha	RS	63196
Caucaia	CE	94106
Americana	SP	121998
Bagé	RS	100133
Caxias	MA	125509
Juazeiro do Norte	CE	135620
Campinas	SP	664566
Fortaleza	CE	1307608
Porto Alegre	RS	1125478
São Luís	MA	449433

ANEXO 2

ELENCO DE VARIÁVEIS CONSIDERADAS NO ESTUDO

ELENCO DE VARIÁVEIS CONSIDERADAS NO ESTUDO

- Domicílios particulares ocupados duráveis, situação urbana
- Domicílios particulares ocupados duráveis, situação rural
- Domicílios particulares ocupados duráveis, total (urbana + rural)
- Domicílios particulares ocupados rústicos, situação urbana
- Domicílios particulares ocupados rústicos, situação rural
- Domicílios particulares ocupados rústicos, total (urbana + rural)
- Domicílios particulares permanentes já pagos
- Domicílios particulares permanentes não pagos
- Domicílios particulares permanentes alugados
- Domicílios particulares permanentes cedidos
- Domicílios particulares permanentes outra condição de ocupação
- Domicílios particulares permanentes alugados por até 1/2 salário mínimo (SM)
- Domicílios particulares permanentes com aluguel mensal maior que 1/2 SM e até 1 SM
- Domicílios particulares permanentes com aluguel mensal maior que 1 até 3 SM
- Domicílios particulares permanentes com aluguel mensal maior que 3 SM
- Domicílios particulares permanentes próprios em aquisição
- Domicílios particulares permanentes próprios com prestação mensal até 1/2 SM
- Domicílios particulares permanentes próprios em aquisição c/ prestação de 1/2 a 1 SM
- Domicílios particulares permanentes próprios em aquisição c/ prestação de 1 a 3 SM
- Domicílios particulares permanentes próprios em aquisição c/ prestação maior que 3 SM
- Domicílios particulares permanentes com canalização interna, rede geral
- Domicílios particulares permanentes com canalização interna, poço ou nascente
- Domicílios particulares permanentes com outra forma de abastecimento de água
- Domicílios particulares permanentes sem canalização interna, rede geral
- Domicílios particulares permanentes sem canalização interna, poço ou nascente
- Domicílios particulares permanentes sem canalização interna, outra forma de abastecimento de água
- Domicílios particulares permanentes com abastecimento de água ignorado
- Domicílios particulares permanentes com instalação sanitária só do domicílio, rede geral
- Domicílios particulares permanentes com instalação sanitária só do domicílio, fossa séptica
- Domicílios particulares permanentes com instalação sanitária só do domicílio, fossa rudimentar

- Domicílios particulares permanentes com instalação sanitária só do domicílio, outra forma de instalação
- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, rede geral
- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, fossa séptica
- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, fossa rudimentar
- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, outra forma de instalação
- Domicílios particulares permanentes sem instalação sanitária
- Domicílios particulares permanentes com fogão
- Domicílios particulares permanentes com fogão improvisado
- Domicílios particulares permanentes com fogareiro
- Domicílios particulares permanentes que não têm equipamento para cozinha
- Domicílios particulares permanentes com equipamento ignorado
- Domicílios particulares permanentes que usam gás canalizado para cozinhar
- Domicílios particulares permanentes que usam lenha para cozinhar
- Domicílios particulares permanentes que usam carvão para cozinhar
- Domicílios particulares permanentes que usam óleo ou querosene para cozinhar
- Domicílios particulares permanentes que usam álcool para cozinhar
- Domicílios particulares permanentes que usam eletricidade para cozinhar
- Domicílios particulares permanentes que não têm combustível para cozinhar
- Domicílios particulares permanentes com iluminação elétrica, com medidor
- Domicílios particulares permanentes com iluminação elétrica, sem medidor
- Domicílios particulares permanentes com telefone
- Domicílios particulares permanentes com rádio
- Domicílios particulares permanentes com geladeira
- Domicílios particulares permanentes com televisão em cores
- Domicílios particulares permanentes com televisão em cores e em preto e branco
- Domicílios particulares permanentes com televisão em preto e branco
- Domicílios particulares permanentes com automóvel para uso particular
- Domicílios particulares permanentes com automóvel para trabalho
- Domicílios particulares permanentes com tempo de residência inferior a 1 ano
- Domicílios particulares permanentes com tempo de residência no domicílio de 1 ano
- Domicílios particulares permanentes com tempo de residência no domicílio de 2 anos
- Domicílios particulares permanentes com tempo de residência no domicílio de 3 a 6 anos

- Domicílios particulares permanentes com tempo de residência no domicílio de 7 a 10 anos
- Domicílios particulares permanentes com tempo de residência no domicílio de 11 anos ou mais

ANEXO 3

GRÁFICOS

GRÁFICO 1
PERDAS DE EFICIÊNCIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO < 4 MIL HABITANTES

GRÁFICO 2
FRAÇÃO AMOSTRAL NECESSÁRIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO < 4 MIL HABITANTES

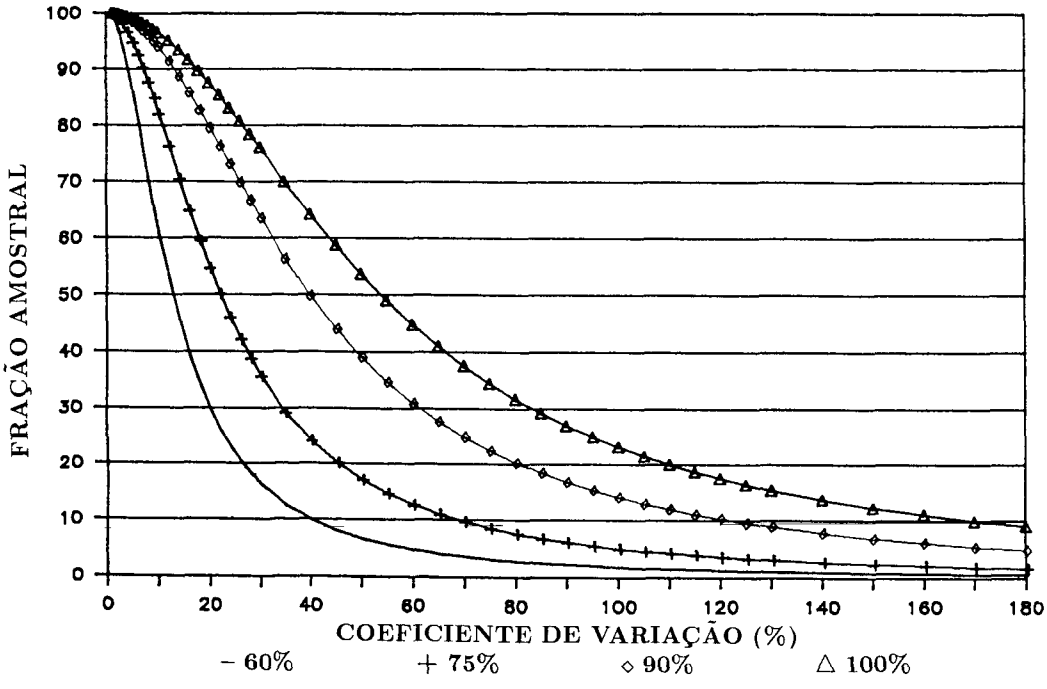


GRÁFICO 3
PERDAS DE EFICIÊNCIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 4 A 14 MIL HABITANTES

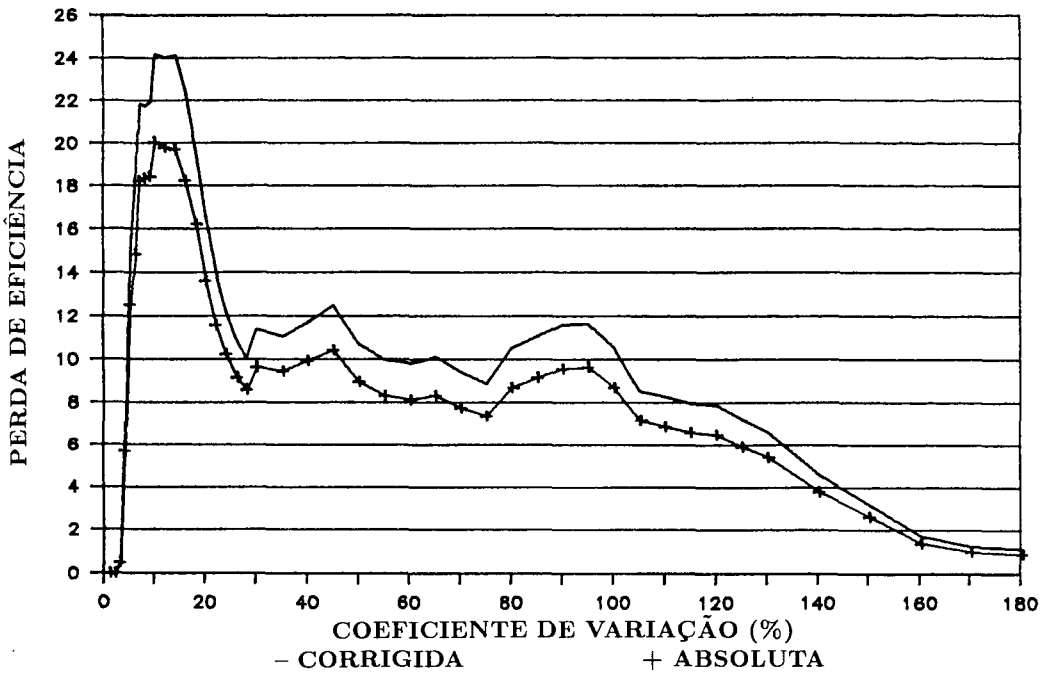


GRÁFICO 4
FRAÇÃO AMOSTRAL NECESSÁRIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 4 A 14 MIL HABITANTES

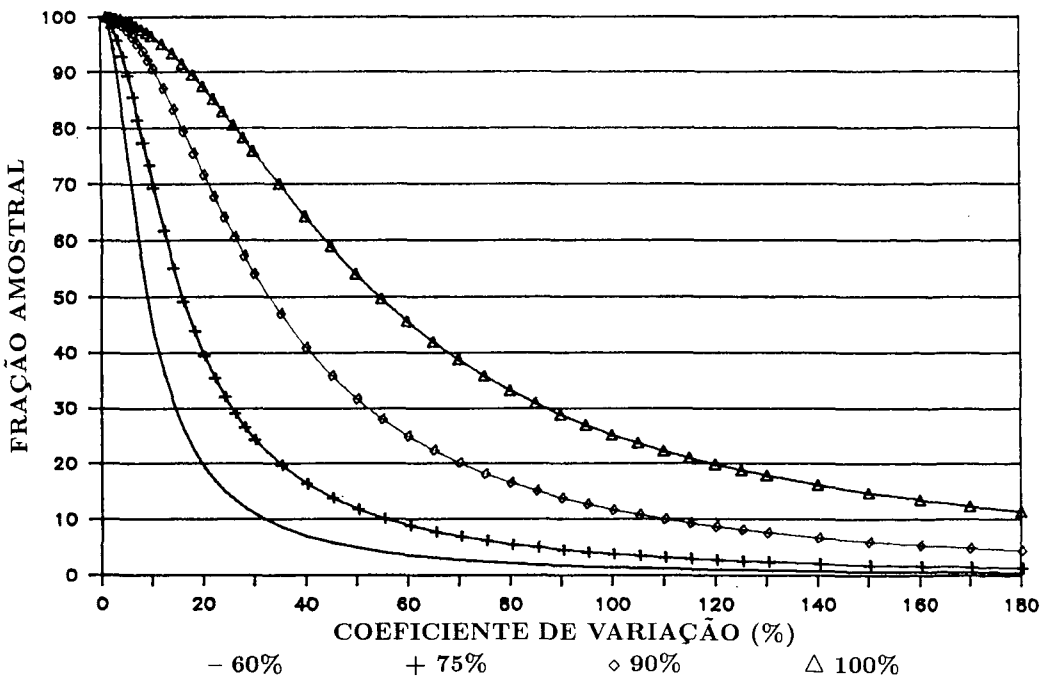


GRÁFICO 5
PERDAS DE EFICIÊNCIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 15 A 30 MIL HABITANTES

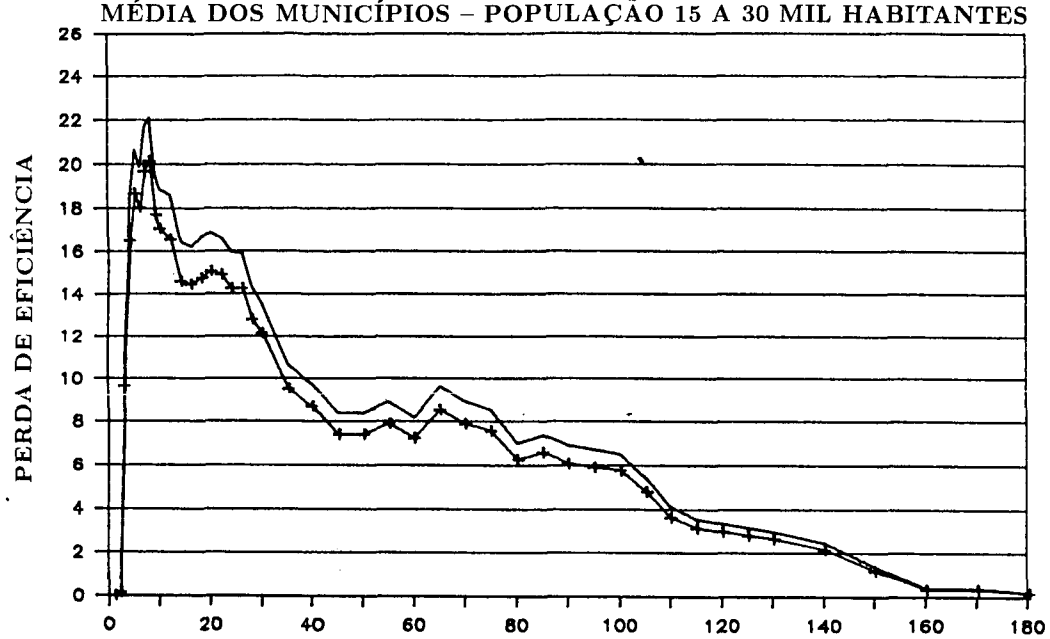


GRÁFICO 6
FRAÇÃO AMOSTRAL NECESSÁRIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 15 A 30 MIL HABITANTES

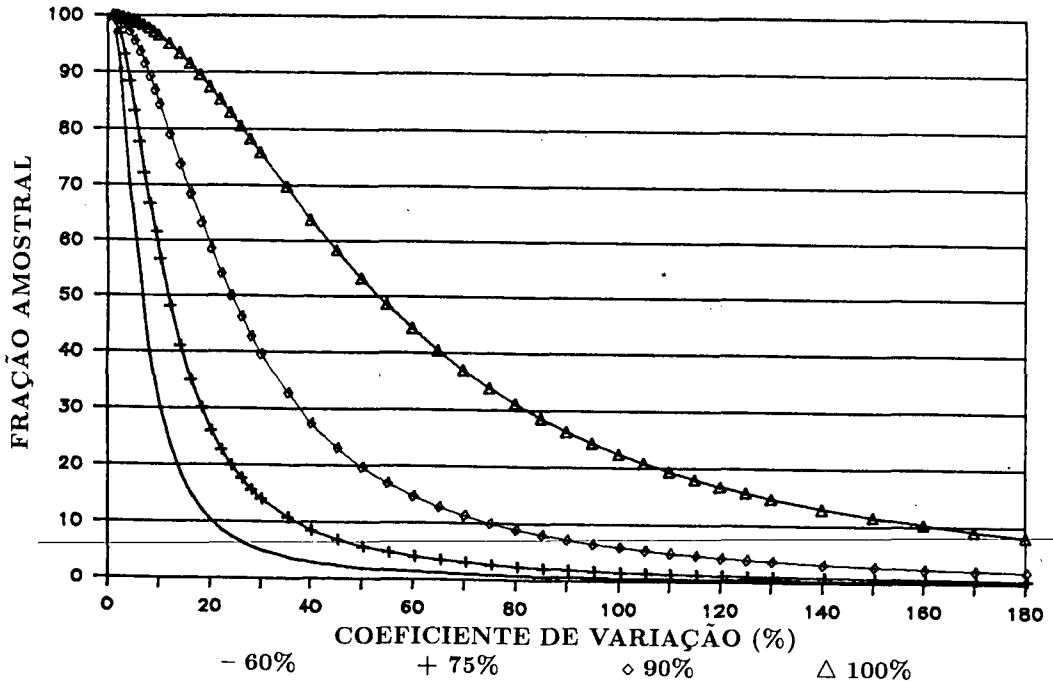


GRÁFICO 8
FRAÇÃO AMOSTRAL NECESSÁRIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 60 A 100 MIL HABITANTES

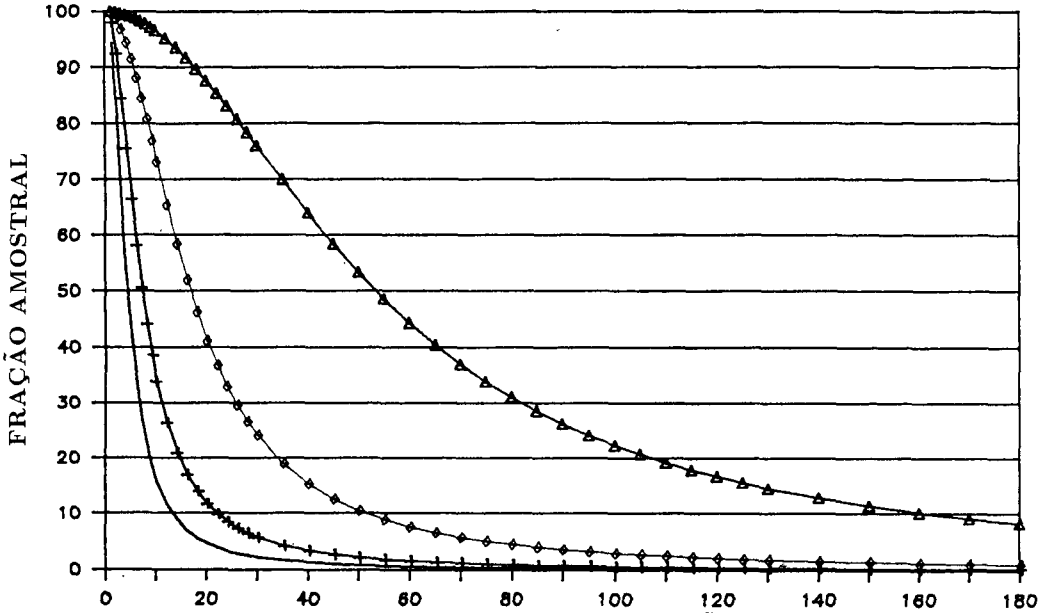


GRÁFICO 7
PERDAS DE EFICIÊNCIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 60 A 100 MIL HABITANTES

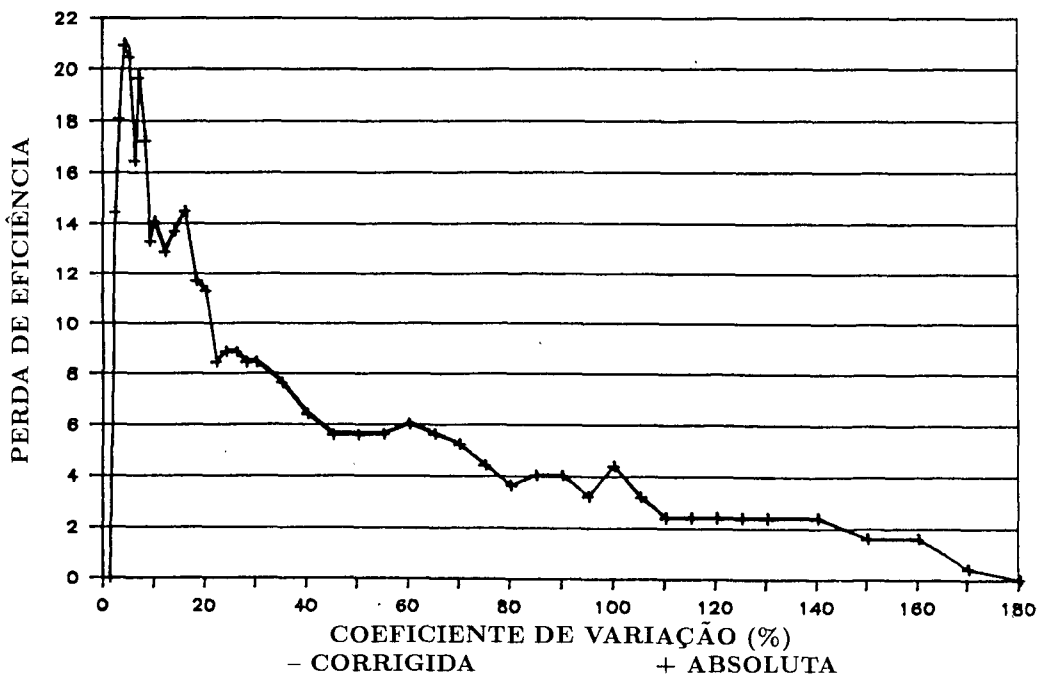


GRÁFICO 9
PERDAS DE EFICIÊNCIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 100 A 140 MIL HABITANTES

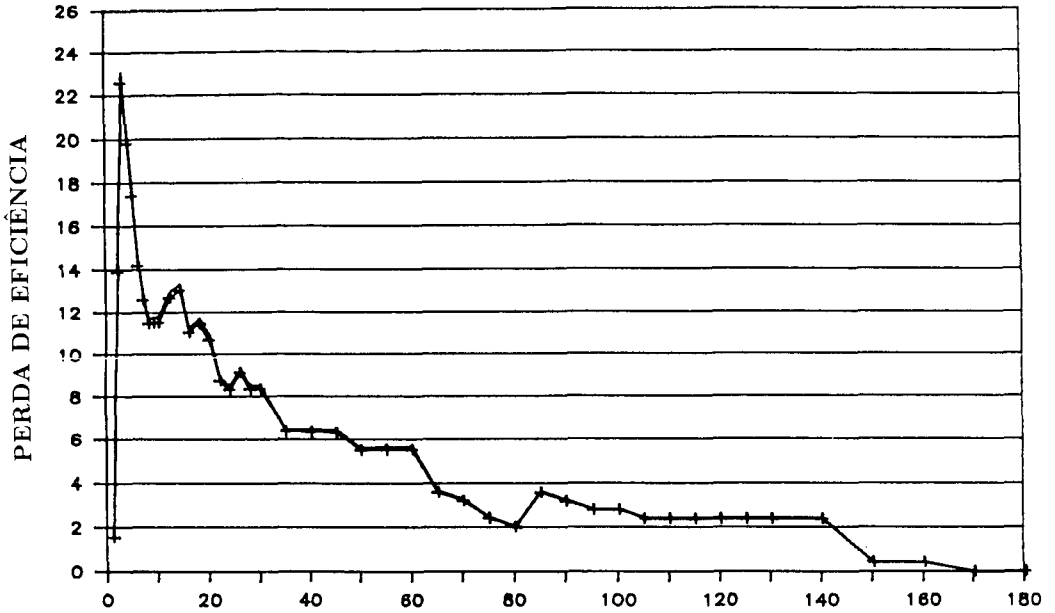


GRÁFICO 10
FRAÇÃO AMOSTRAL NECESSÁRIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO 100 A 140 MIL HABITANTES

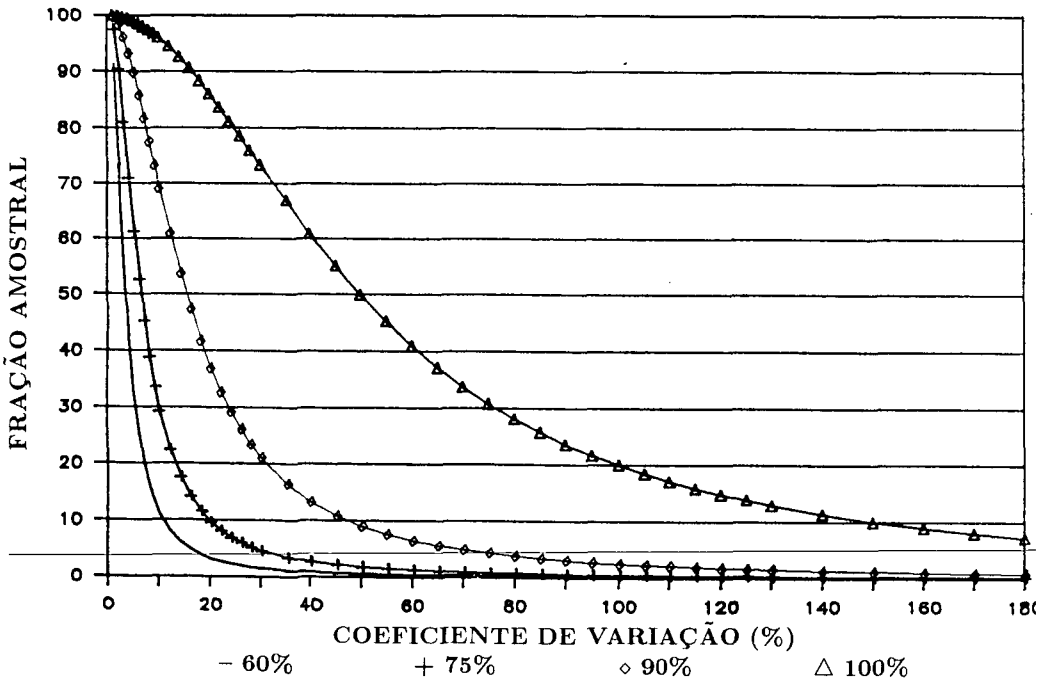


GRÁFICO 11
PERDAS DE EFICIÊNCIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO > 400 MIL HABITANTES

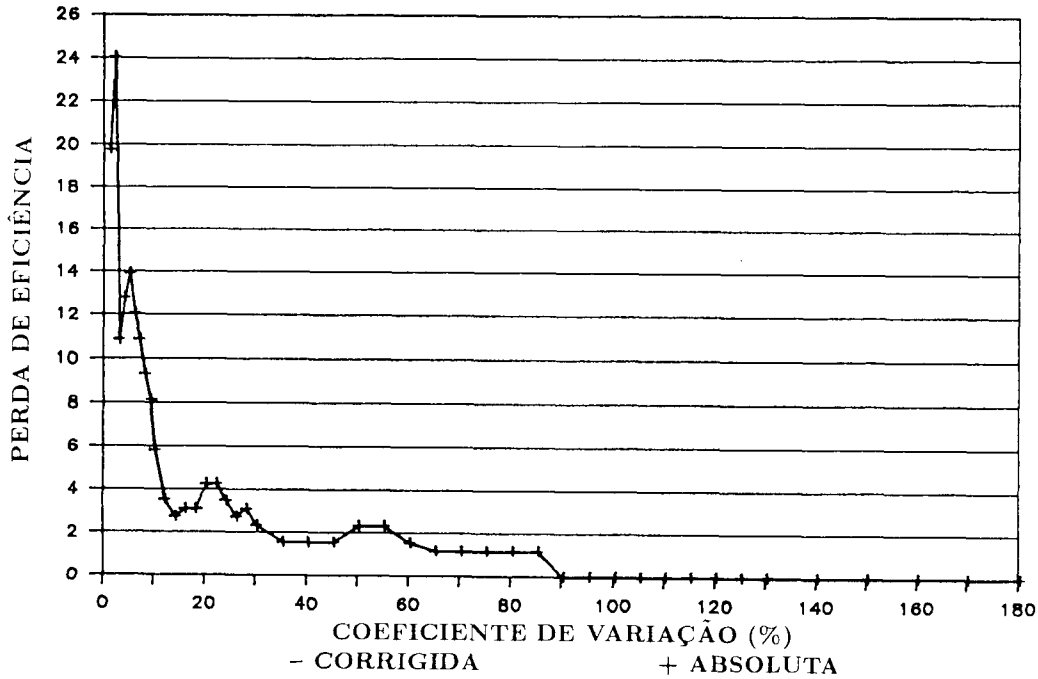
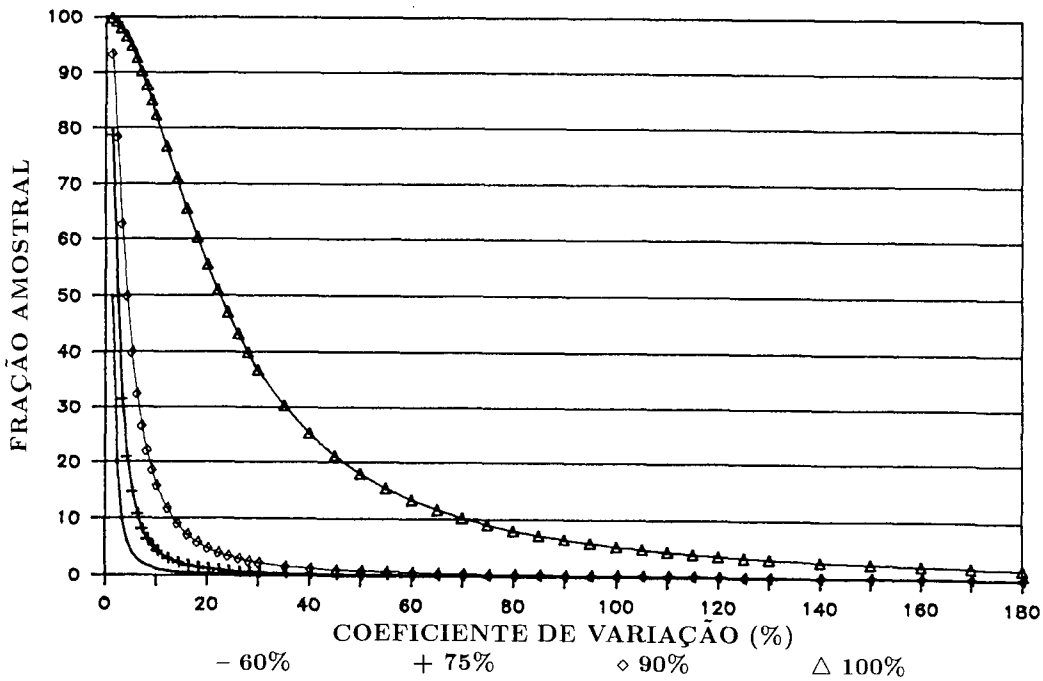


GRÁFICO 12
FRAÇÃO AMOSTRAL NECESSÁRIA
MÉDIA DOS MUNICÍPIOS – POPULAÇÃO > 400 MIL HABITANTES



BIBLIOGRAFIA

[1] Fundação Instituto Brasileiro de Geografia e Estatística. *Metodologia do censo demográfico de 1980.* Série Relatórios Metodológicos/IBGE, ISSN 0101-2843, V.4, Rio de Janeiro, 1983.

[2] COSTA, LUIZ NERY & LIMA, JOSÉ MATIAS. *Avaliação da Consistência das Estimativas de Características de Domicílios Obtidas Através de uma Amostra de 10% Simulada a partir da Amostra de 25% do CD-80* (Documento Interno) – DPE/DEPOP/NME, 1988.

[3] BIÇKEL, PETER J. & DOKSUM, KJEKK A. *Mathematical Statistics: basic ideas and selected topics.* San Francisco, Düsseldorf, Holden-Day, c-1977.

[4] COCHRAN, WILLIAM GEMMELL. *Sampling Techniques* - 3rd. ed. New York, J. Wiley, c-1977.

RESUMO

Neste trabalho é desenvolvida uma medida de perda de eficiência da precisão das estimativas decorrente da redução da fração de amostragem.

A medida, aqui apresentada, foi desenvolvida no sentido de suprir as deficiências da medida clássica de eficiência relativa entre estimadores, para atender aos propósitos do trabalho, ou seja, analisar até que ponto a perda de eficiência representada pela redução da fração de amostragem poderá comprometer a qualidade das estimativas obtidas através da amostra.

É também desenvolvido um estudo referente à fração de amostragem necessária para garantir um nível prefixado de precisão na estimação de características de interesse.

Uma aplicação da metodologia desenvolvida é incluída, tomando-se por base as características de domicílios investigadas no Censo Demográfico de 1980 que tiveram suas estimativas divulgadas a nível de município e as frações de amostragem que vêm sendo cogitadas para utilização no Censo Demográfico de 1990.

ABSTRACT

A measure of efficiency loss for the precision of estimates implied by sampling fractions reduction is described here. The measure introduced aims to avoid deficiencies of the usual measure of relative efficiency between estimators.

A study of the sampling fractions needed to guarantee a given precision level to estimate characteristics of interest is also performed.

An application of the methodology suggested is included. The example is based on household characteristics investigated in the Demography Census of 1980, with estimates published at a county level and for the set of alternative sampling fractions for the 1990 Demography Census.

GENERALIZED EDIT AND IMPUTATION SYSTEM: Overview and Applications

J. G. Kovar and P. Whitridge*

1 — INTRODUCTION

In 1985, Statistics Canada undertook a major project with the goal of redesigning all of the Bureau's economic surveys. As part of this Business Survey Redesign Project (BSRP), the development of generalized software was emphasized, in an attempt to conserve resources and eliminate duplication (Outrata and Chinnappa, 1989). The main goal of the Generalized Survey Function Development (GSFD) project was to develop generalized tools that could be adapted easily to the majority of business and social surveys which would undergo redesign. The systems would be based on a limited set of standardized methodological approaches designed to improve timeliness, reduce respondent burden and minimize resources in the production process. In addition, these systems would be flexible enough to permit new modules to be incorporated as they were developed, and would be portable across various system architectures and sites (Doucet, 1988). The systems to be developed were: Generalized Sampling System, Generalized Data Collection and Capture System, Generalized Edit and Imputation System, and Generalized Estimation System.

In developing the generalized systems, the task of edit and imputation has been broken into two stages: preliminary editing, which is done at the data collection and capture stage, followed by automatic edit and imputation. This is different from what has been done traditionally, when the approach consisted of error detection followed

*Statistics Canada 11C RHC, Tunney's Pasture, Ottawa, Ontario, K1A 0T6

R. Bras. Estat., Rio de Janeiro, 51(195): 85–100, jan./jun. 1990.

by manual correction as the records were received and reviewed. In this traditional approach, several courses of action were used for the correction of a record, including following-up the respondent, manually supplying ad hoc values, overriding the edit, excluding the record or imputing.

In the GSFD, it was decided that only records that were unresolved after the preliminary edit stage and those with lesser impact would be sent to the Generalized Edit and Imputation System (GEIS), as a last resort. In the GEIS, all attempts would be made to resolve the remaining inconsistencies by automatic imputation.

GEIS, while still undergoing some development, has been in use at Statistics Canada for the past two years. Experience has been gained concerning how applications should best approach using GEIS. This paper concentrates on the applications that have used GEIS, representing both small and large surveys and administrative data.

The paper is divided into five parts. Section 2 provides an overview of GEIS. A description of each major component of the system, including editing, error localization and imputation is given. Section 3 describes the general points that should be addressed by an application considering using GEIS. Section 4 outlines the experiences of different applications: the 1991 Census of Agriculture, Income Tax Data, and the Annual Motor Carrier Freight Survey. The paper concludes with a short summary of the insights gained from these projects.

2 — OVERVIEW OF GEIS

GEIS consists of three major components: editing, error localization, and imputation. The objective of editing is to determine whether a given data record contains invalid, missing, inconsistent or questionable responses. In other words, editing is the act of error detection. Imputation is the task of replacing an invalid, missing, inconsistent or questionable value with a plausible one. Effectively, imputation is then error correction. The two functions are linked through error localization, which is the process of determining which fields to impute. See Figure 1, reproduced here from Kovar (1990a).

Historically, editing has been a manual process, with the edits taking the form of if-then-else conditions. Rules of this type effectively perform both edit and imputation: the if condition is the edit and the then or else action is the imputation. The edits are applied sequentially to the data, with corrections being made for each failing edit. These corrections are often subjective and usually not reproducible. In certain circumstances, corrected fields can be subsequently changed if they fail further edits.

The philosophy behind GEIS is based on the Fellegi-Holt approach to edit and imputation (Fellegi and Holt, 1976), which recommends that the minimum amount of respondent data be changed. To achieve this, GEIS considers the edits as a set of

rules which defines a feasible region with good records inside the region and records which require imputation outside the region. The fields that require imputation are determined as a function of the edit failures for each individual record. In GEIS, the editing, error localization and imputation are separate steps (Giles and Patrick, 1986). These steps are described in more detail below.

2.1 – Editing

Editing in GEIS, consists of iteratively specifying, analyzing, and applying the edits. The edits are assumed to be linear and the data continuous and non-negative. If necessary, it is often possible to transform the data or the edits to satisfy these conditions. In addition, GEIS assumes that all follow-ups with the respondents have already been performed at the collection and capture stage, and that nothing more is to be gained by re-contact with the respondent or by referring to the questionnaire. This allows the system to be fully automatic, requiring no manual intervention in the production process.

The specification of the edits is done interactively, utilizing two or three different screens. Through any one of these screens the user specifies the edit identifier, whether the edit is a pass or fail condition, and the edit itself, by providing the variables with their coefficients and the constant (GEIS Development Team, 1989). There is also a facility to update the edits, attach comments to the edits, and automatically date the changes. Other screens are used to group the edits. These groups are used to create edit sets that will be applied together either to sections of questionnaires (e.g. income versus expenses), to subsets of the population (e.g. different industry groups) or in some circumstances to different edit and imputation functions (e.g. error localization versus imputation). During the edit specification the system performs some syntax verification including checking as to whether arithmetic operators have been correctly specified, and whether all variables referenced are, in fact, part of the questionnaire.

Further edit analysis is possible as a result of the assumptions that the edits are linear and the data are non-negative. Linear programming techniques are used to analyze the edit set beyond mere syntax (Sande, 1979). When the check edits function is applied, GEIS verifies the consistency of the edits, that is, it ensures that the set of edits is not self-contradictory. For consistent edit sets, the system also identifies redundant edits, if any, that is, edits which do not further restrict the feasible region of data values in the presence of the other edits. By identifying the redundant edits, the system identifies the minimal set of edits: a set of edits which defines the same region as the original set, but whose further processing is more efficient due to its reduced size.

Finally, the system generates the acceptable ranges for all variables, the extreme

points of the feasible region, and the set of implied edits (Schiopu- Kratina and Kovar, 1989). The acceptable ranges identify the maximum and minimum values a variable can assume in the presence of the specified edits. The extreme points module generates the coordinates of the vertices of the feasible region. These vertices can be thought of as the worst possible data records, that is, data records that would pass all the edits, but which are in the corners of the acceptance region specified by the edits. Such records may suggest to the user that some edits should be changed, or other, more restrictive edits should be added to the existing set. By contrast, the implied edits module generates linear combinations of the input edits, thus uncovering conditions which are being imposed on the variables, but which have not been stated explicitly in the original set of edits. Implied edits may indicate that some variables are being overly constrained and thus suggest that some edits should be relaxed or perhaps removed entirely. All of these diagnostics can aid the analyst in verifying that the specified edits are meaningful, and act as a check on the correct entry of the edits (Giles, 1987 and 1988; Sande, 1988).

Once the user is satisfied with the edits, the edits should be fine-tuned preliminary or past data, or test data if this is not possible. The last module of the edit component applies the edits to a set of data and provides counts of edit failures. Counts indicating edits which fail too often or not often enough may suggest that some changes are necessary. If the edit set is changed as a result, the specification, analysis and application steps should be repeated. Often a number of cycles through the three steps is necessary before a satisfactory set of edits is established. The analysis of edits should be carried out well before the final data is available, preferably as soon as the survey questions have been finalized. This task can often be accomplished on a microcomputer, as large amounts of data are not necessary, and the interactive nature of this architecture can be exploited.

2.2 – Error Localization

The second major component of GEIS is the error localization. This module identifies the minimum set of fields that require imputation for each particular record failing at least one edit. The fields to impute are a direct function of the acceptable data on the record and the edit set. In order to take into account varying levels of reliability of fields the user may introduce weights. For example, if a respondent is more likely to supply a valid total than the components of the total to a question, then the total would be assigned a higher weight, and hence would be less likely to be imputed. The actual algorithms used to perform this task are based on the work of Chernikova (1964, 1965) Rubin (1973) and Sande (1979). They have been documented in detail by Schiopu-Kratina and Kovar (1989); the particular implementation is described by

the GEIS Development Team (1990).

2.3 – Imputation

Once the fields to be imputed have been determined, what remains is simply the imputation. The objective is to impute new values in such a way as to preserve the underlying structure of the data and to ensure that each resulting data record will pass all the required edits, whenever the method allows for this. In other words, the objective is not to reproduce the true micro-data values, but rather to establish internally consistent records that will yield sensible aggregate estimates.

GEIS provides three main types of imputation (Giles and Patrick, 1986). The first is deterministic imputation, where a value is supplied if there is only one possible way to fix the fields, given the edits and the valid responses on the record. An example would be if components were provided but the total was left blank. In that case the total would be deterministically imputed as the sum of the components. All records imputed in this fashion will automatically satisfy the edits.

The second type of imputation uses imputation estimators, which estimate the missing values using previous observations, means, trends, or ratios, one variable at a time. Note that while these model based estimators are often intuitively appealing, the resultant record does not necessarily satisfy the edits. Details of the methods may be found in Giles (1986) and Kovar (1990a), while the exact formulae are described by the GEIS Development Team (1990).

The last type of imputation is a donor method using the nearest neighbour technique. For this purpose, matching variables (those used in the distance function) are first determined for each record in error. These variables are either system-generated as a function of the edits (Schiopu-Kratina and Kovar, 1989), or user-specified. The matching fields are then transformed to uniform marginals to make all variables independent of scale, and a search tree is created (Friedman, Bentley and Finkel, 1977). The tree is traversed for each record in error and the nearest neighbours are found. Here nearest is in terms of a distance based on the transformed matching variables. The closest record is used to supply the missing fields, and the imputed record is re-edited. If the edits are not satisfied, then the next closest neighbour is tried, until either the record is successfully imputed, or the supply of nearest neighbours is exhausted. Details of the method can be found in Sande (1979), while the implemented algorithms are described by the GEIS Development Team (1990).

GEIS provides reports throughout, to help managers track the process. These performance measures include counts of edit failures by edit, record and field; frequency of imputation by field and method; identity of donors used; match fields used; number of times a record was used as a donor; number of records that had a given number of

fields imputed; the number of imputations attempted; etc. These reports are not only useful in documenting what has happened, but also in establishing future directions for both the given application as well as the development of the system.

3 — APPLICATION OF GEIS

It can be assumed that most surveys will require an edit and imputation system. However, what edits and what methods of imputation to use are decisions that must be considered individually for each application. The main objective of GSFD is to provide tools that can be adapted for the majority of surveys. It is hoped that the methodologists can take the systems for granted and concentrate their time on developing comprehensive edit and imputation strategies and testing alternatives, rather than writing specifications and testing new systems.

Any application considering the use of a generalized system must first determine whether or not the system is appropriate for that particular application. For example, GEIS imputes continuous numeric variables, so it might not be suitable for an application that involves a large number of categorical or qualitative variables. Usually a feasibility study is carried out to determine whether or not the application should proceed with the implementation of the generalized system.

Having decided to use GEIS, the primary issues that must be addressed in the development of an edit and imputation strategy include determining and fine-tuning the edits, choosing imputation methods, and designing an evaluation of the entire process. GEIS incorporates functions to help the methodologist analyze and fine-tune the edits, but the original edits themselves must be determined outside of the system. There are different imputation methods available, but the applications must decide which methods suit them and specify values for numerous parameters. The question of how to evaluate the imputation is left to the users, though the system makes it easier by providing some summary tables. The next sections address these three issues in order.

3.1 – Development of Edits

When an application is considering the use of GEIS, there is a tendency to want to implement the same traditional edit and imputation methods that were in place previously. This is not recommended, since, typically, the old edits are if-then-else rules with the error localization and imputation actions implicit.

This type of rule can often be linearized, but it is no longer possible to specify particular imputation actions for each rule. As well, these traditional edits usually

identify invalid conditions rather than clean records. Even though the system accepts such conflict rules as input, the hazard in this approach is that any unforeseen condition is implicitly considered acceptable, since it was not specified as being incorrect. In the case of continuous variables, it is generally easier to start by describing a clean record rather than by trying to enumerate all possible invalid conditions.

Several possible sources of edits must be considered when an application is defining its edit requirements. The primary source is the subject matter expert. Subject matter knowledge is highly variable and somewhat subjective, since it depends upon the human expert, but it can be extremely useful in identifying possible combinations of variables that should be considered together. Usually, the experts would specify context sensitive relationships, for example BULLS < COWS/10, or L-OF-MILK < 100*COWS, and sensible bounds, for example 25K < SALES < 10M. As well, edits of the form "if A>0 then B>0", referred to as existency edits, are often indicated, for example "if SALES were reported, then PURCHASES must be reported", or "if L-OF-MILK were reported then COWS must be reported". Note that these latter, conditional or logical edits, can be approximated by a linear edit as for example "PURCHASES > 0.0001*SALES".

The second main source involves a logical examination of the questionnaire. Certain relationships are evident, for example, components that sum to a total and variables that are subsets of others (e.g. number of cows milked <= total number of cows). As well, the structure of the questionnaire will identify logical blocks of variables, and quantity/quality pairs which are clearly related (e.g. # of tractors, their $ value).

Previous or related surveys are another source that can be useful. Typically the edits from these other surveys, regardless of the form, can be helpful in identifying relationships to examine. For example, tax data covers all industries, so the rules used to edit the retail sector of tax data could be useful for retail surveys in general and vice versa. Edits used for previous surveys should be adjusted based on their effectiveness. In some cases the constants require recalibration, for example because of inflation.

Perhaps the most powerful source of information available in specifying edits is data analysis. Principal Components analysis techniques can be used to group the variables into clusters. Correlations can be calculated to determine which variables within a cluster are related. Graphical analysis or other exploratory data analysis techniques can be used to specify bounds to be used in edits. Statistical edits based on the distribution of certain variables or the ratio of certain variables can also be identified (Hidiroglou and Berthelot, 1986). For a detailed description of all of these techniques, see Bilocq and Berthelot (1990).

Due to operational constraints, GEIS cannot edit more than 40 variables at one time. This often means that an application must divide the variables into groups of no more than 40. The edits are then grouped according to the variables involved. This grouping of variables can be chosen to correspond to sections of a questionnaire (e.g.

livestock, crops, expenses) or to some other logical arrangement of the variables. Care must be taken to ensure that a variable is not imputed more than once, but there are facilities in GEIS to help take care of this.

When the edit rules have been determined, they should be analyzed using the facilities available in GEIS, as discussed in Section 2. If the application includes any variables that could have valid negative values, then these variables would need to be transformed by the addition of a constant to make them positive. Any edits involving these transformed fields would need to incorporate the same constant.

Once the edit rules have been analyzed, GEIS can be used to apply the edits to preliminary data (test data, pilot survey data, historical data, or data from other sources) to fine-tune them based on the actual failure rates. Edits that fail in a very high percentage of cases, or never fail at all, usually require adjustment of the constants, or complete removal. Review with subject matter experts is often very useful at this stage. As edits are added, removed, or changed, they should be re-analyzed as described in Section 2.

Finally, it should be emphasized again, that all edit failures in GEIS are resolved by means of imputation of one or more fields on the record. In other words, GEIS does not recognize notions such as warning edits, and as such, the pass/fail edits of GEIS must be thought out very carefully. It is also of utmost importance that subject matter personnel feel comfortable with the approach in general and the edit set in particular (Kovar, 1990d). Accepting and understanding the edits and their implications makes the subject matter expert's job of data analysis much easier.

3.2 – Imputation Strategy

The imputation strategy in GEIS consists of specifying methods of imputation to be used such that the final product is a complete fully-imputed set of data. Ideally, relationships between variables and higher moments of distributions should be preserved, but this is not always possible. The best imputation strategy depends upon the characteristics of an individual application. The issues to consider include the level of aggregation at which imputation should take place, the choice of imputation methods and certain technical, imputation-specific questions as discussed below.

It is important to establish carefully the level at which to impute, in order to ensure that geographical and classification structures are preserved. Often this leads to a large number of imputation groups. By contrast, it is also desirable to avoid having too many imputation groups, as each imputation group must be processed individually and this increases the number of computer jobs to be managed and decreases the number of potential donors. A compromise must therefore usually be reached.

GEIS incorporates three different types of imputation methods, as described in Sec-

tion 2. For most applications, the donor method is recommended, though repeated subannual surveys whose characteristics are highly correlated over time may be better served by methods which make use of past information (Kovar, MacMillan and Whitridge, 1988). Many applications specify the donor method as the primary one, but use imputation estimators as back-ups in case no appropriate donor is available. Surveys with large numbers of variables find the imputation estimators cumbersome to use, since a method including any necessary auxiliary information must be specified for each variable requiring imputation. If estimators are to be used, then the order in which they are applied is important, since, for example, it is desirable to impute a variable before using it as auxiliary information for imputing another variable. Correlations between variables should be examined when auxiliary information is required.

Certain other imputation questions still remain. First, donor imputation requires the specification of post-imputation edits to be used to determine if an imputation was successful. These edits could be the same as the original edits used in error localization, or they could be a relaxed version of the same set. For example, it might be desirable to relax an equality edit into a range bounded by two inequality edits, as it is unlikely that a donor-imputed record would satisfy all equality edits. Second, if estimators are used, some applications require post-imputation edits to verify the imputation, but this is not easily implemented in GEIS. Some re-processing of records through error localization is necessary. Third, if donor imputation is used, then matching fields can be specified by the user if desired. This facility can be used to bring fields that are not part of the block of variables being edited into the search for a good donor. This is often useful in bringing into play a global measure of size, such as the Gross Business Income or Total Sales variables. Fourth, minimum imputation criteria can be specified, e.g. number of records used to calculate a mean or trend, or minimum donor population size. This is useful in controlling the reliability and stability of the values to be imputed. Finally, it is sometimes necessary to exclude certain records from the donor population, such as zero values, or suspicious or outlying values. All these points must be considered as part of the imputation strategy.

3.3 – Evaluation Strategy

The system provides some tables that can be used for evaluation. As well, the data should be analysed to produce counts such as the number of times a reported value was increased, or decreased, or remained unchanged, and the estimates of the corresponding totals. The tables typically would be produced at several different levels of aggregation, perhaps corresponding to the estimates to be tabulated. They can be used both as input to feasibility studies, as well as post processing documentation.

Because of the nature of the software environment (ORACLE), as well as the available choice of computer platforms, the users may monitor the edit and imputation process more effectively and thoroughly.

Evaluating the impact of the imputation is often difficult, since it requires a pre-specified idea of the true values, and of how much imputation is acceptable. These are very subjective measures. For example, acceptable imputation rates depend upon the response rates, the reasons for imputation, and which specific field is being imputed. Moreover, the impact of imputation must be measured carefully, using the proper subsets of data. For example, applications with a high rate of partial non-response will be replacing zeroes by positive values, and the impact will seem to be in a positive direction only.

The evaluation is best done co-operatively between the subject matter experts who are responsible for the data being published and the methodologist who is designing the edit and imputation system. It is a highly subjective exercise which varies from application to application and tends to be built upon with experience.

4 — EXPERIENCES WITH GEIS

GEIS has been in use at Statistics Canada for the last two years, while it has been undergoing continued development. Experience gained from the first applications has helped determine directions for this development, as the product is made more efficient and user friendly. This section describes the experiences of three applications: the Census of Agriculture, Income Tax Data, and the Annual Motor Carrier Freight Survey.

4.1 – Census of Agriculture

In September, 1989 the Census of Agriculture established a working group to study the feasibility of implementing GEIS for edit and imputation of the 1991 Census. Various processing scenarios were examined and an overall recommendation was made. The primary problems associated with the Census were the large volume of records (300,000 farms) and the large number of variables (320) to be edited and imputed. Of equal concern was the known heterogeneity of the population.

It was decided that the best way to proceed for this application would be to develop an actual prototype and to completely pass it through the edit and imputation system. The section of the questionnaire dealing with livestock was selected and eight imputation regions from across Canada were chosen. The livestock section had 29 variables, so it was ideal for one edit group. It was felt that livestock was a good choice in that

there would be a reasonable number of relationship edits that could be derived and that editing and imputing this section was of at least average difficulty. The eight imputation regions involved 35,000 records (of 1986 data), so they would adequately test the volume. Since the regions were from across Canada, they would represent different types of farms: beef cattle in the province of Alberta, pork producers in the province of Quebec, and dairy farms in the province of Ontario.

The questionnaire for the Census has a number of tick boxes to indicate whether or not a farm has certain characteristics, such as fruit trees, field crops, or cattle. If the box is ticked yes but no data are provided, then the entire section of the questionnaire should be identified to be imputed. This action is really the result of a response pattern that should be detected by the capture system, but presently is not. As a result, a pre-processor program was written to help resolve this issue.

Edits were determined based on subject matter edits that had been used in the previous Census, current subject matter knowledge, intense data analysis and examination of the questionnaire itself. The data analysis yielded ratio edits between pairs of variables. The analysis was based upon a combination of data from the 1986 Census and data from a pilot test that took place in 1989. The new set of edits was presented to the subject matter experts and explained graphically. It was felt that these edits might better identify the units to be imputed than the edits used in 1986 and would react correctly to the heterogeneity of the population.

The prototype simulated different processing scenarios that might be used in production. The regions were all run at once to simulate processing one entire province and the livestock section was replicated and run several times at once to simulate the entire questionnaire. Different combinations of edit and data groups were then run to determine the optimal scenario. The impact of the imputation was evaluated and it was felt that the overall quality of the data after imputation would be as good if not better than that of the 1986 Census of Agriculture, which was considered excellent (Statistics Canada, 1990).

The overall conclusions of the feasibility study were positive. The working group recommended that GEIS be used for the 1991 Census of Agriculture and this was accepted by senior management.

4.2 – Income Tax Data

Historically at Statistics Canada, automatic imputation is not performed on the Income Tax Data that is used as a frame for various business surveys. The situation is changing as economic surveys are encouraged to take advantage of administrative sources of data especially for small businesses. Studies have recently taken place to determine if it would be feasible to edit and impute this data (Michaud and Bureau,

1988; Bureau, Michaud and Kovar, 1988).

Based on the results, it was decided to use GEIS to perform automatic edit and imputation on Income Tax Data from Tax Year 1988. The issues to be addressed included differentiating between zeroes and missing values and developing an edit and imputation system that would satisfy all potential users of the tax data, both those interested in one industry such as Construction, and those whose interests covered all industries such as Small Business Statistics. The system had to be optimal for the set of all industries, which meant that it would probably not be the best for any given industry. The problem was also a question of volume: 200,000 records were involved with 24 variables each. The file was divided into imputation groups according to industry; one edit group only was required.

In Canada, businesses supply their financial statements in whatever format they choose; Revenue Canada Taxation only requires that the data be complete. However, when Statistics Canada transcribes the financial statements, there are often items that business have grouped together or specified in a form that is incompatible with the transcription form. As a result, these items are then captured as zeroes, when they might really be missing values. The main objective of edit and imputation for tax data is to identify these missing values and replace them with valid values. The edits to perform this task are largely conditional rules that are not suited to implementation in GEIS. This issue of missing values is really one of non-response that should be identified at the capture stage. To solve this problem a program was written to pre-process the data and supply GEIS with flags indicating which values were missing and should be imputed.

Edits were developed by subject matter experts for their areas of interest: some were for one specific industry, and others covered all industries. The edits were assembled into one group for each industry then analyzed. Any inconsistencies or redundancies were resolved co-operatively between the appropriate experts.

The edit and imputation of Tax Year 1988 was run early in the summer of 1990. The preliminary results were positive, although the evaluation of the impact of imputation has not yet been completed.

4.3 – Annual Motor Carrier Freight Survey

The Annual Motor Carrier Freight Survey is currently undergoing a redesign and edit and imputation was identified as one area which should be addressed. The system in place relies on clerical staff manually performing detailed edits, follow-ups and imputation. For this reason, a study was undertaken in February 1990 to investigate the feasibility of using GEIS for this application. Various imputation methods were examined in order to gain information that would help determine an optimal imputation

strategy.

The feasibility study concentrated on one section of the questionnaire for one class of motor carriers. The twenty-nine variables on the Balance Sheet were considered for 1300 records. A large number of edits were specified, mostly defined by the accounting structure of the Balance Sheet. The edits were complex and recursive, some dealing with a subsequent subtotal. There were many equalities. A pre-processor program was required to correctly error localize the cases where a total was provided but the components were all missing, so that the entire set of components would be identified to be imputed. Specific problems that had to be addressed concerned the large amount of non-response and the complex pattern of edits required by the survey. Some problems were experienced in the error localization module during the study due to the large number of variables and edits.

As a result of the feasibility study, it was recommended that GEIS could be used the Annual Motor Carrier Freight Survey, if some minor enhancements were made to the error localization module of GEIS (Gossen, 1990).

5 — SUMMARY

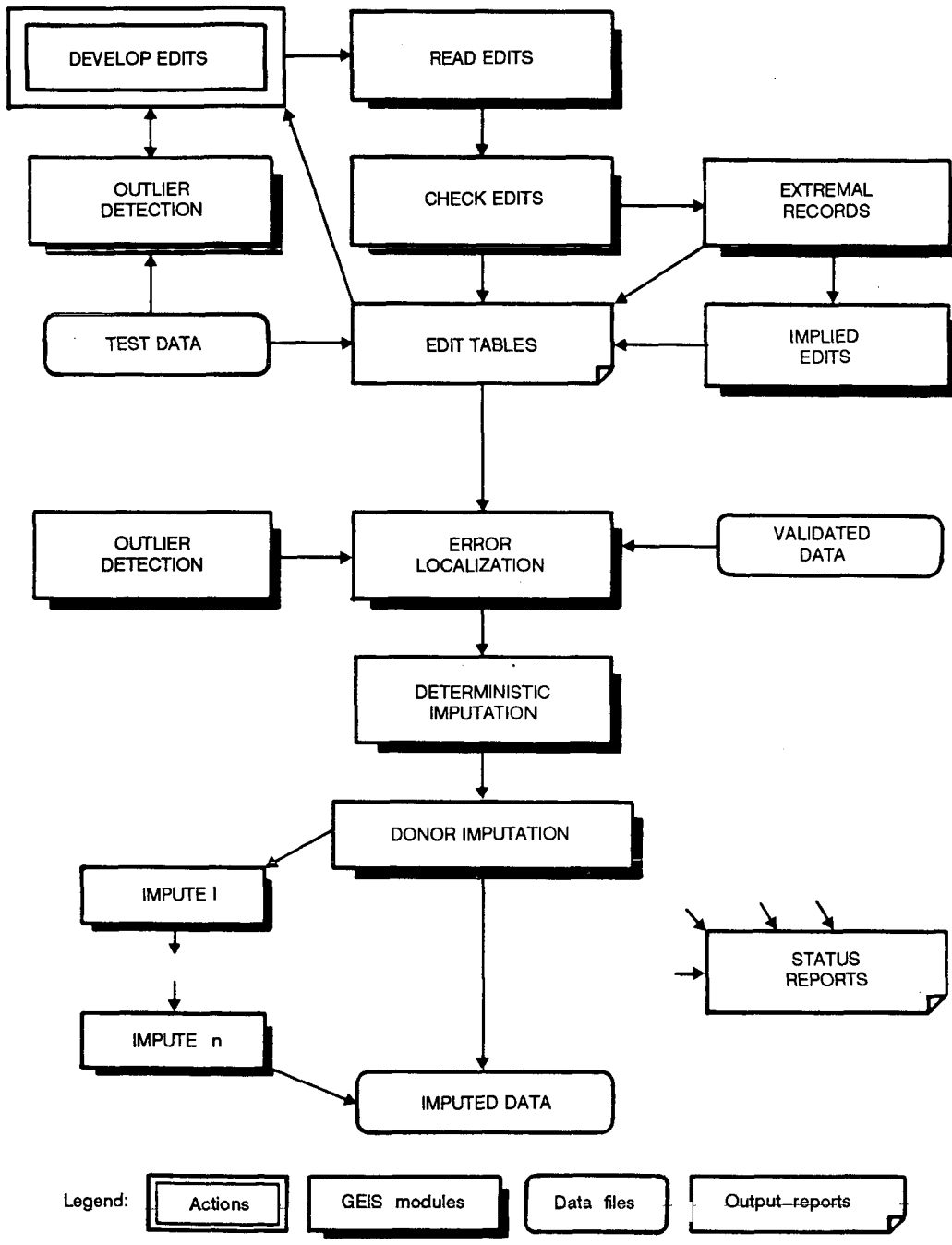
The overall experience to date with GEIS has been positive. The system is appreciated by methodologists since it allows them to spend more time developing edits and an imputation strategy instead of writing specifications and testing new systems. The applications that have studied GEIS have provided valuable input to the GEIS development team on how to improve its functionality and how to make the system more user-friendly. Many enhancements currently being programmed were designed as a direct result of requests from the users.

A common concern about the system involves the need for a pre-processor program to massage the data before it can be processed by GEIS. This was the case for all three of the applications described here. It is felt that such a program will not be necessary once a sophisticated capture system, such as the Generalized Data Collection and Capture System (DC2), is in place. This system would correctly identify all missing components from a total as requiring imputation, rather than just one field as would be the case with GEIS.

Surveys that have considered the use of GEIS have tended to prefer simply implementing existing edit rules rather than taking the opportunity to re-think the way the edit and imputation has been done. The need for a pre-processor to handle conditional edits, as in the case for the tax application, is a direct result of this problem. It would be more efficient if applications were to rethink the edit and imputation strategy and develop a set of linear edits that take advantage of the way GEIS functions. Similar edit requirements to those currently in place could then be used in GEIS, although

the actual edit specifications would look very different. This problem should resolve itself with time, as applications become more accustomed to working with generalized systems.

Figure 1: A schematic diagram of a possible arrangement of the major modules of GEIS



BIBLIOGRAPHY

BILOCQ, F., BERTHELOT, J.-M. (1990). Analysis on grouping of variables and on detection of questionable units. Statistics Canada, Methodology Branch Working Paper No. BSMD 90-005E/F.

BUREAU, M., MICHAUD, S., KOVAR, J.G. (1988). Edit and imputation of tax data. Proceedings of the Section on Business and Economic Statistics, American Statistical Association Meetings, New Orleans, 372-375.

CHERNIKOVA, N.V. (1964). Algorithm for finding a general formula for the non-negative solutions of a system of linear equations. USSR Computational Mathematics and Mathematical Physics 4, 151-158.

CHERNIKOVA, N.V. (1965). Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. USSR Computational Mathematics and Mathematical Physics 5, 228-233.

DOUCET, J.E. (1988). Redefining software portability: A new design objective for applications. Statistics Canada Technical Report.

FELLEGI, I.P., HOLT, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association 71, 17-35.

FRIEDMAN, J.H., BENTLEY, J.L., FINKEL, R.A. (1977). Algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software 3, 209-226.

GILES, P. (1986). Generalized edit and imputation - part II. Statistics Canada Technical Report.

GILES, P. (1987). Towards the development of a generalized edit and imputation system. Presented at the U.S. Bureau of the Census Third Annual Research Conference, Alexandria, Virginia.

GILES, P. (1988). Generalized edit and imputation of survey data. Statistics Canada Special Issue of the Canadian Journal of Statistics 16, Supplement, 57-73.

GILES, P., PATRICK, C. (1986). Imputation options in generalized edit and imputation systems. Survey Methodology 12, 49-60.

GEIS Development Team (1989). Generalized Edit and Imputation System: User's Guide. Statistics Canada Technical Report.

GEIS Development Team (1990). Generalized Edit and Imputation System Specifications. Statistics Canada Technical Report.

GOSSEN, M. (1990). GEIS Feasibility Study for the Annual Motor Carrier Freight Survey. Statistics Canada Technical Report.

HIDIROGLOU, M.A., BERTHELOT, J.-M. (1986). Statistical editing and imputation for periodic business surveys. Survey Methodology 12, 73-83.

KOVAR, J.G. (1990a). Generalized Edit and Imputation System: An overview. Invited presentation at the Workshop on Data Editing and Imputation Methods, Rio de Janeiro, Brazil, February 12-16.

KOVAR, J.G. (1990b). Generalized Edit and Imputation System: Applications. Invited presentation at the Workshop on Data Editing and Imputation Methods, Rio de Janeiro, Brazil, February 12-16.

KOVAR, J.G. (1990c). Generalized Edit and Imputation System: Algorithms. Invited presentation at the Workshop on Data Editing and Imputation Methods, Rio de Janeiro, Brazil, February 12-16.

KOVAR, J.G. (1990d). Automatic editing: A discussion. Presented at the US Bureau of the Census 1990 Annual Research Conference, Washington, DC, March 18-21.

KOVAR, J.G., MACMILLAN, J.H., WHITRIDGE, P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD 88-007.

MICHAUD, S., BUREAU, M. (1988). Edit and imputation of tax data: An overall strategy. Statistics Canada, Methodology Branch Working Paper No. 88-018E.

OUTRATA, E., CHINNAPPA, N. (1989). General survey functions design at Statistics Canada. Invited paper, presented at the 47th Session of the International Statistical Institute, Paris, August 29 - September 6.

RUBIN, D.S. (1973). Vertex generation in cardinality constrained linear programs. Operations Research 23, 555-565.

SANDE, G. (1979). Numerical Edit and Imputation. Presented at the 42nd Session of the International Statistical Institute, Manila, Philippines.

SANDE, I. (1988). A Statistics Canada perspective on numerical edit and imputation in business surveys. Presented at the conference of European Statisticians, Geneva, Switzerland, February 2-5.

SCHIOPU-KRATINA, I., KOVAR, J.G. (1989). Use of Chernikova's algorithm in the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD 89-001.

Statistics Canada Working Group on the use of GEIS for the 1991 Census of Agriculture (1990). Summary of the GEIS Prototype Results. Statistics Canada Technical Report.

RESUMO

Como parte do projeto *Generalized Survey Function Development* do Statistics Canada, os últimos dois anos foram dedicados ao desenvolvimento de um Sistema Generalizado de Crítica e Imputação (GEIS). Este artigo fornece uma visão geral do sistema GEIS e algumas indicações das primeiras experiências em seu uso. A ênfase é dada sobre como aplicar o sistema ao invés de se oferecer descrição detalhada dos algoritmos envolvidos. Em particular, aspectos metodológicos tais como o desenvolvimento de regras de crítica e a estragégia de imputação são considerados. Experiências recentes com várias aplicações são usadas para ilustrar estes pontos. O artigo serve como resumo de três apresentações feitas pelo autor no "Workshop sobre Métodos de Crítica e Imputação" promovido pelo IBGE em fevereiro de 1990.

ABSTRACT

As part of the Generalized Survey Function Development Project at Statistics Canada, the last two years have been spent developing a Generalized Edit and Imputation System (GEIS). This paper gives an overview of GEIS and some indication of the early experiences in its use. Emphasis is placed on how to apply the system rather than on a detailed description of the underlying algorithms. In particular, methodological issues such as the development of edits and the imputation strategy will be addressed. Recent experiences with several applications will be used to illustrate these points. The paper serves as a summary of three presentations made by the first author at the Workshop on Data Editing and Imputation Methods (Kovar, 1990a-c).

FÓRMULAS MATRICIAIS PARA O FATOR DE CORREÇÃO DA ESTATÍSTICA ESCORE

Silvia L. de Paula Ferrari*
e
Gauss M. Cordeiro**

1 — INTRODUÇÃO

Consideremos o vetor $y = (y_1, \ldots, y_n)^\top$ de n observações independentes, cuja distribuição depende de um vetor de parâmetros $\beta = (\beta_1^\top, \beta_2^\top)^\top$, sendo $\beta_1 = (\beta_1, \ldots, \beta_q)^\top$ o parâmetro de interesse q-dimensional e $\beta_2 = (\beta_{q+1}, \ldots, \beta_p)^\top$ o parâmetro de incômodo $(p-q)$-dimensional. O objetivo é testar a hipótese nula $H_o : \beta_1 = \beta_1^{(0)}$ contra a alternativa $H : \beta_1 \neq \beta_1^{(0)}$. Assumimos que o logaritmo da função de verossimilhança total $L = L(\beta)$ seja regular com respeito a todas as derivadas em relação a β até quarta ordem.

Um aperfeiçoamento do teste escore (Rao, 1948) para testar a hipótese acima é examinado neste artigo. Sabe-se que existe uma equivalência assintótica entre este teste, o da razão da verossimilhança e o de Wald, sendo que o teste escore tem uma grande vantagem sobre os demais por requerer somente estimação sob a hipótese nula. Em grandes amostras as estatísticas destes testes têm, sob H_o, distribuição χ^2 com q graus de liberdade.

Sabe-se que, sob H_o, a estatística da razão de verossimilhança, que denotamos por

*Instituto de Matemática e Estatística, USP
**Departamento de Economia, CCSA/UFPE

w, tem média $q[1+b(\beta)]$ até ordem n^{-1}, onde $b(\beta)$ é uma certa função dos parâmetros dada por Lawley (1956). Sabe-se ainda que w pode ser aperfeiçoada por um fator de correção de Bartlett, de tal forma que a estatística modificada $w^* = w[1-b(\beta)]$ tenha distribuição χ^2 com q graus de liberdade até ordem n^{-1} sob H_o. Cordeiro e Ferrari (a ser publicado), baseando-se em resultados de Harris (1985), obtiveram uma estatística escore modificada

$$S_R^* = S_R[1 - b(S_R, A_1, A_2, A_3, q)], \tag{1.1}$$

onde S_R é a própria estatística escore, A_1, A_2 e A_3 são dados por Harris (1985) e b é uma certa função paramétrica dada na Seção 2. Foi mostrado que S_R^* tem distribuição χ^2 com q graus de liberdade até ordem n^{-1} sob H_o. Observa-se que $1 - b(S_R, A_1, A_2, A_3, q)$ é um tipo de correção de Bartlett para a estatística escore. Harris (1985) obteve uma expansão assintótica até ordem n^{-1} para os quantis de S_R sob H_o. Ele mostrou que, até ordem n^{-1}, •

$$z = x[1 + b(x, A_1, A_2, A_3, q)], \tag{1.2}$$

onde $\mathrm{pr}(S_R \leq z) = \mathrm{pr}(\chi_q^2 \leq x)$ com χ_q^2 denotando uma variável aleatória com distribuição χ^2 com q graus de liberdade. Assim, observa-se a existência de uma conexão entre a expansão assintótica dos quantis de S_R sob H_o e a estatística escore modificada. No caso da estatística da razão de verossimilhança, a conexão conhecida é entre a expansão assintótica da média de w sob H_o e a estatística modificada.

Na Seção 2, apresentamos a estatística escore modificada S_R^* obtida por Cordeiro e Ferrari (a ser publicado). Na Seção 3, apresentamos as quantidades A_1, A_2 e A_3, necessárias à obtenção de S_R^*, em notação matricial e fazemos uma análise do custo computacional no cálculo de S_R^*. Na Seção 4, fazemos duas ilustrações numéricas e, finalmente, na Seção 5, analisamos algumas propriedades de S_R^*.

2 — UMA ESTATÍSTICA ESCORE MODIFICADA

Seja $U = U(\beta) = \partial L(\beta)/\partial\beta = (U_1^\top(\beta_1,\beta_2), U_2^\top(\beta_1,\beta_2))^\top$ a função escore total de β e sejam $K = K(\beta) = E\{U(\beta)\, U(\beta)^\top\}$ e $K^{-1} = K^{-1}(\beta)$ a matriz de informação total de Fisher e sua inversa, particionadas segundo a partição de β, isto é,

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, K^{-1} = \begin{pmatrix} K^{11} & K^{12} \\ K^{21} & K^{22} \end{pmatrix} \tag{2.1}$$

Assumimos que K é positiva definida. Seja $\hat{\beta}_1$ a estimativa de máxima verossimilhança de β_1.

A estatística escore para testar $H_o : \beta_1 = \beta_1^{(0)}$ *versus* $H : \beta_1 \neq \beta_1^{(0)}$ (Cox e Hinkley, 1979; Seção 9.3) é dada por $S_R = \tilde{U}_1^\top \tilde{K}^{11} \tilde{U}_1$, onde $K^{11} = K^{11}(\beta)$ é a matriz de

covariância assintótica de $\hat{\beta}_1$, obtida da inversa de K, e sendo as funções assinaladas com um til avaliadas no ponto $\tilde{\beta} = \left(\beta_1^{(0)\top}, \tilde{\beta}_2^{\top}\right)^{\top}$, onde $\tilde{\beta}_2$ é a estimativa de máxima verossimilhança de β_2 restrita a H_o.

Sejam $f(x)$ e $g_m(x)$ as funções densidades de S_R e de uma variável aleatória com distribuição χ^2 com m graus de liberdade, respectivamente. Harris (1985), baseado no trabalho de Hayakawa (1977), obteve uma expansão assintótica até ordem n^{-1} para a função densidade de S_R sob H_o que, escrita de forma conveniente, se reduz a

$$f(x) = g_q(x)(1 + B_0 + B_1 x + B_2 x^2 + B_3 x^3) \tag{2.2}$$

sendo $B_0 = (A_2 - A_1 - A_3)/24$, $B_1 = (3A_3 - 2A_2 + A_1)/(24q)$, $B_2 = (A_2 - 3A_3)/[24q(q+2)]$ e $B_3 = A_3/[24q(q+2)(q+4)]$. As quantidades A_1, A_2 e A_3, dadas por Harris, são funções complicadas de ordem n^{-1} dos cumulantes de derivadas do logaritmo da função de verossimilhança.

Cordeiro e Ferrari (a ser publicado) notaram que a forma da função densidade de S_R dada em (2.2), envolvendo um polinômio de terceiro grau dependendo de três constantes, sugere uma estatística escore modificada definida por

$$S_R^* = S_R[1 - (c + bS_R + aS_R^2)],$$

onde a, b e c são funções dos A's. Eles mostraram que é possível determinar a, b e c de tal forma que, até ordem n^{-1}, a função densidade de S_R^* sob H_o seja idêntica a $g_r(x)$. A demonstração baseia-se no cálculo da função geradora de momentos de S_R^* ou segue da fórmula (1) do Teorema 1 de Cox e Reid (1987), retendo termos até ordem n^{-1}. Assim, foi possível mostrar que a estatística escore modificada tendo distribuição χ^2 até ordem n^{-1} é dada por (1.1) com

$$b(S_R, A_1, A_2, A_3, q) = \frac{A_1 - A_2 + A_3}{12q} + \frac{A_2 - 2A_3}{12q(q+2)} S_R + \frac{A_3}{12q(q+2)(q+4)} S_R^2.$$

Como já observado na Seção 1, existe uma estreita conexão entre a estatística modificada S_R^* e a expansão assintótica para os quantis da distribuição de S_R^* deduzida por Harris (1985).

A obtenção de S_R^* depende obviamente do cálculo de A_1, A_2 e A_3. Mas, dependendo do modelo associado aos dados, pode ser bastante difícil calcular estas quantidades. Na Seção 3, apresentamos as expressões algébricas dos A's dadas por Harris e propomos expressões equivalentes, porém em notação matricial, que têm custo computacional reduzido.

3 — FÓRMULAS MATRICIAIS PARA A_1, A_2 E A_3

Inicialmente, vamos introduzir alguma notação. Os subscritos que aparecem nas expressões a seguir variam pelos inteiros $1, \ldots, p$. Sejam $U_r = \partial L/\partial\beta_r$, $U_{rs} = \partial^2 L/\partial\beta_r\partial\beta_s$ e assim por diante. Adotamos a notação padrão para os cumulantes de derivadas do logaritmo da função de verossimilhança: $k_{rs} = E(U_{rs})$, $k_{rst} = E(U_{rst})$, $k_{r,s} = E(U_rU_s)$, $k_{rs,t} = E(U_{rs}U_t)$, $k_{rs,tu} = E(U_{rs}U_{tu}) - k_{rs}k_{tu}$, $k_{r,s,tu} = E(U_rU_sU_{tu}) - k_{r,s}k_{tu}$ e $k_{r,s,t,u} = E(U_rU_sU_tU_u) - k_{r,s}k_{t,u} - k_{r,t}k_{s,u} - k_{r,u}k_{s,t}$. Os elementos da matriz de informação total K são $k_{r,s} = -k_{rs}$ e os de sua inversa são denotados por $k^{r,s}$. Sejam

$$A = \begin{pmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{pmatrix}, \quad M = K^{-1} - A, \tag{3.1}$$

onde K_{22}^{-1} e K^{-1} são obtidos de (2.1). Os elementos (i,j) das matrizes M e A são denotados por m_{ij} e a_{ij}, respectivamente. Denotemos ainda por Σ' e Σ as somas sobre todos os β's e sobre os dados, respectivamente.

Das expressões de Harris (1985) temos

$$\begin{aligned} A_1 = {} & 3\Sigma'(k_{ijk} + 2k_{i,jk})(k_{rst} + 2k_{rs,t})a_{ij}a_{st}m_{kr} \\ & - 6\Sigma'(k_{ijk} + 2k_{i,jk})k_{r,s,t}a_{ij}a_{kr}m_{st} \\ & + 6\Sigma'(k_{i,jk} - k_{i,j,k})(k_{rst} + 2k_{rs,t})a_{js}a_{kt}m_{ir} \\ & - 6\Sigma'(k_{i,j,k,r} + k_{i,j,kr})a_{kr}m_{ij}, \end{aligned} \tag{3.2}$$

$$\begin{aligned} A_2 = {} & -3\Sigma' k_{i,j,k}k_{r,s,t}a_{kr}m_{ij}m_{st} \\ & + 6\Sigma'(k_{ijk} + 2k_{i,jk})k_{r,s,t}a_{ij}m_{kr}m_{st} \\ & - 6\Sigma' k_{i,j,k}k_{r,s,t}a_{kt}m_{ir}m_{js} \\ & + 3\Sigma' k_{i,j,k,r}m_{ij}m_{kr}, \end{aligned} \tag{3.3}$$

$$\begin{aligned} A_3 = {} & 3\Sigma' k_{i,j,k}k_{r,s,t}m_{ij}m_{kr}m_{st} \\ & + 2\Sigma' k_{i,j,k}k_{r,s,t}m_{ir}m_{js}m_{kt}. \end{aligned} \tag{3.4}$$

Note que na expressão de Harris para A_2, o termo $-6K.,.,.^*M^*M^*J^*K.,.,.,.$ está incorreto, devendo ser substituído por $-6K.,.,.^*M^*M^*J^*K.,.,.,.$.

Como se pode observar, as expressões para A_1, A_2 e A_3 são funções complicadas dos cumulantes k's do logaritmo da função de verossimilhança. Cordeiro, Ferrari e Paula (1990) obtiveram fórmulas matriciais para os A's para diversos testes em modelos lineares generalizados (McCullagh e Nelder, 1989) e mostraram como calcular estas quantidades através do sistema GLIM.

A seguir apresentaremos expressões matriciais simples para os A's que podem ser facilmente implantados em programas de computador como o REDUCE, pois envolvem

somente operações de multiplicação e soma de matrizes e vetores. O número total de matrizes envolvidas nos cálculos é $2p^2 + 3p + 13$, sendo $2p^2 + 3p + 6$ matrizes $p \times p$, 2 matrizes $p \times p^2$, 3 matrizes $p^2 \times p$ e 2 matrizes $p^2 \times p^2$.

Denotamos por A_{lm} a m-ésima parcela de A_l. Definimos $b_{jk}^{(i)} = k_{ijk} + 2\,k_{i,jk}$ como o elemento (j,k) da matriz $B^{(i)}$. Nota-se que $B^{(i)}$ é uma matriz $p \times p$ simétrica. Assim, de (3.2),

$$A_{11} = 3\,\Sigma' a_{ij} b_{jk}^{(i)} m_{kr} b_{rs}^{(t)} a_{st},$$

ou ainda,

$$A_{11} = 3 \sum_{i,t} a_i^\top B^{(i)} M\, B^{(t)} a_t,$$

onde $a_i^\top$ é a i-ésima linha da matriz A definida em (3.1). Agora é fácil ver que

$$A_{11} = 3 1^\top F M F^\top 1,$$

onde F é uma matriz $p \times p$ cuja i-ésima linha é $a_i^\top B^{(i)}$ e 1 é um vetor $p \times 1$ de uns.

Para a obtenção de A_{12}, definimos $c_{jk}^{(i)} = k_{i,j,k}$ com o elemento (j,k) da matriz $C^{(i)}$, simétrica e de ordem p. Assim,

$$A_{12} = -6\Sigma' a_{ij} b_{jk}^{(i)} a_{kr} c_{rs}^{(t)} m_{st},$$

ou seja,

$$A_{12} = -6 \sum_{i,t} a_i^\top B^{(i)} A\, C^{(t)} m_t,$$

sendo $m_t^\top$ a t-ésima linha da matriz M definida em (3.1). A expressão acima conduz a

$$A_{12} = -6\, 1^\top F A G^\top 1,$$

onde G é uma matriz $p \times p$ cuja i-ésima linha é $m_i^\top C^{(i)}$.

Definimos $d_{jk}^{(i)} = k_{ijk} - k_{i,j,k}$ como o elemento (j,k) da matriz $D^{(i)}$, simétrica e de ordem p. Assim, A_{13} pode ser escrita como

$$A_{13} = 6\Sigma' m_{ir} b_{rs}^{(t)} a_{sj} d_{jk}^{(i)} a_{kt},$$

ou

$$A_{13} = 6 \sum_{i,t} m_i^\top B^{(t)} A\, D^{(i)} a_t.$$

Sejam

$$J = \begin{pmatrix} m_1^\top B^{(1)} & \dots & m_1^\top B^{(p)} \\ \vdots & & \\ m_p^\top B^{(1)} & \dots & m_p^\top B^{(p)} \end{pmatrix}$$

e

$$H = \begin{pmatrix} D^{(1)}a_1 & \dots & D^{(p)}a_1 \\ \vdots & & \vdots \\ D^{(1)}a_p & \dots & D^{(p)}a_p \end{pmatrix}$$

matrizes $p \times p^2$ e $p^2 \times p$, respectivamente, e

$$\tilde{A} = \begin{pmatrix} A & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & A \end{pmatrix}$$

uma matriz $p^2 \times p^2$ bloco diagonal. Observa-se que o i-ésimo elemento da diagonal da matriz $J\tilde{A}H$ é $\sum_t m_i^\top B^{(t)} A D^{(i)} a_t$ e, portanto,

$$A_{13} = 6 \text{ tr}(J\tilde{A}H),$$

onde tr indica o traço da matriz.

Para a obtenção da última parcela de A_1, definimos o elemento (i,j) da matriz $E^{(k,r)}$, simétrica de ordem p como $e_{ij}^{(k,r)} = k_{i,j,k,r} + k_{i,j,kr}$. Temos

$$A_{14} = -6 \sum_{k,r} a_{kr} \sum_{i,j} m_{ij} e_{ij}^{(kr)}$$

que conduz a

$$A_{14} = -6 \sum_{k,r} a_{kr} \text{ tr}(M E^{(k,r)}).$$

Definindo $l_{ij} = a_{ij} \text{tr}(M E^{(i,j)})$ como o elemento (i,j) da matriz L, simétrica de ordem p, temos

$$A_{14} = -6\, 1^\top L 1.$$

Tem-se, então,

$$A_1 = 3\, 1^\top F M F^\top 1 - 6\, 1^\top F A G^\top 1 + 6 \text{tr}(J\tilde{A}H) - 6\, 1^\top L\, 1, \qquad (3.5)$$

onde as matrizes envolvidas foram definidas acima.

A obtenção de A_2 e A_3 em forma matricial a partir de (3.3) e (3.4) é semelhante à de A_1. Temos

$$A_2 = -3\, 1^\top G A G^\top 1 + 6\, 1^\top F M G^\top 1 - 6 \text{tr}(P\tilde{M}S) + 3\, 1^\top R 1, \qquad (3.6)$$

onde

$$P = \begin{pmatrix} m_1^\top C^{(1)} & \dots & m_1^\top C^{(p)} \\ \vdots & & \vdots \\ m_p^\top C^{(1)} & \dots & m_p^\top C^{(p)} \end{pmatrix}$$

e

$$S = \begin{pmatrix} C^{(1)}a_1 & \dots & C^{(p)}a_1 \\ \vdots & & \vdots \\ C^{(1)}a_p & \dots & C^{(p)}a_p \end{pmatrix}$$

são matrizes de ordens $p \times p^2$ e $p^2 \times p$, respectivamente,

$$\tilde{M} = \begin{pmatrix} M & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & M \end{pmatrix}$$

é uma matriz $p^2 \times p^2$ bloco diagonal e R é uma matriz simétrica de ordem p cujo elemento (k,r) é $r_{kr} = m_{kr}\,\mathrm{tr}(MQ^{(k,r)})$ com $Q^{(k,r)}$ sendo uma matriz também simétrica e de ordem p com elemento (i,j) igual a $q_{ij}^{(k,r)} = k_{i,j,k,r}$. As demais matrizes envolvidas foram definidas anteriormente. Finalmente, temos

$$A_3 = 3\ 1^\top GMG^\top 1 + 2\,\mathrm{tr}(P\tilde{M}T), \tag{3.7}$$

onde

$$T = \begin{pmatrix} C^{(1)}m_1 & \dots & C^{(p)}m_1 \\ \vdots & & \vdots \\ C^{(1)}m_p & \dots & C^{(p)}m_p \end{pmatrix}$$

A seguir fazemos uma análise do custo computacional de obtenção de A_1, A_2 e A_3 através das fórmulas matriciais. Definimos este custo como o número total de produtos envolvidos no cálculo destas quantidades.

Sejam A e B duas matrizes de ordem p. Note que o custo do produto de A por B é p^3, o de seu traço é p^2 e que $1^\top A1 = \sum_{i,j} a_{ij}$, onde a_{ij} é o elemento (i,j) da matriz A, tem custo nulo.

Supomos que as parcelas envolvidas em A_1, A_2 e A_3 são calculadas seqüencialmente na ordem em que elas aparecem nas expressões (3.5) a (3.7). Desta forma, o cálculo do custo de uma determinada parcela levará em conta o fato de que as parcelas anteriores já foram calculadas e, portanto, que os produtos já efetuados podem ser aproveitados sem custo adicional.

Analisemos em detalhes o custo da primeira parcela de A_1, ou seja, $A_{11} = 31^\top FMF^\top 1$. O custo de F é p vezes o custo do produto de $a_1^\top$ por $B^{(1)}$, ou seja, é igual a p^3. Agora, os custos dos produtos de F por M e de FM por $F^\top$ são também iguais a p^3. No cálculo de A_{11} é envolvido ainda mais um produto, o de $1^\top FMF^\top 1$ por 3. Logo, o custo de A_{11} é igual a $3p^3 + 1$.

O custo de $A_{12} = -61^\top FAG^\top 1$ é calculado de forma análoga ao de A_{11}, lembrando que a matriz F já foi obtida e, portanto, que o custo de sua obtenção aqui é nulo. Assim, temos que o custo de A_{12} é $2p^3 + 1$.

Notemos agora que o i-ésimo elemento da diagonal de $J\tilde{A}H$ é $\Sigma_t m_i^\top B^{(t)} A D^{(i)} a_t$ e que o custo de $A_{13} = 6\,\mathrm{tr}(J\tilde{A}H)$ é dado por 1 mais p vezes o custo deste elemento, que envolve os custos dos produtos de $m_i^\top$ por $B^{(t)}$, de $m_i^\top B^{(t)}$ por A, de $D^{(i)}$ por a_t e de $m_i^\top B^{(t)} A$ por $D^{(i)} a_t$. Assim, temos que o custo de A_{13} é igual a $3p^4 + p^3 + 1$.

Finalmente, o custo de $A_{14} = -61^\top L1$ é dado por 1 mais p^2 vezes o custo de $\ell_{ij} = a_{ij}\,\mathrm{tr}(M\,E^{(i,j)})$, ou seja, é igual a $p^4 + p^2 + 1$. Temos então que o custo de A_1 é $4p^4 + 6p^3 + p^2 + 4$.

Os custos de A_2 e A_3 são obtidos de maneira semelhante ao de A_1 e são dados, respectivamente, por $3p^4 + 4p^3 + p^2 + 4$ e $3p^3 + 2$, de forma que o custo total da obtenção de A_1, A_2 e A_3 através das fórmulas matriciais é $7p^4 + 13p^3 + 2p^2 + 10$, isto é, de ordem p^4. O custo total a partir das expressões do Harris (1985) para os A's é $0(p^6)$.

4 — ILUSTRAÇÕES NUMÉRICAS

Consideremos os dados apresentados por Cox e Snell (1981; pgs. 148–150) relativos a tempos de sobrevivência y desde o diagnóstico de 17 pacientes sofrendo de leucemia. Eles consideram um modelo de regressão exponencial com $\mu_i = \exp\{\alpha + \beta(x_i - \bar{x})\}$, onde μ_i é o tempo médio de sobrevivência desde o diagnóstico e x_i é o logaritmo na base 10 da contagem inicial de células brancas. O interesse é testar a hipótese nula $H_0 : \beta = 0$ *versus* a alternativa $H : \beta \neq 0$. McCullagh e Nelder (1989; pg. 464) consideram a estatística da razão de verossimilhança w e sua versão modificada $w^* = w(1 - b)$, onde $1 - b$ é o fator de correção de Bartlett, para o teste destas hipóteses. Os valores de ambas as estatísticas indicam que H_0 é rejeitado ao nível nominal de 1%. Cordeiro, Ferrari e Paula (1990) obtiveram S_R e os A's para o teste de H_0 *versus* H. Denotando $\bar{s}_a = n^{-1}\Sigma(x_i - \bar{x})^a$, $a = 2,3,4$, e $\bar{s} = n^{-1}(\Sigma x_i y_i - n\bar{x}\bar{y})$ mostra-se que $S_R = n\,\bar{s}^2/(\bar{y}^2\bar{s}_2)$, $A_1 = -12/n$, $A_2 = -18(2 - \bar{s}_4/\bar{s}_2^2)/n$ e $A_3 = 20\bar{s}_3^2/(n\bar{s}_3^3)$, o que conduz aos valores $S_R = 5,681$, $A_1 = 0,7058$, $A_2 = 0,1803$ e $A_3 = 0,0008$. Através da expressão (1.2) calcula-se o valor crítico corrigido correspondente ao nível nominal de 1% e obtem-se o valor 6,365, enquanto que o não-corrigido obtido diretamente da distribuição χ^2 com 1 grau de liberdade vale 6,635.

A seguir mostraremos que A_1, A_2 e A_3 podem ser obtidos das expressões matriciais (3.5)–(3.7) sem qualquer dificuldade.

O logaritmo da função de verossimilh nça de α e β é dado por

$$L = \quad n\alpha \quad \exp(=\alpha)\Sigma y_i \exp\{-\beta(x_i - x)\}.$$

Devemos montar 21 matrizes: 14 de dimensão 2×2, 2 de dimensão 2×4, 3 de dimensão 4×2 e 2 de dimensão 4×4. Através das propriedades da distribuição exponen-

cial, os cumulantes de certas derivadas do logaritmo da função de verossimilhança são calculados a estes conduzem a

$$A = \frac{1}{n}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad M = \frac{1}{n}\begin{pmatrix} \bar{s}_2^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad B^{(1)} = -n\begin{pmatrix} \bar{s}_3 & \bar{s}_2 \\ \bar{s}_2 & 0 \end{pmatrix},$$

$$B^{(2)} = -n\begin{pmatrix} \bar{s}_2 & 0 \\ 0 & 1 \end{pmatrix}, \quad F = -\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad G = 2\begin{pmatrix} \bar{s}_3/\bar{s}_2 & 1 \\ 0 & 0 \end{pmatrix},$$

$$J = -\begin{pmatrix} \bar{s}_3/\bar{s}_2 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad E^{(2,2)} = 4n\begin{pmatrix} \bar{s}_2 & 0 \\ 0 & 1 \end{pmatrix},$$

$$L = \frac{4}{n}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad P = 2\begin{pmatrix} \bar{s}_3/\bar{s}_2 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$Q^{(1,1)} = 6n\begin{pmatrix} \bar{s}_4 & \bar{s}_3 \\ \bar{s}_3 & \bar{s}_2 \end{pmatrix}, \quad R = 6\begin{pmatrix} \bar{s}_4/\bar{s}_2^2 & 0 \\ 0 & 0 \end{pmatrix},$$

$$H = -3\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad S = 2\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \bar{s}_2 & 0 \\ 0 & 1 \end{pmatrix},$$

$$T = 2\begin{pmatrix} \bar{s}_2/\bar{s}_2 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{M} = \frac{1}{n}\begin{pmatrix} \bar{s}_2^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{s}_2^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\tilde{A} = \frac{1}{n}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad C^{(i)} = -2B^{(i)} \text{ e } D^{(i)} = 3B^{(i)}, i = 1,2.$$

Note que as matrizes $E^{(1,1)}$, $E^{(1,2)}$, $E^{(2,1)}$, $Q^{(1,2)}$, $Q^{(2,1)}$, e $Q^{(2,2)}$ não precisam ser calculadas, pois estas serão multiplicadas por elementos de M ou A que valem zero.

De (3.5)–(3.7) e das matrizes acima mostra-se facilmente que as expressões para A_1, A_2 e A_3 coincidem com as obtidas por Cordeiro, Ferrari e Paula (1990). De (1.1), temos $S_R^* = 5{,}939$ o que conduz a um nível descritivo, obtido diretamente da distribuição χ^2 com 1 grau de liberdade, de 1,48%. Note que o nível descritivo não-corrigido do teste baseado na estatística S_R é de 1,71%. Assim, embora a decisão, ao nível de significância de 1% não seja alterada se usarmos a estatística S_R^* em lugar de S_R, o teste baseado em S_R^* é o que indica maior evidência contra H_0.

Consideremos agora os dados de Jorgensen (1961) referentes ao número de falhas por semana (y) de uma peça complexa de um equipamento eletrônico. A peça foi analisada por 9 semanas divididas em 2 regimes de operação. As variáveis auxiliares x e u representam os tempos gastos nos regimes 1 e 2, respectivamente. Jorgensen assumiu uma distribuição de Poisson para y_i com média $\mu_i = \beta_1 x_i + \beta_2 u_i$, além de assumir independência das observações. O interesse é testar $H_{o1} : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ e $H_{o2} : \beta_2 = 0$ *versus* $H_2 : \beta_2 \neq 0$. Este exemplo foi analisado por Wang (1985), que considerou a estatística escore e sua versão gráfica para testar as hipóteses acima.

Neste exemplo, utilizamos a notação

$$r_i = x_i - u_i \frac{\bar{x}}{\bar{u}}, \qquad \bar{s}_1 = \frac{1}{n}\Sigma \frac{x_i y_i}{u_i}, \qquad \bar{s}_2 = \frac{1}{n}\Sigma \frac{x_i^2}{u_i}$$

$$\bar{s}_3 = \frac{1}{n}\Sigma \frac{r_i^2}{u_i}, \quad \bar{s}_4 = \frac{1}{n}\Sigma \frac{r_i^4}{u_i^3} \quad \text{e} \quad \bar{s}_5 = \frac{1}{n}\Sigma \frac{r_i^3}{u_i^2}.$$

O logaritmo da função de verossimilhança é dado por:

$$L = -\Sigma(\beta_1 x_i + \beta_2 u_i) + \Sigma y_i \log(\beta_i x_i + \beta_2 u_i) - \Sigma \log y_i!$$

o que conduz ao estimador de máxima verossimilhança de β_2 sob H_{o1} dado por $\tilde{\beta}_2 = \bar{y}\bar{u}^{-1}$. A estatística escore para testar H_{o1} *versus* H_1 é

$$S_R = \frac{n\bar{u}^2}{\bar{y}(\bar{u}\bar{s}_2 - \bar{x}^2)}(\bar{s}_1 - \bar{y}\bar{x}/\bar{u})^2.$$

De (3.5)–(3.7) deduz-se, sem dificuldade, que $A_1 = 0$ e A_2 e A_3 são funções do parâmetro de incômodo β_2 que, substituído por $\tilde{\beta}_2$, leva a

$$\tilde{A}_2 = \frac{3\bar{u}^3}{n(\bar{u}\bar{s}_2 - \bar{x}^2)^2\bar{y}}(\bar{s}^2/\bar{u} + \bar{s}_4)$$

e

$$\tilde{A}_3 = \frac{5\bar{u}^4\bar{s}_5^2}{n(\bar{u}\bar{s}_2 - \bar{x}^2)^3\bar{y}}.$$

O cálculo de S_R com os dados de Jorgensen conduz ao valor 39,96 que é altamente significante. Note que, neste caso, não há necessidade prática de se calcular S_R^* dada a alta significância de S_R. Para o teste da hipótese H_{o2} *versus* H_2, $S_R = 1,831$, $A_1 = 0$, $\tilde{A}_2 = -8,948 \times 10^{-3}$, $\tilde{A}_3 = 4,558 \times 10^{-3}$ e $S_R^* = 1,830$.

Observa-se, então, que S_R e S_R^* são bastante próximos de forma que, neste exemplo, a correção da estatística é desprezível.

5 — ALGUMAS PROPRIEDADES DE S_R^*

Algumas propriedades da estatística escore modificada S_R^* em (1.1) merecem ser apontadas. Em primeiro lugar, notamos que, enquanto S_R é uma variável aleatória não negativa, S_R^* pode assumir valores menores do que zero. Outro aspecto a ser mencionado é que S_R^* é uma função crescente de S_R se e somente se $A_2 = A_3 = 0$. Estas parecem ser características indesejáveis para a estatística S_R^*. A seguir passamos a analisar estes pontos com maior profundidade. Para tal, mostraremos primeiramente que a quantidade A_3 em (3.4) é não negativa. Este é um fato importante para a análise que se segue.

Suponhamos que $K = nI$, onde K é a matriz de informação total definida em (2.1) e I é a matriz identidade de dimensão $p \times p$. De (3.1) temos que os elementos de M são $m_{ij} = n^{-1}$ se $i = j$ e $i \leq q$ e $m_{ij} = 0$ caso contrário. Agora, de (3.4) temos que

$$A_3 = n^{-3}\left\{3\sum_{r=1}^{q}\left(\sum_{i=1}^{q}k_{r,i,i}\right)^2 + 2\sum_{i,j,r=1}^{q}k_{i,j,r}^2\right\}.$$

Logo, $A_3 \geq 0$ se $K = nI$. Seguindo a argumentação de Cordeiro (1987; pg. 268) temos que, como K é positiva definida, existe uma matriz P não singular $p \times p$ de constantes desconhecidas tal que $P^\top K^{-1} P = n^{-1} I$ sendo K avaliada no verdadeiro valor do vetor de parâmetros β. Então, a matriz de informação do vetor paramétrico $\beta^* = P^\top \beta$ é n vezes a matriz identidade. Portanto, sem perda de generalidade, podemos assumir que $K = nI$ e, portanto, $A_3 \geq 0$.

Passamos a analisar o fato de que S_R^* pode assumir valores negativos. Lembremos que os A's são de ordem n^{-1} e suponhamos, inicialmente, que $A_3 = 0$. Se $A_2 = 0$, temos, de (1.1), que $S_R^* \geq 0$ se e somente se $1 - A_1/(12q) \geq 0$. Se $A_1 \leq 0$, esta condição está satisfeita. O mesmo ocorre se $A_1 > 0$ e se n for suficientemente grande. Se $A_2 > 0$, é fácil de ver que $S_R^* \geq 0$ se e somente se $S_R \leq L_1$, onde

$$L_1 = \frac{12q(q+2)}{A_2} + (q+2)\left(1 - \frac{A_1}{A_2}\right).$$

Negligenciando a parcela de ordem 1, temos $L_1 \simeq 12q(q+2)/A_2$. Lembrando que S_R tem distribuição assintótica χ^2 com q graus de liberdade sob H_o, notamos que, para n grande, se $S_R > 12q(q+2)/A_2$, temos forte evidência em favor da hipótese alternativa. Neste caso, não haveria necessidade prática de se calcular a estatística corrigida. A metodologia de correção só se faz necessária quando existe uma possibilidade razoável de aceitar H_o. Um forte indicativo de rejeição de H_o torna a estratégia de correção inútil do ponto de vista prático. Em resumo, se $A_3 = 0$ e $A_2 > 0$ e se n for suficientemente grande, as amostras que não evidenciam forte rejeição de H_o produzem valores de S_R^* não negativos.

Suponhamos agora que $A_3 = 0$ e $A_2 < 0$. De (1.1), temos que $S_R^* \geq 0$ se e somente se $S_R \geq L_1$. Mas $L_1 \leq 0$ para n suficientemente grande e, neste caso, a condição necessária e suficiente é satisfeita.

Vejamos o que ocorre se $A_3 > 0$. De (1.1), é fácil ver que $S_R^* \geq 0$ se e somente se $LI \leq S_R \leq LS$ onde $LI = a_1 - a_2$ e $LS = a_1 + a_2$ com

$$a_1 = (q+4)\left(1 - \frac{A_2}{2A_3}\right)$$

e

$$a_2 = \left\{ \frac{12q(q+2)(q+4)}{A_3} + (q+4)^2 \left(1 - \frac{A_2}{2A_3}\right)^2 + (q+2)(q+4)\left(\frac{A_2 - A_1}{A_3} - 1\right) \right\}^{1/2}$$

Note que a expressão entre chaves em a_2 é positiva para n suficientemente grande. Negligenciando a_1, que é de ordem 1 e as parcelas desta ordem entre as chaves de a_2, temos que $LI \simeq -a_3$ e $LS \simeq a_3$, onde $a_3 = [12q(q+2)(q+4)/A_3]^{1/2}$. Assim, para n suficientemente grande, a condição necessária e suficiente para termos $S_R^* \geq 0$ se reduz a $S_R \leq a_3$. Mas, se S_R não satisfaz tal condição, há forte evidência contra H_o de forma que não há necessidade de se calcular S_R^*.

Da análise acima concluímos que, do ponto de vista prático e desde que o tamanho da amostra não seja muito pequeno, o fato de S_R^* poder assumir valores negativos não apresenta nenhum problema, pois só ocorre em situações em que a evidência contra H_o é bastante forte, fazendo com que não seja necessária a utilização da estratégia de correção da estatística.

Passamos a analisar agora outra característica de S_R^* aparentemente indesejável, ou seja, o fato de S_R^* não ser necessariamente uma função crescente de S_R. Propomos estudar a expressão em (1.1) como função de S_R e determinar para que valores de S_R a primeira derivada desta função é positiva. Este estudo é análogo à análise de positividade de S_R^* e, portanto, será omitido. A conclusão a que se chega é novamente a de que, se o tamanho da amostra não for muito pequeno, S_R^* não será uma função crescente de S_R somente em situações em que a evidência contra H_o é bastante forte, ou seja, quando a estratégia de correção da estatística não é necessária do ponto de vista prático.

NOTA

Este artigo é baseado em parte da tese de doutoramento de Silvia L. de Paula Ferrari a ser submetida ao Instituto de Matemática e Estatística da Universidade de São Paulo, sob orientação de Gauss M. Cordeiro. Este trabalho foi financiado parcialmente pelo CNPq/Brasil.

BIBLIOGRAFIA

CORDEIRO, G.M. (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, 74, 265–274.

CORDEIRO, G.M. e FERRARI, S.L.P. (a ser publicado). A modified score test statistic having chi-squared distribution to order n^{-1}. A aparecer na *Biometrika*, 78.

CORDEIRO, G.M., FERRARI, S.L.P. e PAULA, G.A. (1990). Improved score tests for generalized linear models. *Relatório Técnico RT-MAE-9016* IME-USP e submetido ao *J. R. Statist. Soc.* B.

COX, D.R. e HINKLEY, D.V. (1979). *Theoretical Statistics*. Wiley, New York.

COX, D.R. e REID, N. (1987). Approximations to noncentral distributions. *Canad. J. Statist.*, 15, 105–114.

COX, D.R. e SNELL, E.J. (1981). *Applied Statistic*. Chapman and Hall, London.

HARRIS, P. (1985). An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, 72, 653–659.

HAYAKAWA, T. (1977). The likelihood ratio criterion and the asymptotic expansion of its distribution. *Ann. Inst. Statist. Math.*, 29, 359–378.

JORGENSEN, D.W. (1961). Multiple regression analysis of a Poisson process. *J. Amer. Statist. Assoc.*, 56, 235–245.

LAWLEY, D.N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43, 295–303.

McCULLAGH, P. e NELDER, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

RAO, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philosophical Soc.*, 44, 50–57.

WANG, P.C. (1985). Adding a variable in generalized linear models. *Technometrics*, 27, 273–276.

RESUMO

Apresentam-se fórmulas matriciais para calcular fatores de correção para a estatística escore no teste de uma hipótese nula geral composta. Mostra-se que as fórmulas matriciais têm vantagem computacional em relação à notação tensorial da expansão assintótica de Harris (1985) para a distribuição da estatística escore. Combinando essas fórmulas com o uso de um sistema algébrico computadorizado como o REDUCE tem-se a maneira mais fácil de se calcular a expansão assintótica da distribuição da estatística escore. Apresentam-se algumas propriedades de uma estatística escore modificada, que tem distribuição qui-quadrado até ordem n^{-1}, onde n é o tamanho da amostra, sendo dois exemplos numéricos ilustrados.

ABSTRACT

We give simple matrix formulae for computing Bartlett correction factors for the score statistic for testing a general composite null hypothesis. We show that the matrix formulae have computational advantage over the tensor notation of Harris' (1985) asymptotic expansion for the distribution of the score statistic. These formulae combined with the use of algebraic computerized system such as REDUCE provide the easiest way to compute the asymptotic expansion for the distribution of the score statistic. We present some properties of a modified score statistic, which has chi-squared distribution to order n^{-1}, where n is the sample size, and two numerical examples are illustrated.

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) análise de dados;
j) crítica e imputação de dados;
l) disseminação e confiabilidade de dados;
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:

Djalma G. C. Pessoa
Editor Responsável – RBEs
ENCE
Rua André Cavalcanti, 106
Bairro de Fátima
20231 – 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) autor(es), 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 parágrafo, sem fórmulas, com no máximo 150 palavras.

3. O artigo deve ser dividido 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 tal 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 títulos.

7. Serão aceitos originais processados por editores de texto tais como CW, Word, Carta Certa, WP e WS.

Impresso no CDDI/Departamento de Editoração e Gráfica — IBGE, em fevereiro de 1992, O. S. 658-20/91

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, econômica e territorial do País.

VOCÊ PODE OBTER ESSAS PESQUISAS, ESTUDOS E LEVANTAMENTOS EM TODO O PAÍS

No Rio de Janeiro procure o
Núcleo de Atendimento Integrado - NAT do
Centro de Documentação e Disseminação de
Informações - CDDI

Rua General Canabarro, 666
CEP 20271 - Maracanã - Rio de Janeiro - RJ
Tel.: (021)284-0402
Telex: 2134128 - Fax: (021)234-6189

Nos Estados procure o
Setor de Documentação e Disseminação de Informações - SDDI
dos Escritórios Estaduais

O IBGE possui, ainda, agências localizadas nos
principais Municípios.