

Presidente da República
José Sarney

Ministro-Chefe da Secretaria de Planejamento e Coordenação
João Batista de Abreu

Secretário-Geral
Ricardo Luís Santiago

**FUNDAÇÃO
INSTITUTO BRASILEIRO
DE GEOGRAFIA
E ESTATÍSTICA – IBGE**

Presidente
Charles Curt Mueller

Diretor-Geral
David Wu Tai

Diretor de Pesquisas
Lenildo Fernandes Silva

Diretor de Geociências
Mauro Pereira de Mello

Diretor de Informática
José Sant'Anna Bevilaqua

**REVISTA BRASILEIRA
DE ESTATÍSTICA**

Editor Responsável
Djalma Galvão Carneiro Pessoa
Conselho Editorial

Cláudio Considera
Diretoria de Pesquisas

Elisa Caillaux
Diretoria de Pesquisas

Hélio Migon
Escola Nacional de Ciências Estatísticas

Kaizô Beltrão
Escola Nacional de Ciências Estatísticas

Marilourdes Lopes Ferreira
Diretoria de Geociências

Pedro Luis N. Silva
Diretoria de Pesquisas

Victor Hugo Gouvêa
Diretoria de Pesquisas

Valéria da Motta Leite
Diretoria de Pesquisas

SECRETARIA DE PLANEJAMENTO E COORDENAÇÃO DA PRESIDÊNCIA DA REPÚBLICA
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA — IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175

R. bras. Estat. Rio de Janeiro, v. 49, n.º 191, p. 1-127, jan./jun. 1988.

REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE e da Sociedade Brasileira de Estatística

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 Cavalcanti, 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.
- Os pedidos de assinatura e número avulso ou atrasado devem ser endereçados para:

Centro de Documentação e Disseminação de Informações
Av. Beira Mar, 436 – 6º andar – Centro
Rio de Janeiro – RJ – CEP: 20 021
Tel.: (021) 533-3094

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

Criação: Programação Visual e Capa
Pedro Paulo Machado

© IBGE

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística – ano 1, n. 1 (jan./mar., 1940)-
Rio de Janeiro : IBGE, 1940-
Trimestral. (1940-1986), semestral (1987-)
Órgão oficial do IBGE e da Sociedade Brasileira de Estatística.
Continuação de : Revista de economia e estatística
Sumários e índices acumulados de autor — assunto: v.1/40
(1940) no v. 43, n. 169 (jan./mar. 1982)
ISSN 0034-7175 = Revista brasileira de estatística
1. Estatística — Periódicos. I. IBGE

IBGE. Gerência de Documentação e Biblioteca
RJ-IBGE/88-05

CDU 31(05)

SUMÁRIO

ARTIGOS

OUTLIERS E ROBUSTEZ — 5

Oscar H. Bustos

UMA NOVA METODOLOGIA PARA AVALIAÇÃO DE IMÓVEIS UTILIZANDO MODELOS LINEARES GENERALIZADOS — 27

Rubens Alves Dantas
Gauss Moutinho Cordeiro

APLICAÇÃO DE ANÁLISE MULTIVARIADA A UM PROBLEMA DE DOSAGEM DE FÓSFORO PARA OVINOS — 47

Lúcia Silva Kubrusly
Dorinha M. S. S. V. Marcondes
Cláudio T. Bornstein

PESQUISA AGROPECUÁRIA DO PARANÁ SAFRA 1986/87 — 55

Charles Curt Mueller
Gil Silva
Álvaro Gonzales Villalobos

ESTUDO DAS FERRAMENTAS PARA APURAÇÃO DE DADOS — 85

Dulce Maria Rocha Barbosa
Reina Marta Hanono

CRÍTICA DE RAZÕES NO CENSO INDUSTRIAL DE 1985 — 101

José Carlos C. Pinheiro
Renato Martins Assunção

RESENHAS BIBLIOGRÁFICAS

OUTLIERS IN STATISTICAL DATA — 119

CLASSIFICATION AND REGRESSION TREES — 121

INSTRUÇÕES BÁSICAS PARA PREPARO DOS ORIGINAIS — 125

ISSN 0034-7175

OUTLIERS E ROBUSTEZ*

Oscar H. Bustos (IMPA/CNPq)**

1 — O QUE É UM OUTLIER?

Quando realizamos uma experiência e medimos uma certa variável, nunca podemos estar absolutamente seguros de que um determinado dado seja realmente uma manifestação genuína do fenômeno que queremos estudar. Sem dúvida, a confiabilidade provável de uma observação se baseia em sua relação com o conjunto das outras observações dessa variável realizadas em circunstâncias similares. Assim, um *outlier* dentro de um conjunto de dados será uma observação (ou conjunto de observações) *que parece ser inconsistente com a maioria dos dados do conjunto*. A frase grifada é de vital importância. Destaca o aspecto mais importante: uma observação ser ou não considerada como *outlier* é algo dependente do juízo subjetivo do analista.

A preocupação com *outliers* é antiga. Sua história remonta às primeiras tentativas realizadas pelo homem para basear suas de-

cisões sobre inferências extraídas de um conjunto de dados obtidos ao observar um certo fenômeno. Uma visão rápida sobre a literatura estatística nos permite ver que há 200 anos atrás, a prática de descartar observações que parecem anômalas ou discordantes com a massa de dados era coisa comum.

Poderíamos esperar que depois de tanto tempo houvésssemos chegado a um entendimento e clareza objetiva sobre este assunto, e que os métodos disponíveis agora para tratar com *outliers* fossem fundamentados sobre sólidas bases objetivas, aceitas por todos. A verdade é que estamos bem afastados de satisfazer tal esperança. Apesar do muito que se tem escrito sobre este assunto, a noção de *outlier* permanece tão vagamente assentada como o era no Século XVIII.

É certo que têm surgido e são empregados numerosos procedimentos objetivos para manejar *outliers*. Não obstante, sua aplicação se faz, geralmente, depois que uma mera inspeção visual dos dados conduzir a uma identificação *subjetiva* da presen-

* Nota do editor. Escrito a convite, este artigo objetiva oferecer uma retrospectiva sobre *Outliers* e *Robustez* e será continuado nos dois números subsequentes desta Revista.

** Pesquisador do IMPA — Instituto de Matemática Pura e Aplicada.

ça de tais pontos. Isto é, a decisão de empregar um critério objetivo está frequentemente baseada sobre uma reação inicial subjetiva a respeito dos dados.

Para fixar idéias, consideremos o seguinte exemplo extraído de Beckman e Cook (1983). Suponhamos que as seguintes 10 observações são consideradas como realizações independentes de uma variável aleatória comum com distribuição normal:

$$\begin{array}{cccccc} -1.64, & -1.33, & -1.10, & -.57, & -.27 & \\ & & & & & (1.1) \\ 1.04, & 1.56, & 1.84, & 2.04, & 4.99 & \end{array}$$

Entre estas observações, existem *outliers*? E se os há, quais podem ser consideradas como tais? Para responder a estas questões ajudemo-nos com um gráfico:

— * — * — * — * — * — * — * — * — * —

A partir deste, parece que a observação 4.99 é demasiadamente grande, e alguém poderia também questionar a validade da segunda maior observação. As respostas definitivas às questões propostas, poderíamos adiá-las até depois da realização de um teste de detecção e/ou identificação de *outliers*. De qualquer maneira, em geral, a decisão de empregar ou não um tal teste depende dessa primeira reação subjetiva sobre a validade da maior (ou das maiores) observação. Mais adiante voltaremos a este exemplo.

Collet e Lewis (1976) dão a conhecer um interessante estudo sobre a subjetividade na declaração de *outliers*. Desse trabalho eles concluíram que o fato de declarar ou não um ponto (ou conjunto de pontos) como *outlier* depende de diversos fatores: apresentação dos dados (se é realizada por meio de gráficos, ou se a amostra está ordenada, etc.), experiência do analista e até de insuspeitados fatores psicológicos.

Hawkins (1980) levanta objeção a esta conclusão, por serem os "analistas" consultados por Collet e Lewis estatísticos de pouca ou nenhuma experiência. Ele contrapõe a tal conclusão o resultado de um excelente trabalho realizado por Relles e Rogers (1977): os estatísticos de experiência podem ser considerados como testes de detecção de *outliers*, bastante eficientes.

De qualquer forma, não podemos esperar que a maioria dos laboratórios estatísticos contem com um plantel de especialistas de alto nível aos quais podemos submeter qualquer conjunto de dados que chegue para ser analisado. Mas o que é mais importante para substanciar a dúvida sobre a validade de nos conformarmos com o juízo subjetivo de um analista, por mais experiência que ele tenha, é que todos esses estudos se referiam a conjuntos de dados univariados de tamanhos entre pequeno e moderado. Em conjuntos de dados tão grandes como os que aparecem atualmente, devido à facilidade de processamento brindada pelos computadores, ou em contextos mais complicados como os de análise de regressão, análise multivariada, desenho de experimentos, séries temporais, etc., uma inspeção visual valiosa pode ser praticamente impossível.

Surge assim a imperiosa necessidade de dispormos de critérios objetivos para manejar *outliers*. Muitos desses critérios, atualmente em uso, estão baseados sobre modelos probabilísticos, assumidos *a priori* do registro das observações que supostamente deveriam explicar o mecanismo de geração dos dados. A este respeito, existe certa tendência perigosa, pelo menos dentro da Estatística Paramétrica "clássica", que consiste em dar uma grande relevância ao processamento de dados orientado para a estimação de parâmetros ou comprovação de hipótese sobre eles, relegando a um segundo plano a investigação sobre se os dados apóiam ou não as considerações de índole extra-estatística que levaram à adoção de um determinado modelo.

Com referência ao assinalado no último parágrafo, é bem ilustrativo o que dizem Tiku e colaboradores em Tiku, Tan e Balakrishnan (1986). Citam ali Gore e outros (1977), os quais afirmam que 32 de 62 informes publicados entre janeiro e março de 1976 no British Medical Journal continham erros no uso de estatísticas, pelo fato de usarem um modelo sem uma crítica prévia. Também citam o trabalho de Gibbons e Freund (1980), onde se destaca que, em 28 de 30 trabalhos publicados nos volumes correspondentes a março e abril de 1978 do Agronomy Journal, a metodologia

estatística usada estava pobremente documentada, ou era incompleta e até em alguns casos totalmente inadequada.

Outro exemplo bem recente de tipos de erros a que pode conduzir essa prática de descartar dados por não concordarem com o subjetivamente "esperado", o temos no porque os cientistas da NASA não perceberam a falta de ozônio nas camadas superiores da atmosfera terrestre, apesar de esta falta estar manifestada nos dados enviados pelos satélites. Podemos ler o seguinte na pág. 42 da *Time* (1987), 19 de outubro:

"For several years NASA's scientists failed to accept data on the Antarctic ozone hole that was before their eyes. The reason: computers prescreening data from monitoring satellites had been programmed to dismiss as 'suspicious presumably wild data' showing a 30% or greater drop in ozone levels. After British scientists reported the deficit in 1985, NASA went back to its computer records, finally recognizing that the satellite data had been showing the hole all along..."

(Tradução aproximada: "Durante vários anos os cientistas da NASA erraram em aceitar os dados a respeito do "buraco" de ozônio sobre a Antártida que estava diante de seus olhos. A razão: os computadores que processam os dados provenientes de satélites haviam sido programados para descartar como dados suspeitos, presumivelmente "erros grosseiros", aqueles que mostrassem uns 30% ou mais de falta de ozônio (com relação ao nível que se supunha *a priori*). Depois que (certos) cientistas britânicos deram a conhecer o déficit em 1985, a NASA revisou registros gravados pelos computadores, reconhecendo finalmente que os dados dos satélites já haviam mostrado esse "buraco").

Vemos assim que, além de critérios objetivos para tratar *outliers*, necessitamos de uma bateria de técnicas para assegurar a concordância entre um conjunto de dados e o modelo probabilístico que supostamente se ajusta aos dados. Neste sentido, julgamos muito bem-vinda a revitalização do estudo de técnicas robustas que permitam tratar os dados com um maior grau de confiança sobre um espectro mais amplo de modelos probabilísticos que o usualmente permitido pela estatística "clássica".

Antes de passarmos ao estudo e estabelecimento de critérios objetivos sobre *outliers*, devemos deixar claro o significado de diversos termos aparentemente sinônimos: *outliers*, "observações contaminantes", "observações extremas", etc.

Já neste ponto enfrentamos a falta de acordo entre os estatísticos. Beckman e Cook (1983) adotam as seguintes definições:

"Observação discordante": qualquer observação que apareça como surpreendente ou discrepante da maioria, a juízo do analista.

"Observação contaminante": uma que não seja realização da população probabilística que se suponha ajustar aos dados.

Outlier: um termo coletivo que se refira tanto a uma observação discordante como a uma contaminante.

Por outro lado, Barnett (1983), ao discutir o trabalho recém-citado de Beckman e Cook, faz uma breve, mas notável digressão em torno da nomenclatura a ser usada. Destaca, em primeiro lugar, a falta de precedentes firmes sobre os quais devesse basear uma terminologia correta. A seguir, propõe uma base que julga razoável para distinguir os diferentes termos conforme as diversas necessidades. Para Barnett, e também para Barnett e Lewis (1984), um *outlier* é uma observação (ou conjunto de observações) que parece surpreendentemente discrepante das restantes, à vista do analista. Este elemento de "surpresa" leva implícito um juízo sobre qual deve ser a distribuição que gera a amostra (distribuição-alvo). Neste sentido, podemos dizer que uma técnica de identificação de *outliers* será um método objetivo para definir "observação surpreendente". Seguindo com a terminologia, para Barnett, "extremo" em dados univariados é uma observação com o menor (ou maior) valor. Por último, "contaminante" tem o mesmo significado que para Beckman e Cook (1983).

De nossa parte, adotaremos a nomenclatura sugerida no parágrafo anterior.

Agora vamos formalizar em alguma medida o já dito.

Suponhamos ter uma amostra $x_1, \dots, x_n$ de uma distribuição F . Seja $x_{(1)} \leq \dots \leq x_{(n)}$ a amostra ordenada. As observações $x_{(1)}$ e $x_{(n)}$

são "extremos". Elas serão *outliers* conforme a sua relação com a distribuição F (veja a Figura 1).

Suponhamos agora que todas as observações provêm de F exceto duas, que provêm de outra distribuição G; estas serão as "contaminantes". Elas poderiam ser "extremos" ou não, *outliers* ou não (veja a Figura 2).

A discussão em torno da nomenclatura não é banal. É preciso ter bem clara a natureza de cada observação (quer dizer, como poderia haver sido gerada) para saber como tratá-la.

Quando trabalhamos com dados estruturados, como no caso das análises de regressão, aparece um elemento a mais: as chamadas "observações influentes". Voltaremos sobre este assunto mais adiante, mas desde agora devemos deixar bem assentado que uma observação influente, por mais extrema que seja, pode não ser contaminante se a experiência tiver sido planejada para que esse ponto aparecesse onde está.

2 — POR QUE ESTUDAR OS OUTLIERS?

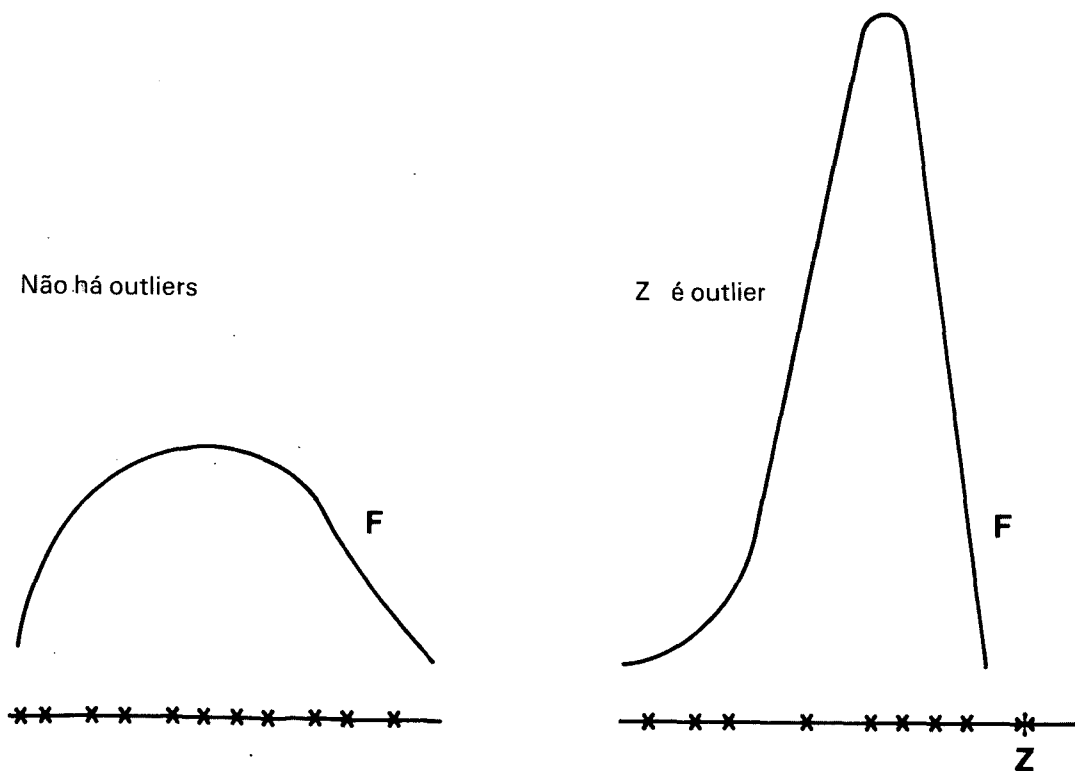
Uma resposta óbvia já foi adiantada no início da seção anterior: a presença de *outliers* é, sob um ponto de vista realista, (quase) inevitável. Esta é possivelmente a razão mais importante que tem motivado o estudo deste tema ao longo da história da Estatística. Necessitamos de técnicas objetivas para o manejo de dados.

De qualquer modo, a partir de um certo tempo, têm-se manifestado outras razões igualmente importantes. Vejamos algumas delas.

2.1 — Interesse nos outliers por si mesmos

Existem situações onde o problema estatístico de interesse para o analista consiste em obter inferências válidas acerca das observações *outliers* e não tanto em estimar os parâmetros da população-alvo.

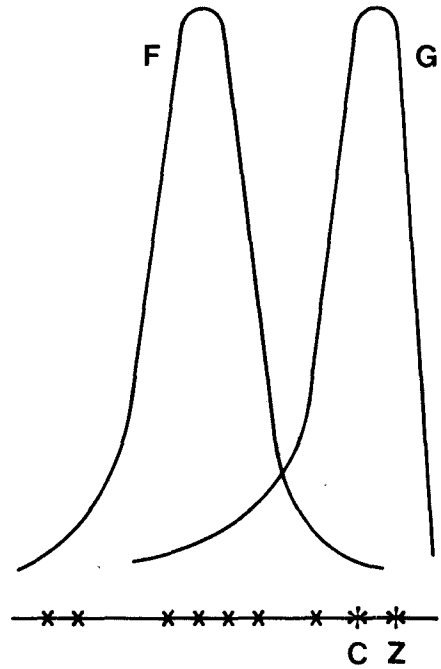
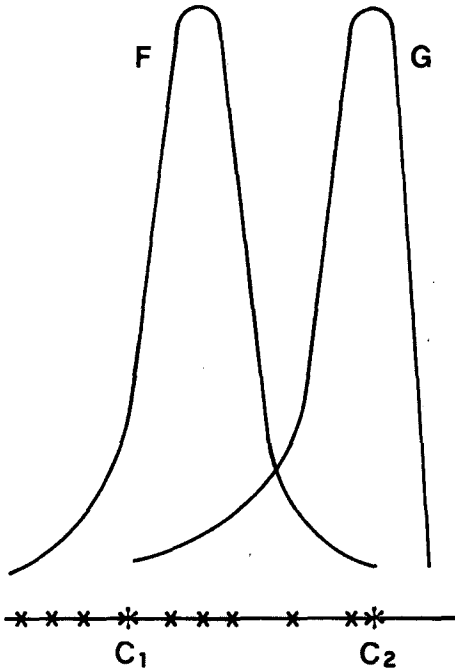
FIGURA 1



C_1 é contaminante
(não-outlier)
 C_2 é contaminante,
extremo (não-outlier)

FIGURA 2

Z é extremo outlier
não contaminante
C não extremo outlier,
contaminante

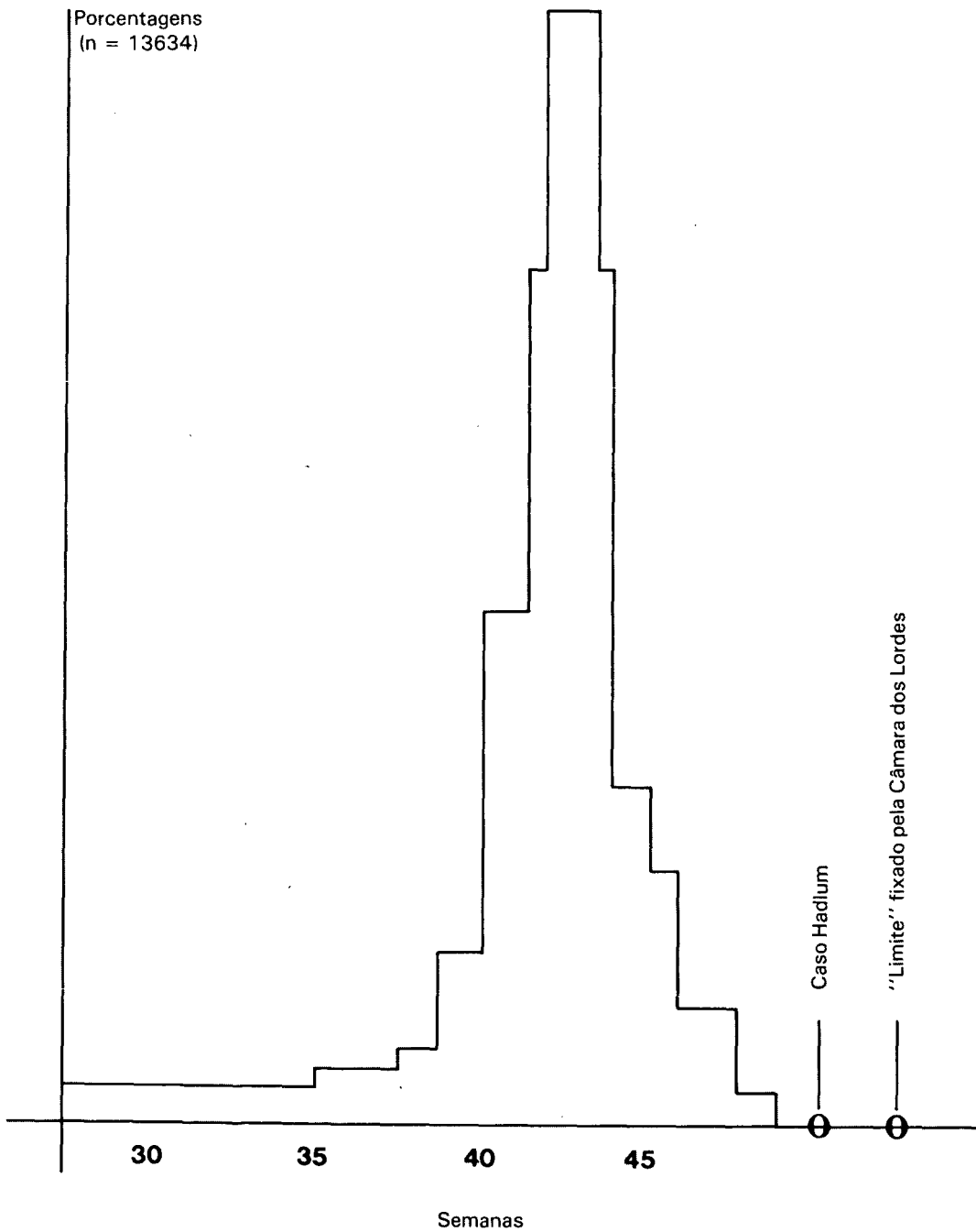


A título de exemplo (um tanto engraçado), vejamos o seguinte caso citado, entre outros, por Barnett e Lewis (1984). Refere-se a um pleito judicial que teve lugar na Grã-Bretanha em 1949, e que foi denominado Hadlum vs. Hadlum. O Sr. Hadlum apresentou uma demanda de divórcio acusando a sua mulher de adultério. Para sustentar tal acusação aduziu que sua mulher havia dado à luz 349 dias depois que ele deixara o país para ir à guerra. Como a média do período de gestação é de 280 dias ao Sr. Hadlum pareceu que 349 era surpreendentemente grande, e assim o classificou como outlier. De nenhuma maneira ele queria que sua observação fosse “desprezada” (descartada ou ignorada). Pelo contrário, pretendia que se lhe desse importância. O Sr. Hadlum não teve sorte. Com efeito, os juízes lhe deram sentença contrária, pois reconheceram que, embora 349 dias para uma gestação não fosse algo habitual, tampouco estava fora da amplitude da possibilidade científica. Não obstante, em casos anteriores, se havia

julgado que um período de 340 dias de gestação era algo impossível, à luz dos últimos resultados da Medicina. A Câmara dos Lordes pôs um ponto final formal a esta controvérsia, estabelecendo que o “limite” do período em questão devia ser 360 dias. De qualquer forma, ninguém realizou um estudo estatístico para estabelecer esses limites ou fundamentar suas alegações. Em 1975, R. Chamberlain realizou um estudo sobre nascimentos na Grã-Bretanha e sobre o qual Barnett fez uma análise estatística. A Figura 3 exibe um histograma dos períodos de gestação estimados, em uma amostra de tamanho 13 634 nascimentos ocorridos na Inglaterra junto com os pontos correspondentes aos 349 dias da Sra. Hadlum e os 360 do “limite” estabelecido pela Câmara dos Lordes.

Obviamente, o ponto 349 é “surpreendente”, diremos então que é um *outlier*. Agora, a questão de interesse é decidir se este ponto é um “extremo” e, como tal, um proveniente da distribuição que se supõe

FIGURA 3



corresponder à variável aleatória $X =$ "período de gestação da mulher"; ou se se trata de um ponto "contaminante", isto é, um ponto proveniente de outra distribuição. De fato, o Sr. Hadlum não se interessa em saber qual é a média e/ou a variância da distribuição da variável X .

2.2 — Detecção de certos fenômenos pouco comuns em um contexto dado

Pensemos em situações onde a presença de *outliers* interessa, para saber se se produz ou não um fenômeno estranho. Por

exemplo, Beckman e Cook (1983) citam o seguinte caso. Há poucos anos atrás um satélite russo se destruiu quando passava sobre a região central do Canadá. Em um esforço conjunto, os governos desse país e dos Estados Unidos realizaram um trabalho de localização dos restos desse satélite por meio de um procedimento que envolvia a medição de níveis de radioatividade, enquanto sobrevoavam os possíveis lugares de queda de tais restos. Níveis altos de radiação com relação aos níveis da maioria dos lugares medidos eram considerados como indicadores de lugares prováveis de presença de restos do satélite. A metodologia estatística básica, neste caso, consistiu em uma técnica de detecção de *outliers* em uma amostra de grande tamanho. Não havia interesse em estimar a média do nível de radiação na zona central do Canadá, exceto para decidir se certos níveis eram ou não *outliers*.

2.3 — Indicadores para diagnósticos

Em certas análises, onde não se tem certeza sobre a bondade do modelo (distribuição-alvo) que supostamente explica o fenômeno em estudo, a presença de *outliers* constitui freqüentemente uma ferramenta valiosa para detectar as falhas do modelo e/ou do conjunto de dados observados. A partir daí, podemos obter uma reformulação mais realista do modelo e decidir uma metodologia mais adequada para a coleta de dados. Por exemplo, os *outliers* podem revelar que não é possível supor aditividade e homocedasticidade no modelo. Na verdade, *toda* análise de dados deveria incluir técnicas de diagnósticos de *outliers* que pudessem fornecer informação sobre o uso apropriado da mesma e a certeza de suas conclusões.

2.4 — Acomodamento de (a convivência com) *outliers*

Aqui centramos o interesse em estimar, ou fazer inferências sobre parâmetros da população-alvo. Suspeitamos, à luz dos fatos e/ou das condições do problema, que existem pontos “contaminantes”, cuja presença distorce gravemente a otimalidade das técnicas clássicas, que são recomendáveis sob o modelo “puro”. Tratamos

então de usar outras técnicas (robustas, resistentes, etc.) que dêem resultados válidos para inferências sobre os parâmetros do modelo “puro” não só sob esse modelo e sim, também, sob outros modelos “próximos” a ele. Em tais casos, os parâmetros da distribuição (ou distribuições) contaminantes não são de interesse em si, sendo considerados como parâmetros perturbadores. Uma característica de quase todas estas técnicas, quando são usadas em problemas de estimação e/ou de testes de hipóteses, é que não desprezam os *outliers* e sim que os “acomodam”, de maneira tal que os valores estimados não resultem muito distorcidos pela influência deles. Deste modo, ao usar tais técnicas, o analista não está interessado, principalmente, em identificar *outliers*, e sim em usar toda a informação disponível a partir da amostra a fim de entender a ação da mesma globalmente. Elas não servem para estudar os *outliers* em particular. Porém, tem-se começado a usar os pesos que dão a cada ponto, como informação para identificá-lo [ver, por exemplo: Prescott (1980), Cook e Beckman (1980), Iglewicz e Martinez (1982), Leroy e Rousseeuw (1984), entre outros].

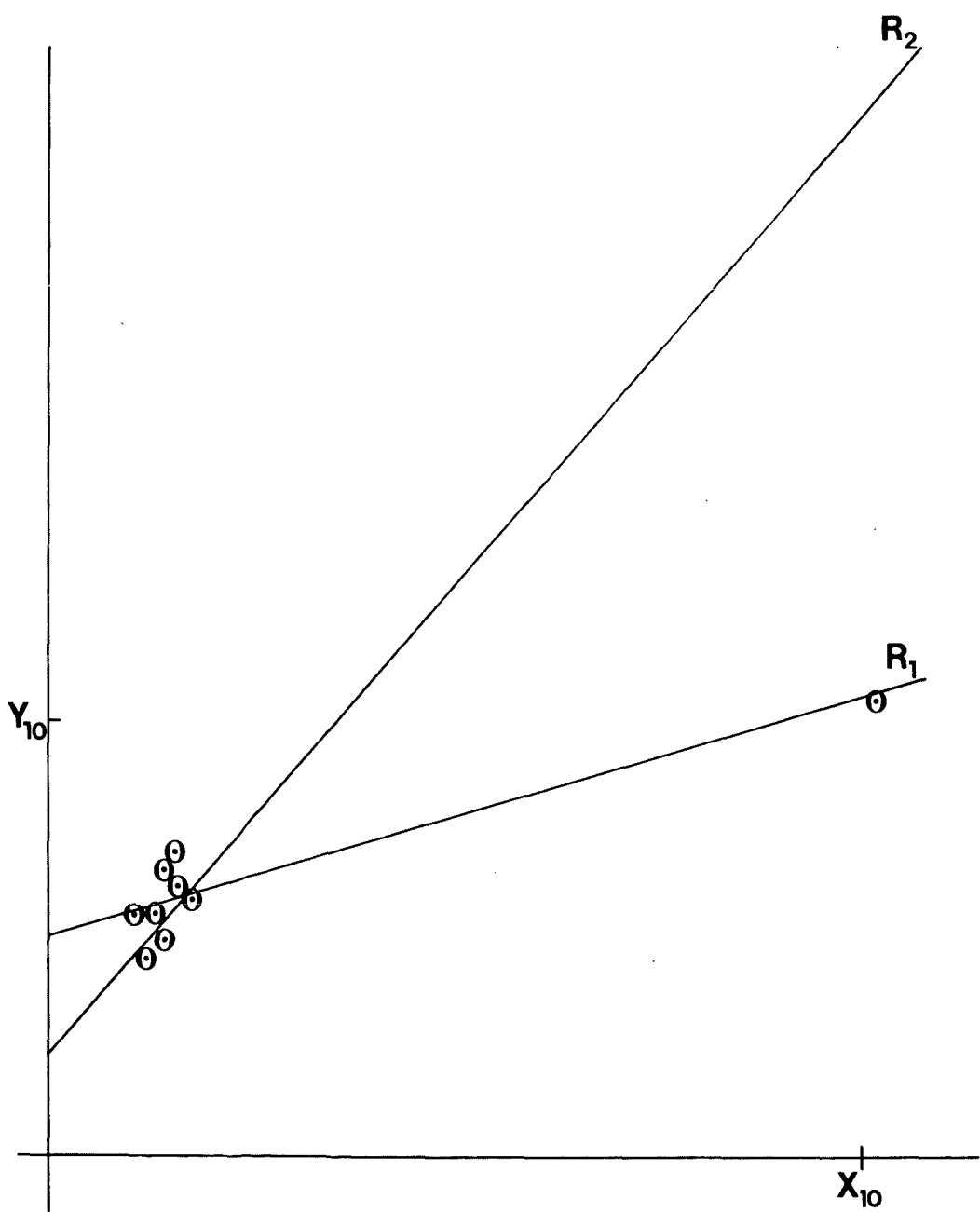
2.5 — Manejo de pontos influentes

Estes pontos são, em um sentido amplo, aqueles que determinam fortemente os resultados de uma técnica estatística. Para entender o que queremos dizer, vejamos a Figura 4. A reta R1 corresponde à de mínimos quadrados, obtida ao considerar a observação (x_{10}, y_{10}) , enquanto que a R2 é a obtida ao excluir tal ponto. Neste caso, dizemos que o ponto (x_{10}, y_{10}) é um ponto influente (com respeito à técnica de mínimos quadrados).

Apesar de o gráfico anterior ter sido inventado a propósito, existem numerosos exemplos reais onde a ação de um ou de poucos pontos é decisiva [ver, por exemplo, Kerr (1982)]. Quando estes pontos são retirados da amostra, os valores estimados variam significativamente.

Agora, estes pontos não são necessariamente *outliers*, nem tampouco os *outliers* têm por que ser pontos influentes [ver, para mais detalhes: Chatterjee e Hadi (1986)]. Porém, é útil encarar os pontos influentes

FIGURA 4



como tipos particulares de *outliers*. Frequentemente, seu estudo tem servido para manifestar certos fenômenos não suspeitados que ocorriam ao realizar uma determinada experiência. Por exemplo, Rayleigh [citado por Hampel e outros (1986), pág. 28] notou que 7 de 15 medidas do peso atômico do nitrogênio estavam em um grupo separado. Isso o levou a concluir a existência do argônio.

3 — COMO PODERÍAMOS EXPLICAR O APARECIMENTO DE OUTLIERS

Para extrair a maior quantidade possível de informação valiosa a partir de uma massa de dados, haja ou não *outliers*, devemos dar muita importância ao estudo do mecanismo no modelo que, supomos, explica a geração de *todos* esses dados, os "bons" e os "maus".

Como já dissemos, a manifestação de *outliers* é um assunto de natureza subjetiva, que depende fundamentalmente do modelo subjacente à amostra que temos em mente. Assim, nosso parecer acerca da presença de *outliers* em um conjunto de dados pode variar bastante com os diferentes modelos probabilísticos (distribuição-alvo) que razoavelmente assumíssemos antes de realizar a experiência e obter os dados. Neste sentido, o principal propósito de um conjunto de métodos estatísticos para examinar *outliers* é, em termos amplos, fornecer meios de assegurar se nossa declaração subjetiva, acerca da presença de tais observações em um certo conjunto particular de dados, tem implicações objetivas, práticas e importantes para as análises que realizarmos posteriormente à experiência.

Vejamos agora algumas dessas causas mais plausíveis que explicam a aparição de *outliers*.

Em primeiro lugar, observações "surpreendentes" podem aparecer simplesmente por causas determinísticas. Por exemplo: leituras, registros ou cálculos errados. Na maioria de tais casos, é óbvio o que devemos fazer: retirar os dados errados ou substituí-los por valores corretos. Porém, cuidado!, o método de substituição de um

dado, possivelmente "errado", por um possivelmente "correto", deve obedecer a critérios definidos com toda precisão e sem ambigüidade. Quando as circunstâncias não são tão nítidas, onde suspeitamos, mas não asseguramos, que um determinado *outlier* seja um dado errado, devemos pensar que tal ponto foi causado pela ação de um mecanismo aleatório. Aí, devemos usar um modelo probabilístico para tentar sua explicação.

Que classes de modelos probabilísticos poderíamos sugerir? Para responder a esta pergunta devemos pensar nas causas possíveis da variabilidade dos dados. Estas "fontes de variabilidade" podemos classificar, assim, segundo Barnett e Lewis (1984):

3.1 — Variabilidade inerente (aos dados)

Essa variabilidade é expressada através da "distribuição verdadeira". É uma característica "natural" da população que estudamos (população no sentido estatístico). Reflete as propriedades de distribuição do modelo correto que descreve o mecanismo de geração das observações. Assim sendo, não podemos exercer controle sobre ela, as observações aparecem por direito próprio.

3.2 — Variabilidade ocasionada por erros de medição

As observações poderiam ser afetadas pelas imperfeições dos instrumentos ou processos de medição. Devemos considerar aqui os inevitáveis erros de arredondamento e certos erros de registro de dados numéricos dificilmente detectáveis. Este último tipo de erro é cada vez mais freqüente quando lidamos com grandes massas de dados: erros provenientes de perfurações de cartões, de entradas de dados "via-teclado", de formatos de leitura inadequados [perigo de usar o formato F do FORTRAN, segundo o destacado por Atkinson (1985)]. O experimenter poderia exercer algum tipo de controle sobre esta variabilidade.

3.3 — Erros provenientes da forma como realizamos uma experiência (erros de "execução")

Certo tipo de variabilidade poderia provir da coleta imperfeita dos dados. Talvez

tenhamos escolhido inadvertidamente uma amostra viciada ou incorporado indivíduos de uma população diferente da que queremos amostrar. É claro que ser cuidadoso na etapa de amostragem pode reduzir a variabilidade, mas poderia ser difícil ou impossível saber se existem ou não tais erros de execução. Em tais casos, possivelmente seja mais adequado trabalhar com um modelo subjacente capaz de tratar convenientemente com essa possibilidade de presença de dados anômalos.

Pois bem, o analista supõe que essa variabilidade é a expressão de um certo modelo probabilístico, modelo subjacente. É com respeito a este que se julgará se uma certa observação é ou não "surpreendente", isto é, *outlier*. É aqui que tais pontos podem ser causados por uma falta de concordância entre o modelo assumido e o verdadeiro. Mas devemos ter em conta também que esses pontos surpreendentemente afastados da massa de dados podem apresentar-se mesmo no caso de concordância entre ambos os modelos, simplesmente porque a distribuição que aparece no modelo é do tipo de "caudas pesadas". Quando existe uma discordância podemos pensar que o modelo suposto é inadequado para explicar os dados em sua totalidade (incapacidade global) ou somente para explicar a aparição de certos pontos (incapacidade local). Dentro da primeira, podemos colocar diversas razões: escalas inadequadas para as variáveis, matrizes de planejamento não balanceadas ou mal condicionadas nos modelos de regressão. Já como motivos de incapacidade local podemos pensar que a distribuição assumida não contempla a possível aparição de erros ou pontos influentes. Hawkins (1980) propõe certos modelos probabilísticos que poderiam explicar, nos casos mais freqüentes de amostras de uma população univariada, a aparição de pontos afastados da maioria. Uma formalização do mecanismo gerador de *outliers*, ligeiramente diferente da proposta de Hawkins (1980), podemos ver em Beckman e Cook (1983). Esta proposta está baseada em uma anterior de Neyman e Scott (1971). Finalmente, uma classificação exhaustiva de modelos possíveis para explicar não tanto *outliers* e, sim, mais especificamente, "contamina-

ções", encontraremos no recomendável livro de Barnett e Lewis (1984).

Aqui veremos somente a de Hawkins, apenas para ter uma idéia das possibilidades de formalização.

Diz Hawkins (1980) que, com referência a gêneses de dados univariados, podemos distinguir, em geral, dois possíveis mecanismos de geração de dados afastados:

3.3.1 — Os dados provenientes de uma distribuição com "caudas" pesadas

As observações afastadas não têm porque ser consideradas errôneas. Os *outliers* aparecem por direito próprio. A este respeito podemos destacar duas famílias de distribuição: as "absolutamente propensas a *outliers*" (*absolutely-outlier-prone*) e as "relativamente propensas a *outliers*" (*relatively-outlier-prone*).

Definição: Seja F uma função de distribuição acumulada sobre R , (x_n) uma sucessão de v.a.i.i.d. com distribuição F , e para cada n , $x_{n:1}, \dots, x_{n:n}$ as estatísticas de ordem associadas com $x_1, \dots, x_n$. Dizemos que:

a) F é absolutamente propensa a *outliers* se existem $\epsilon > 0$, $d > 0$ e n_0 tais que

$$P(x_{n:n} - x_{n:n-1} > \epsilon) > d \text{ para todo } n > n_0.$$

b) F é relativamente propensa a *outliers* se existem $c > 1$, $d > 0$ e n_0 tais que

$$P(x_{n:n}/x_{n:n-1} > c) > d \text{ para todo } n > n_0.$$

Se F não é absolutamente (relativamente) propensa a *outliers*, dizemos que é absolutamente (relativamente) resistente a *outliers*.

3.3.2 — Os dados provêm de duas distribuições

Uma gera os dados "bons" (isto é, a maioria), e a outra os dados "maus" ou *outliers*. É costume chamar a primeira distribuição "pura" e a outra "contaminante". Por sua vez, podemos subclassificar este tipo de modelo em duas categorias, onde:

a) O número de *outliers* é conhecido.

b) Uma observação vem da distribuição "contaminante" com probabilidade p e da "pura" com probabilidade $(1 - p)$. Neste ca-

so, se f_0 é a densidade da "pura" e f_1 , a da "contaminante", a densidade da distribuição que gera as observações será

$$f = (1 - p) \cdot f_0 + p \cdot f_1.$$

Habitualmente, tratamos de estimar parâmetros de f_0 ou às vezes, também, de f_1 e raramente de toda a f .

4 — O QUE FAZER COM OS OUTLIERS?

A resposta a esta questão está obviamente relacionada com nossa crença acerca do mecanismo gerador da amostra e também com os objetivos da análise.

Em primeiro lugar, suponhamos que nosso interesse esteja dirigido a inferir características do modelo básico ou subjacente (aos dados "bons").

Se assumimos um mecanismo do tipo 3.3.1 da Seção 3 e conhecemos a forma da distribuição supostamente verdadeira, então os *outliers*, apesar de sua aparência (afastados do comum dos dados), são observações válidas dessa distribuição. Em tal caso, deveríamos usar técnicas que soubermos serem ótimas sob esse modelo; por exemplo, estimadores de máxima verossimilhança.

Poderíamos pensar que os fenômenos reais sujeitos à distribuição de "caudas" pesadas como as Cauchy ou outras são raros. Barnett e Lewis (1984) dizem que tal idéia é errônea, ao menos em Biologia. Também, nesta Ciência, não é rara a aparição de distribuições marcadamente assimétricas. Sustentando esta afirmação, esses autores dão um exemplo de um conjunto de dados referentes ao número de indivíduos de uma certa espécie de mariposa noturna capturados em Rothamsted:

11, 54, 5, 7, 4, 15, 560, 18, 120, 24, 3, 51, 3, 12, 84.

Aqui, o valor 560 é, de fato, um dado genuíno inerentemente característico ao fenômeno natural que se estuda. Para ressaltar sua característica de "surpreendente", vejamos a Figura 5.

Se houvéssemos imaginado como distribuição verdadeira uma normal, a dis-

crepância entre o imaginado e o observado seria evidente apenas ao olharmos o gráfico (Normal-plot) da Figura 6.

Com efeito, se imaginássemos que os logaritmos dos dados seguem uma normal, quer dizer, se os dados tivessem uma distribuição log-normal, aí a discrepância seria marcadamente menor. Vejamos o gráfico (Log-normal-plot) na Figura 7.

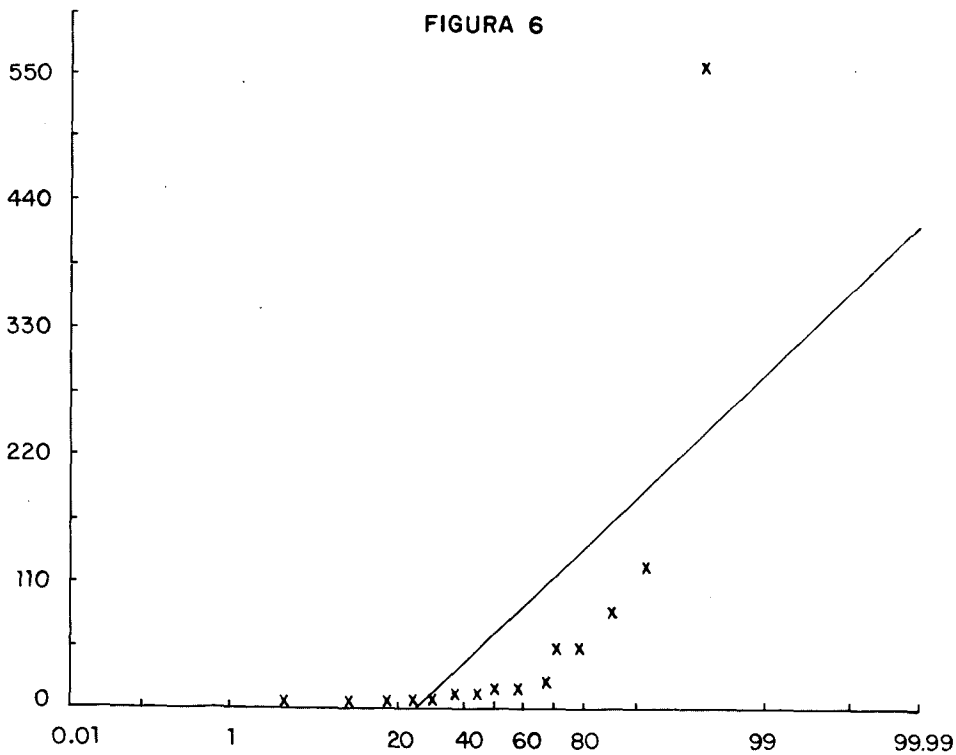
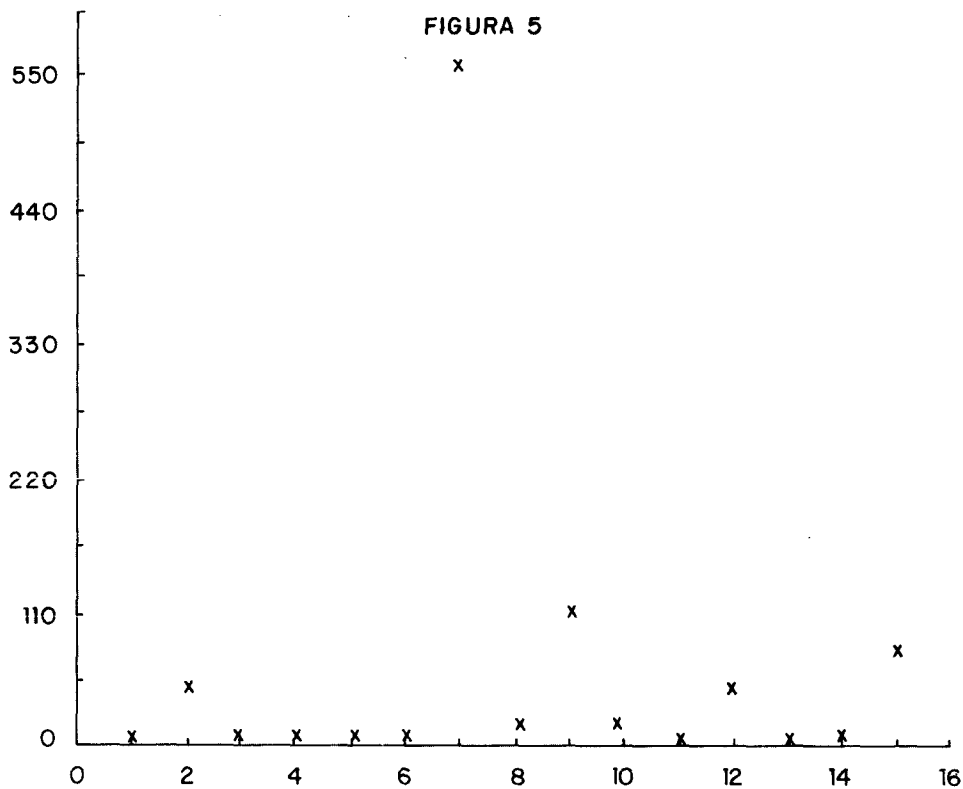
Continuando com o tema desta seção, suponhamos que o mecanismo de geração das observações será do tipo anterior, mas que não tenhamos certeza acerca da forma da distribuição verdadeira. Assim, teremos uma dúvida quanto à distribuição que adotamos como subjacente. Por outro lado, imaginemos que não queiramos ou não tenhamos elementos suficientes para realizar uma análise que conduza a uma redefinição do modelo. Neste caso, os *outliers* seriam considerados como valores perturbadores e buscaríamos métodos robustos de análises a fim de minimizar seu impacto.

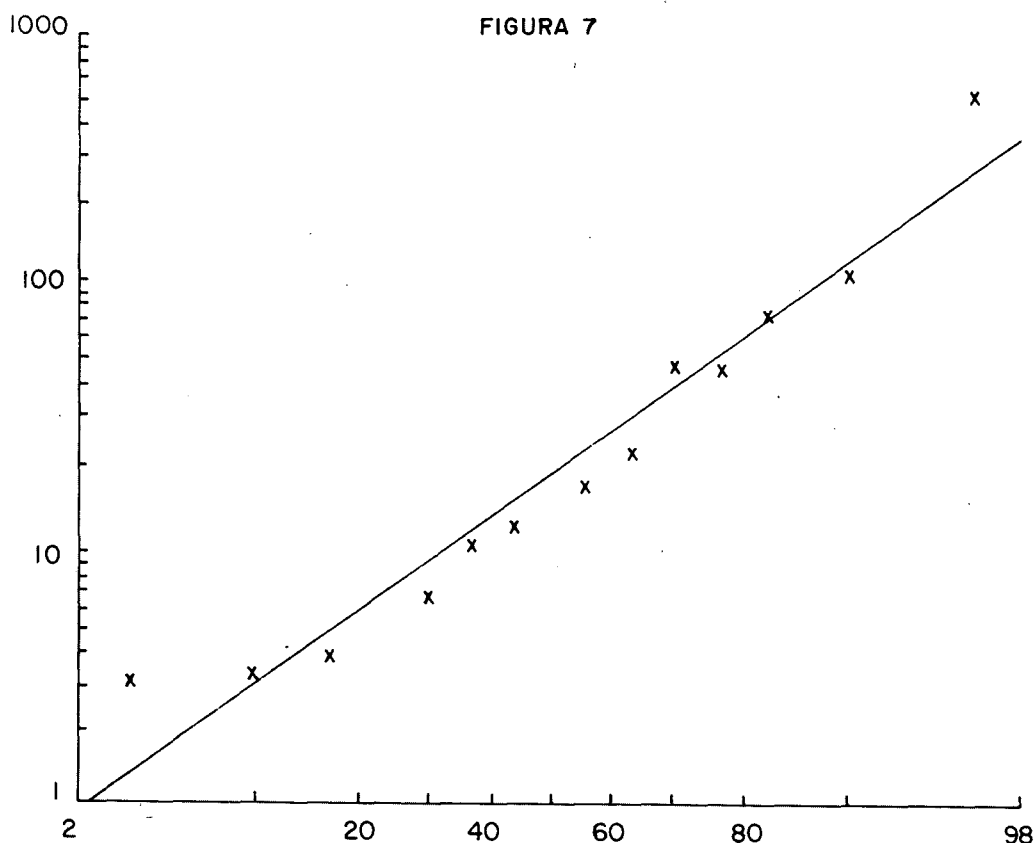
Um exemplo típico desta situação está mostrado no já mencionado Barnett e Lewis (1984). Trata-se de um conjunto de dados consistente de 15 observações efetuadas por Herndon em 1846 do semidiâmetro vertical do planeta Vênus. Pierce em 1852 e Chauvenet em 1863 analisaram os resíduos, ao redor da média amostral, para ver se havia ou não *outliers* que deviam ser desprezados, supondo normalidade. O conjunto em questão é:

-0.30	0.48	0.63	-0.22	0.18
-0.44	-0.24	-0.13	-0.05	0.39
1.01	0.06	-1.40	0.20	0.10

A observação de menor valor, que poderia ser considerada como *outlier* tendo em mente um modelo normal, seria ou não "contaminante"? Pierce e Chauvenet consideraram que sim e a descartaram em suas análises posteriores. Não obstante, Barnett e Lewis (1984) aplicaram alguns testes de discordância a esses dados e não concluíram a mesma coisa que Pierce e Chauvenet. Moral: em dúvida, é melhor não retirar o *outlier* e sim melhor "acomodá-lo". Voltaremos a isto mais adiante.

Por ora, e apenas por simples curiosidade, vejamos alguns gráficos [Figuras 8.a, 8.b e 8.c], trabalhando com esse conjunto de dados.





Nas Figuras 8.b e 8.c, não está nada claro que devemos preferir o modelo normal ao logístico. Essa dúvida subsistiria ainda que realizássemos certos testes de "bondade de ajuste". Mas não entremos em detalhes, neste momento.

Agora, se retirarmos a observação -1.4 , obteremos os gráficos mostrados nas Figuras 9.a, 9.b e 9.c, que também nos deixam na mesma dúvida anterior. De passagem, observemos que as médias e medianas dos conjuntos com e sem a observação -1.4 dão:

Média com observação $-1.4 = 0.018$

Média sem observação $-1.4 = 0.119$

Mediana com observação $-1.4 = 0.06$

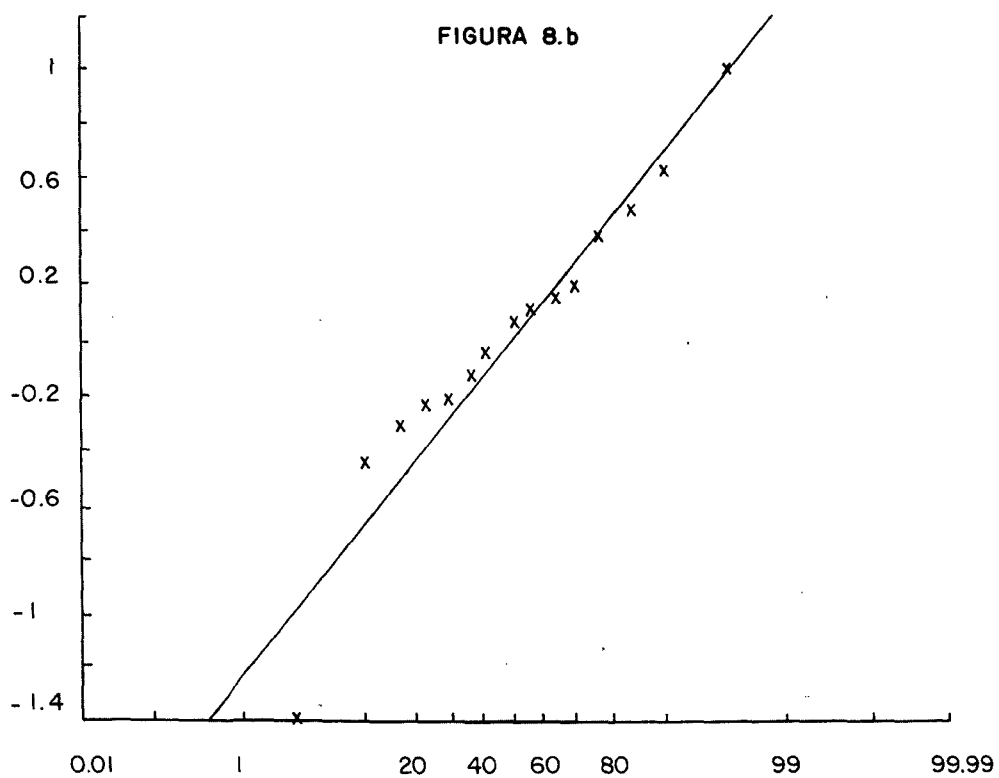
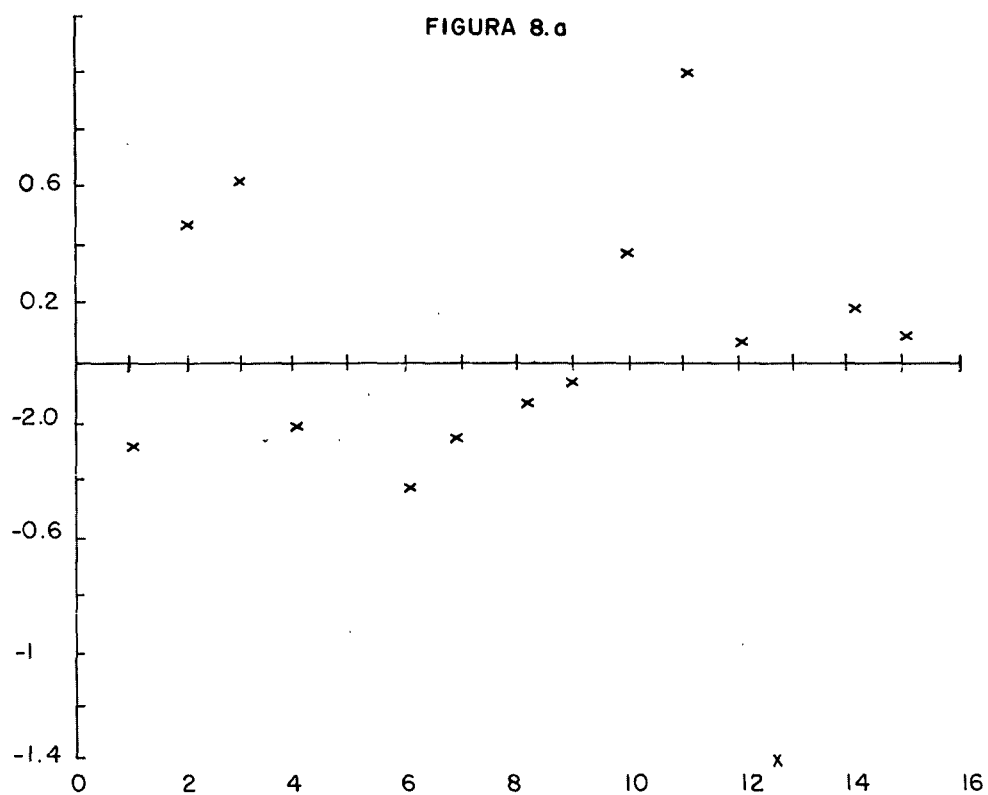
Mediana sem observação $-1.4 = 0.08$

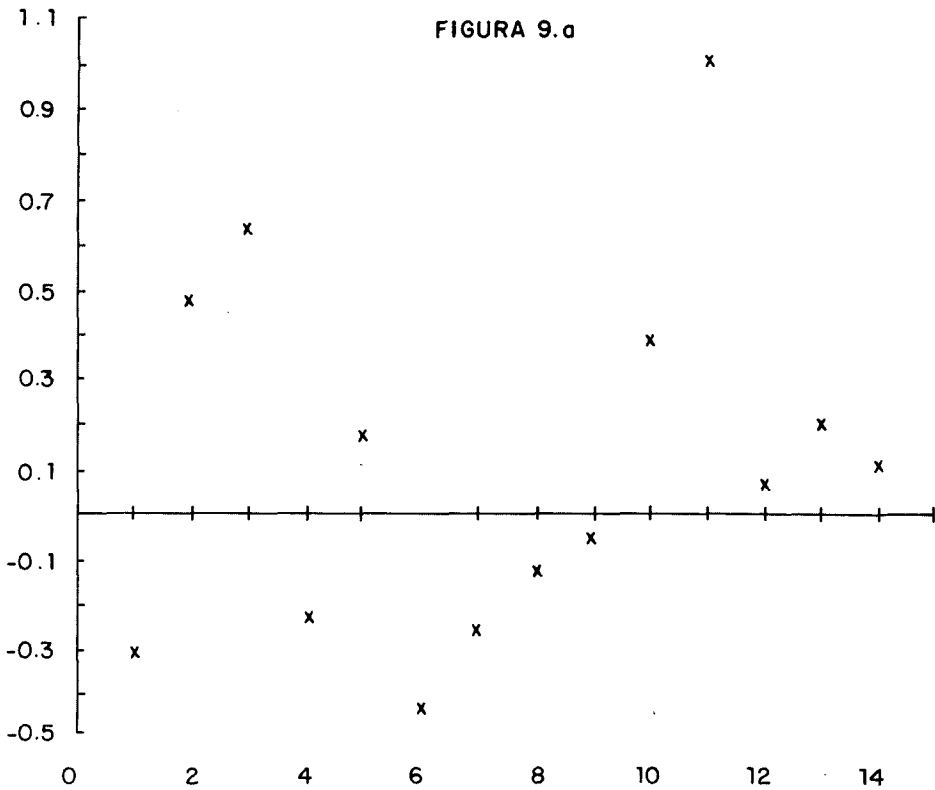
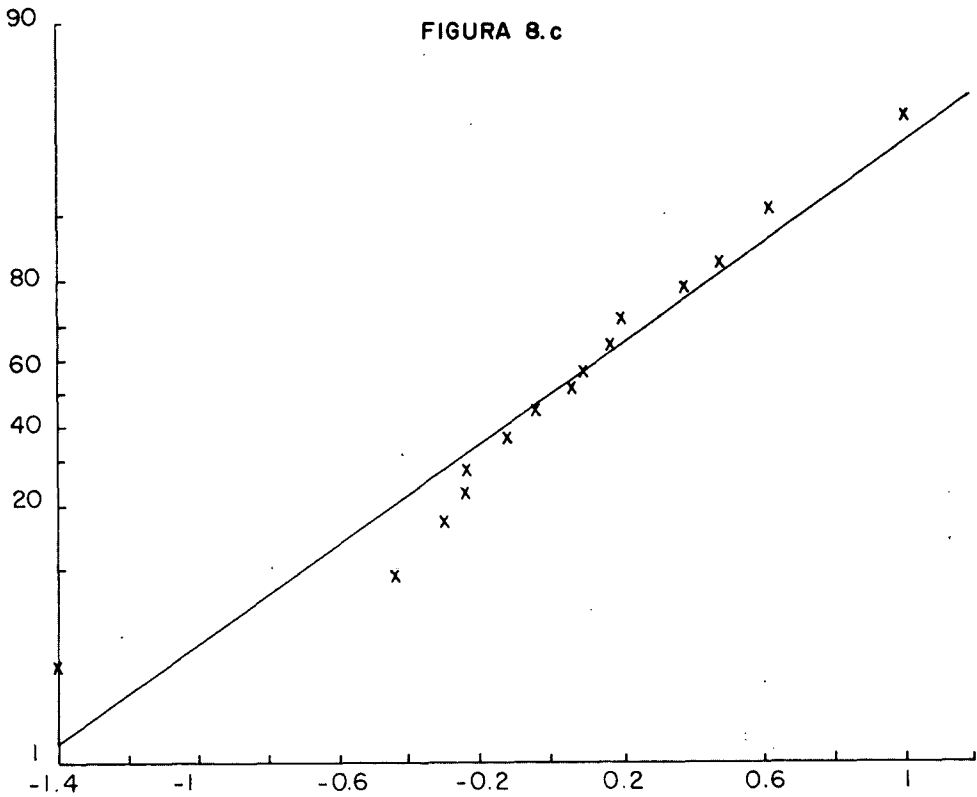
que revelam a maior estabilidade da mediana frente à média.

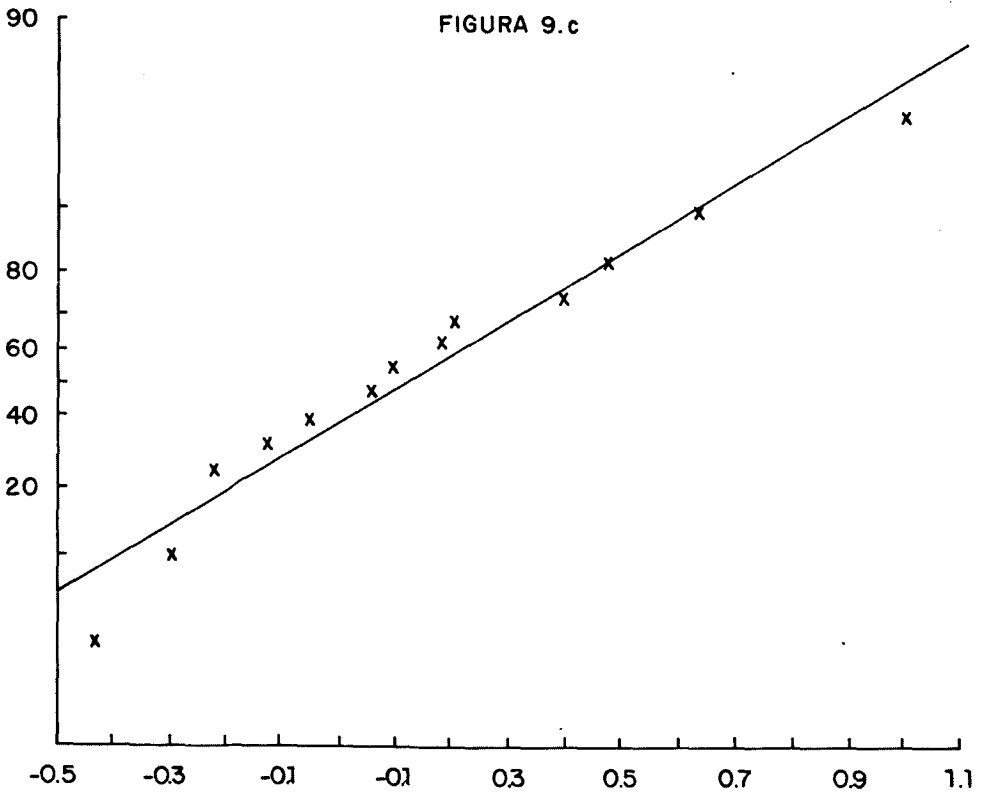
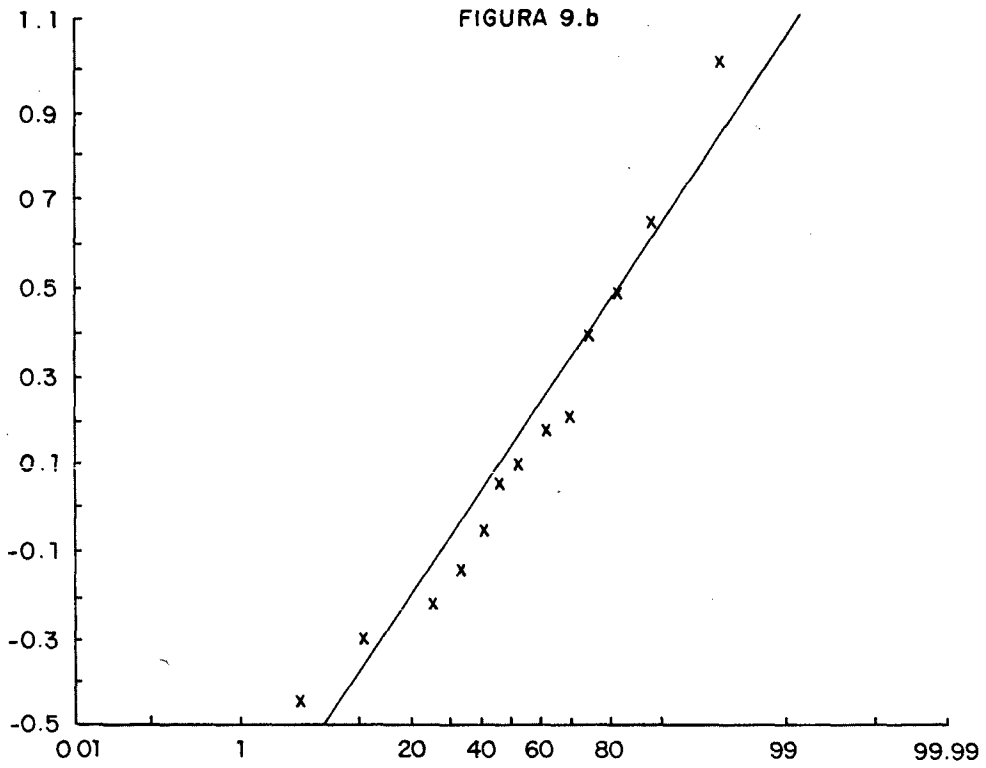
Devemos ter em conta que o procedimento "rejeição-de-outliers" pode ser visto co-

mo um método de estimação robusta. Com efeito, em uma seção posterior, mostraremos como os procedimentos do tipo "rejeição-de-outliers + estimação-clássica" podem considerar-se robustos, ao menos em certos aspectos. Esta idéia foi implementada formalmente por Anscombe (1960) sob um mecanismo do tipo b do 3.3.2 na Seção 3.

Têm-se levantado muitas objeções a este tipo de procedimento. Uma delas se refere à influência marcante que uma ou pouquíssimas observações têm sobre o estimador final. Glaisher (1872) propôs um método alternativo que essencialmente é um da classe de média amostral de observações ponderadas. Outra objeção séria a estes procedimentos é que, freqüentemente, não detectam outliers "mascarados". Estes, ao serem usados como parte da amostra, ocasionam distorções sérias nos resultados alcançados pelas técnicas clássicas. Sobre este assunto, voltaremos mais adiante ao comentar o trabalho de Hampel (1985).







Outro método de estimação que veremos com detalhe mais adiante é o de Winsorização. Este método não despreza *outliers* e sim os “move” a uma posição próxima à massa de dados.

Em última instância, o método de tratamento de *outliers* que usaremos depende do mecanismo suposto para a geração dos dados.

Suponhamos agora que nosso propósito é verificar se o modelo básico que supusemos está de acordo com o verdadeiro e que, em caso contrário, queremos modificá-lo.

Para Beckman e Cook (1983) as técnicas de diagnóstico sobre a debilidade do modelo suposto, baseadas sobre identificação de *outliers*, são particularmente úteis durante a construção iterativa de um modelo. Iterativa no sentido esquematizado no seguinte diagrama:

Modelo inicial → Verificação → Modelo novo → Verificação → ... → Modelo final.

Na próxima seção passaremos em revista algumas técnicas de identificação de *outliers*.

Reciprocamente, um teste de bondade de ajuste pode ser usado, em princípio, para testar a presença de *outliers*. Como exemplo, destacamos que dois testes de bondade de ajuste muito efetivos para detectar *outliers* em uma amostra proveniente de uma população normal, segundo Hawkins (1980), são os baseados nos coeficientes de momentos para simetria e *kurtosis*.

Finalmente, quando os dados são gerados por um mecanismo do tipo 3.3.2 definido na Seção 3, poderíamos estar interessados primeiramente em testar se há observações contaminantes e, em caso afirmativo, seguir averiguando quantas e quais são essas observações. Uma vez classificadas as observações em “puras” e “contaminadas”, continuaríamos com uma análise já sobre as “puras” ou as “contaminadas”. Uma análise sobre as primeiras, o fariamos se estivéssemos interessados principalmente em realizar inferências sobre o modelo suposto verdadeiro. Nossa atenção se focalizaria nas “contaminantes”, se nosso interesse fosse captar a informação contida nelas, como descrito na Seção 2. Recordemos uma vez mais um fato bem conhecido e documentado: muitos descobrimentos impor-

tantes em Astronomia foram realizados como resultado de um estudo cuidadoso dos conjuntos dos “resíduos” obtidos a partir do ajuste de órbitas.

Para terminar esta seção sem ser repetitivo, aconselhamos ao leitor, neste ponto, a recordar o que foi dito sobre o “manejo de pontos influentes”, no final da Seção 2.

5 — PROCEDIMENTOS ESTATÍSTICOS RELEVANTES PARA MANEJAR OUTLIERS

Quando estudamos a história do manejo de *outliers* vemos que os procedimentos propostos até o final do Século XIX podem ser classificados em duas categorias. Uma delas, formada pelos métodos de identificação de *outliers*. A outra categoria, que se desenvolveu paralelamente à primeira, é formada por testes desenhados para decidir se o (ou os) *outlier* identificado deveria ser desprezado ou retido.

Posteriormente apareceram procedimentos para estimar e/ou testar parâmetros da distribuição básica, relativamente livres da influência que os *outliers* poderiam ocasionar. Estas técnicas recebem o nome de procedimentos de “acomodação de *outliers*”. Também são conhecidas como “técnicas robustas” ou “resistentes” contra *outliers*.

Como já sabe o leitor, a “Estatística Robusta” é todo um vigoroso ramo da Estatística que não se limita ao tema *outliers*. Sem a pretensão de rigor, podemos dizer que o propósito principal da Robustez é encontrar técnicas, cuja utilidade não se restrinja ao conjunto de hipóteses que sugerem o modelo suposto, e sim que continuem fornecendo informações de valor sobre o fenômeno estudado, quando aquelas hipóteses forem relaxadas. Uma resenha muito completa do efetuado nesta área até o momento pode ser encontrada no excelente livro de Hampel e outros (1986). Esta resenha apenas se refere ao campo fora de Séries Temporais. Quem desejasse uma resenha sobre Robustez em Séries Temporais poderiam ler Martin e Yohai (1985).

O interesse da Robustez dentro do tema *outliers* reside no desejo natural de encontrar técnicas resistentes contra *outliers*

quando suspeitamos de sua presença em uma amostra. A título de exemplo, é fácil ver a forte influência que as observações extremas têm sobre a variância amostral:

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

Logo, se sabemos que tais observações são contaminações inevitáveis, mas indesejáveis, estaremos inclinados a usar outros estimadores de escala menos sensíveis a esses extremos. Por exemplo, Hampel (1974) sugere o uso da chamada *median deviation* ou "mediana do desvio absoluto da mediana":

$$s_m = \text{mediana} \{ |x_i - \text{mediana} \{x_j\}| \}.$$

Como já dissemos, é importante ter em conta o modelo que *verdadeiramente* gera os dados. Assim, se estivermos interessados em estimar a escala das observações e a distribuição verdadeira é uma Cauchy enquanto que a suposta é uma de caudas menos pesadas, então não é aconselhável usar um procedimento robusto para protegernos das observações extremas, pois em tal caso subestimamos a verdadeira escala das observações e não conseguimos indicadores para corrigir a distribuição suposta. Mas... quem conhece a "verdade"?

Por outra parte, se a distribuição suposta é uma normal e como alternativa estamos pensando em uma distribuição com caudas mais pesadas, então é aconselhável usar um estimador que acomode os *outliers* e não que simplesmente os despreze. Isto é essencial para compreender o espírito subjacente a todas as técnicas robustas. trata-se de uma atitude pessimista em torno da nossa capacidade de conhecer a verdadeira distribuição que gera os dados. As técnicas de "rejeição-de-*outliers*" se fundamentam em uma atitude bem mais otimista: sabemos que o modelo suposto é correto ("bom") e que os dados podem não ser de muito boa qualidade para refletir a distribuição verdadeira; justifica-se assim que queremos desprezar às observações, espúrias.

Do ponto de vista metodológico podemos distinguir duas classes de procedimentos robustos. Uma classe está formada por aqueles que protegem implicitamente contra *outliers*, pois, em geral, dão um peso menor às observações, afastados da maio-

ria. Poderíamos denominar esta classe de "procedimentos robustos-protetores". A outra classe está formada por técnicas de estimação e testes de hipóteses que foram construídas pensando em fazer frente explicitamente aos *outliers* considerados como observações contaminantes.

O outro tipo de método estatístico que aparece usado amplamente é o de "testes de discordância". Para fixar idéias, suponhamos que queremos construir um teste que sirva para averiguar se o extremo superior de uma amostra univariada é um *outlier*. Seja F a distribuição suposta (objetivo) e suponhamos que $x_1, \dots, x_n$ são os valores observados. Estes mesmos valores ordenados de menor a maior seriam:

$$x_{(1)} < \dots < x_{(n)}.$$

A hipótese nula seria:

$$H_0: x_{(n)} \text{ não é um outlier.}$$

Em termos mais precisos:

$H_0: x_{(n)}$ é um valor observado de $X_{(n)} = \max \{X_1, \dots, X_n\}$ onde $X_1, \dots, X_n$ são v.a.i.i.d. com distribuição comum F .

Seja $\alpha < 1/2$. Um teste de discordância de nível $1-\alpha$ será um tal que se R é a região de rejeição (ou região crítica), então:

$$P_F(X_{(n)} \in R) < \alpha.$$

Logo, se $x_{(n)} \in R$, então desprezamos H_0 .

Agora, o que obteremos se rejeitarmos H_0 ? Para podermos estudar propriedades de qualquer teste e/ou compará-lo com outros, necessitamos propor uma hipótese alternativa a H_0 .

Surge assim a necessidade de aprender alguma coisa acerca do que podem ser modelos razoáveis de *outliers* que darão distribuições alternativas à F ou, para ser coerente com a nomenclatura usada na Seção 1, modelos para a contaminação da F . Já dissemos que um estudo bastante completo sobre este assunto pode ser visto em Barnett e Lewis (1984).

Como síntese acerca da relação entre as fontes de variabilidade, a natureza dos *outliers* e os objetivos em mente ao realizar uma análise é conveniente ter presente o resumo, apresentado por Barnett e Lewis (1984).

Para terminar, uma advertência. Ainda que tenhamos apresentado em forma separada as técnicas de acomodamento e as de discordância e, apesar de ambas se basearem sobre filosofias aparentemente diferentes, de modo algum devemos considerá-las como técnicas contrapostas. Pelo contrário, um método de acomodamento pode produzir um método para identificação de *outliers* e vice-versa. Sobre este assunto recomendamos a leitura do trabalho de Prescott (1980).

6 — BREVE HISTÓRIA SOBRE O MANEJO DE *OUTLIERS* ANTES DO SÉCULO XX

Segundo Beckman e Cook (1983) e Barnett e Lewis (1984) destacaremos, brevemente, o que foi realizado por aqueles cientistas dos Séculos XVIII e XIX que, ao tratar de extrair inferências válidas de seus experimentos nas ciências aplicadas a que se dedicavam, tornaram-se os verdadeiros pioneiros da Análise de Dados.

Barnett e Lewis (1984) citam um trabalho de Boscovitch (1775), onde ele decidiu desprezar as duas observações extremas catalogando-as de *outliers*, sem outro motivo que o de julgá-las afastadas das outras.

Bernoulli (1777), referindo-se aos trabalhos de certos astrônomos, questionou a suposição de que os erros de observação deveriam ser identicamente distribuídos. Também condenou a prática de descartar observações discordantes, a menos que se tivesse o conhecimento *a priori* de que essas observações estavam erradas por causa de um acidente.

Bassel e Baeuer, em 1838, declaravam que jamais desprezavam uma observação por ser o resíduo dela, com respeito à média amostral, grande.

Por outra parte, Legendre, em 1805, recomendava a rejeição dos pontos cujo desvio fosse demasiado grande para ser admissível.

Como podemos apreciar, todo esse período se caracterizou pela completa ausência de procedimentos formais.

Quem primeiro publicou um teste objetivo, bastante confuso em sua formulação matemática, foi o astrônomo Peirce em 1852. Em 1855, Gould publicou na *Astronomical Journal* uma tabela, com a qual pretendia ajudar os investigadores na aplicação do teste de Peirce.

Chauvenet, em 1863, propôs um teste objetivo para a rejeição de uma só observação. Este teste, apesar do defeito notável que veremos a seguir, até hoje é usado em muitos textos para engenheiros e cientistas experimentais. O raciocínio de Chauvenet era, aproximadamente, o seguinte: para cada número real x seja $p(x)$ a probabilidade de que uma observação seja $y = x$, então é natural pensar que:

valor esperado de (número de observações maiores que x em um conjunto de n observações) $= n p(x)$.

Buscando x_n tal que $n p(x_n) < 1$, por exemplo: $n p(x_n) = 1/2$, teríamos que rejeitar como errônea uma observação cujo resíduo em módulo fosse maior que $p(x_n)$.

Para ver o efeito desta proposta, aparentemente tão natural, usaremos um pouco de formalismo moderno.

Sejam $X_1, \dots, X_n$ v.a.i.i.d. com distribuição simétrica em torno de zero, então:

$$P(\exists 1 \leq i \leq n: |X_i| \geq x_n) = 1 - P(|X_i| < x_n, \forall i) = 1 - \left(1 - \frac{1}{2n}\right)^n \rightarrow 1 - \exp\left(-\frac{1}{2}\right)$$

Isto nos permite concluir que, se adotarmos o critério de Chauvenet, a probabilidade de rejeitar uma observação equivocadamente (dá-la como "má" quando não o é) é de quase 40%. O que é inadmissível!

Em um recente trabalho de Monte Carlo não publicado, Dias e Castaño (1987) chegaram a conclusões nada alentadoras sobre este critério, inclusive no referente a sua capacidade de destacar observações contaminantes no modelo normal univariado.

Em 1868, Stone introduziu um teste baseado sobre um conceito de índole psicológica que chamou "módulo de cuidado", digamos m , fundamentado na idéia de que um determinado observador frente a uma certa situação amostral comete, em média, um erro em cada n observações que realiza. Como destacam Barnett e Lewis

(1984) e Hawkins (1980), este critério é similar ao de Chauvenet.

Até esse momento, parece que a forma de atuar diante de observações julgadas como anômalas era drástica: rejeitá-las da amostra. O mesmo Stone, um pouco mais tarde, propôs um método que poderíamos considerar como de "acomodamento de outliers", sugerindo simplesmente que se atribuisse um certo conjunto de pesos às observações; mas não formalizou sua proposta.

Parece que foi Glaisher, outro astrônomo, quem publicou em 1873, pela primeira vez, um procedimento do tipo "ponderação" ou "peso" de cada observação [ver detalhes em Barnett e Lewis (1984)].

Esse mesmo método foi publicado independentemente por Edgeworth em 1883, se bem que reconhecendo a prioridade de Stone no assunto.

Outro método de ponderação, baseado em um modelo que admitia explicitamente a possibilidade da equidistribuição dos erros de observação, foi proposto por Newcomb em 1886 [ver detalhes, por exemplo, em James e Bustos (1980)].

Em 1895, Mendeleev, o inventor da tabela periódica dos elementos, nos conta de

um procedimento, por ele adotado, para julgar se um conjunto de medições de uma certa variável era ou não "harmonioso". Este era análogo a uma espécie de técnica de "poda".

Wright (1984) sugeriu o seguinte procedimento: dadas n observações: $x_1, \dots, x_n$, rejeitar x_i se

$$|x_i - \bar{x}| < 3s$$

onde $\bar{x}$ é a média amostral e s é o desvio-padrão amostral. Mais tarde, já no início do Século XX, Wright e Hayford (1906) modificaram ligeiramente este critério e sugeriram algumas recomendações de índole prática para sua aplicação.

Devemos notar que esses critérios e alguns outros sugeridos já na segunda década deste século tinham um grave defeito: não levavam em consideração que uma coisa é a variância populacional e outra é a variância amostral. Ademais, consideravam o comportamento de toda a amostra, em vez de ocupar-se mais apropriadamente com os extremos da mesma. Possivelmente, quem primeiro publicou um trabalho corrigindo tais defeitos foi Irwin (1925).

BIBLIOGRAFIA

- ATKINSON, A.C. Plots, Transformations, and Regression. Clarendon Press, Oxford. 1985.
- BARNETT, V. Commentary on "Outlier...s" by Beckman R. J. and Cook, R.D., *Technometrics*, 25, 119-60. 1983.
- _____. LEWIS, T. Outliers in Statistical Data 2^a. ed., Wiley, New York. 1984.
- BECKMAN, R.J.; COOK, R.D., Outliers...s *Technometrics*, 25, 119-60. 1983.
- BERNOULLI, D. The Most Probable Choice Between Several Discrepant Observations and the Formation Therefrom of Most Likely Induction, In: ALLEN, C.G. (1961), *Biometrika*, 48, 3-13. 1777.
- CASTANO, E.; DIAS, R. Simulação do Teste de Rejeição de Outliers de Chauvenet Usando o Modelo de Contaminação por Slippage, manuscrito, IMPA. 1976.
- CHATTERJEE, S.; HADI, A.S. Influential Observations, High Leverage Points, and Outliers in Linear Regression, *Statistical Science*, 1, 379-416. 1986.
- COLLET, D.; LEWIS, T. The Subjective Nature of Outlier Rejection Procedures, *Applied Statistics*, 25, 228-37. 1976.
- COOK, R.D.; BECKMAN, R.J., Using M-estimators to Identify Outliers, Los Alamos Report LA-UR 80-1418, Los Alamos, New Mexico. 1980.
- GIBBONS, J.D.; FREUND, R.J., Organizations for Statistical Consultancy at Colleges and Universities, *Amer. Statist.*, 34, 140-45. 1980.
- GLAISHER, J.W.L. On the Law of Facility of Errors of Observation and the Method of Least Squares. *Memoires of the Royal Astronomical Society*, 39, 75-124. 1872.
- GORE, S.M. et alii, misuse of Statistical Methods: A Critical Assessment of Articles, In: BMJ from January to March 1976, *British Medical Journal*, 1, 85-7. 1977.

- HAMPEL, F.R. The Influence Curve and its Role in Robust Estimation, *J. Amer. Statist. Assoc.*, 69, 383-93. 1974.
- _____. The Breakdown Points of the Mean Combined with Some Rejection Rules, *Technometrics*, 27, 95-107. 1985.
- _____. Et alii. Robust Statistics, the Approach Based on Influence Functions, Wiley, New York. 1986.
- HAWKINS, D.M. Identification of Outliers Chapman and Hall, London. 1980.
- IGLEWICZ, R.; MARTINEZ, J. Outlier Detection Using Robust Measures of Scale, Technical Report (Theory) 12, Department of Statistics, School of Business Administration, Temple University, Philadelphia, Pennsylvania. 1982.
- IRWIN, J.O. On a Criterion for the Rejection of Outlying Observations, *Biometrika*, 17, 238-50. 1925.
- JAMES, K.L.; BUSTOS, O.H. Procedimentos Robustos, 4to. SIMAPE, 21 a 25 de julho de 1980, Rio de Janeiro. 1980.
- KERR, R.A. Test Fails to Confirm Cloud Seeding Effect, *Science*, 217, 234-36. 1982.
- LEROY, A.; ROUSSEUW, P.J. PROGRESS; A Program for Robust Regression, Research Report No. 201, University of Brussels, Centrum Voor Statistiek en Operationeel Onderzoek. 1984.
- MARTIN, R.D.; YOHAI, V.J. Robustness in Time Series and Estimating ARMA Models, In: HANDBOOK OF STATISTICS, Vol. 5, Hannan, E.J. Krishnaiah, P.R., and Rao, M.M. (eds), Elsevier Science Publishers B.V., New York. 1985.
- NEYMAN, J.; SCOTT, E.L. Outlier Proneness of Phenomena and of Related Distributions, In: OPTIMIZING METHODS, In: STATISTICS, Rustagi, J.S. (ed), Academic Press, New York. 1971.
- PRESCOTT, P. A Review of Some Robust Data Analysis and Multiple Outlier Detection Procedures, *BIAS*, 7, 141-58. 1980.
- RELLES, D.A.; ROGERS, W.H. Statisticians are Fairly Robust Estimators of Locations, *J. Amer. Statist. Assoc.*, 72, 107-11. 1977.
- TIKU, M.L. et alii. Robust Inference, Marcel Dekker Inc., New York. 1986.
- TIME, No. 42, October, 19, 1987, USA.
- WRIGHT, T.W. A Treatise on the Adjustment of Observations by the Method of Least Square, Van Nostrand, New York. 1884.
- _____.; HAYFORD, J.F. Adjustment of Observations, Van Nostrand, New York. 1906.

UMA NOVA METODOLOGIA PARA AVALIAÇÃO DE IMÓVEIS UTILIZANDO MODELOS LINEARES GENERALIZADOS

Rubens Alves Dantas*
Gauss Moutinho Cordeiro*

1 — CONSIDERAÇÕES GERAIS SOBRE ENGENHARIA DE AVALIAÇÃO

A avaliação de imóveis é feita usualmente através do método comparativo, pelo qual o valor de um bem é obtido por comparação com outros de características similares. Ocorre que, após a coleta dos elementos de referência, o avaliador está geralmente de posse de uma amostra aleatória formada por dados de características heterogêneas, sendo imprescindível a sua homogeneização para aplicação adequada do método comparativo. Observa-se no Brasil, que a grande maioria dos profissionais que atuam neste ramo, têm aplicado esta prática sem os cuidados necessários, utilizando-se de fatores empíricos e ponderações de ordem subjetiva, o que implica numa sensível perda do nível de precisão dos seus trabalhos. Um caminho para resolver a questão é, sem dúvida, fundamentar o laudo avaliatório na Metodologia da Pesquisa Científica.

A Pesquisa Científica tem como finalidade tentar conhecer e explicar a natureza dos fenômenos existenciais e desenvolve-se através das seguintes etapas: 1ª) preparação da pesquisa; 2ª) trabalho de campo; 3ª) processamento e análise dos dados; 4ª) interpretação e explicação dos resultados; 5ª) relatório da pesquisa.

Neste trabalho é abordado o assunto ligado à 3ª e 4ª etapas com a introdução da Teoria dos Modelos Lineares Generalizados na Homogeneização de Valores em Engenharia de Avaliação, utilizando-se o Sistema GLIM — Generalized Linear Interactive Modelling — desenvolvido pela Royal Statistical Society. A abordagem necessária para a realização das duas primeiras etapas pode ser encontrada em Dantas (1987, Capítulo 2).

Entende-se por Homogeneização de Valores o tratamento dispensado à amostra coletada, objetivando a retirada das discrepâncias existentes entre as características de cada elemento tomado como referência e o bem avaliando.

*Os autores são, respectivamente, Professor da Universidade Federal de Pernambuco e Caixa Econômica Federal e Professor do Centro de Ciências Sociais Aplicadas/UFPE (Departamento de Economia).

Na avaliação de lotes urbanos, por exemplo, é comum a homogeneização quanto aos aspectos relativos à contemporaneidade, testada efetiva, profundidade equivalente, localização e natureza da informação.

A este respeito, constam da norma para Avaliação de Imóveis Urbanos, NBR-5676, da Associação Brasileira de Normas Técnicas — ABNT, na alínea C, do subitem 7.2, entre outras, as seguintes recomendações: “O tratamento dispensado aos elementos para torná-los homogêneos possibilita conferir aos mesmos:

- equivalência no tempo, mediante a transposição dos valores dos elementos, considerando-se a desvalorização monetária e a valorização efetiva dos imóveis, infirida e justificada esta última por uma equação de regressão;

- equivalência de situação pela transposição dos elementos obedecendo a critérios tecnicamente consagrados e que se comprovem por instrumental matemático”.

A norma citada sugere, ainda, nos seus subitens 9.1.2.1 e 9.1.2.2, que para homogeneização quanto à profundidade, e testada, sejam utilizados modelos multiplicativos aos valores coletados, do tipo:

$$FP = \left(\frac{Pe}{Pa} \right)^{n_2} \quad \text{e} \quad FT = \left(\frac{Fr}{Fe} \right)^{n_1}$$

onde:

- Pe e Fe são as profundidades equivalentes e as frentes efetivas de cada Terreno de Referência;

- Pa e Fr se referem à profundidade equivalente e à frente de referência do lote avaliado;

- n_1 e n_2 — expoentes a serem determinados em função da região abrangida pela pesquisa.

Observando-se estes aspectos, apresenta-se, a seguir, um roteiro para construção de um modelo adequado para homogeneização de valores em engenharia de avaliações, utilizando a Teoria dos Modelos Lineares Generalizados.

Pretende-se, neste artigo, expor, de forma clara e objetiva, as informações necessárias ao engenheiro de avaliações para que, de posse do pacote GLIM, que já está

disponível para microcomputadores de 16 bits compatíveis com o IBM PC, e com alguns conhecimentos básicos de estatística e probabilidade, tenha condições de tratar os dados amostrais de maneira satisfatória. Propositadamente, evitaram-se, neste trabalho, demonstrações e formulações mais complicadas, citando-se as fontes de consulta no momento oportuno. Contudo, tentou-se explorar mais a parte conceitual e de interpretação de resultados.

2 — HOMOGENEIZAÇÃO UTILIZANDO A TEORIA DOS MODELOS LINEARES GENERALIZADOS

2.1 — Considerações Gerais

O Modelo Clássico de Regressão teve origem nos trabalhos de Astronomia elaborados por Gauss no período de 1809 a 1821. Atualmente, a análise de regressão é um dos ramos da teoria estatística mais utilizados na pesquisa científica. É a técnica adequada quando se deseja estudar o comportamento de uma variável (variável dependente) em relação a outras que são responsáveis pela sua formação (variáveis independentes).

Em engenharia de avaliações considera-se, geralmente, como variável dependente o valor unitário à vista do imóvel pesquisado em $Cz\$/m^2$ e como variáveis independentes as respectivas características decorrentes dos aspectos físicos e de localização, bem como a natureza da informação (valor proveniente de oferta ou de negociação efetivamente realizada) e a época de ocorrência do evento.

Observa-se que as variáveis independentes podem ser tanto de natureza *quantitativa* (área, testada efetiva etc.) como de ordem *qualitativa* (natureza da informação, orientação etc.).

A utilização do Modelo Clássico de Regressão, contudo, envolve a comprovação de alguns pressupostos básicos, tais como: variância constante e distribuição normal para o erro. Nelder e Wedderburn (1972) desenvolveram uma classe de modelos bem mais ampla, da qual o Modelo Clássico de

Regressão é um caso particular. Nesses modelos, denominados de Modelos Lineares Generalizados-MLGs, os pressupostos citados não são mais exigidos.

Um MLG é definido por uma distribuição de probabilidade, membro da família exponencial de distribuições, para a variável resposta (componente aleatória), um conjunto de variáveis independentes descrevendo a estrutura linear do modelo (componente sistemática) e uma função de ligação (f) entre a média da variável de resposta (μ) e a estrutura linear (η), digamos $f(\mu) = \eta$.

A análise de dados através dos MLGs é bastante flexível, pois para uma mesma estrutura linear pode-se obter vários modelos, dependendo da distribuição proposta para o erro e a função de ligação escolhida.

A idéia central dos MLGs é transformar as médias dos dados ao invés de transformar as observações, como nos modelos de Box e Cox (1964).

A análise de um MLG é facilitada pela utilização do pacote estatístico GLIM. O sistema GLIM possui um bom manual (Baker e Nelder, 1978) com resumo das suas instruções e, também, aplicações a dados reais.

Apresentam-se, a seguir, de forma resumida, as noções básicas para definição de um MLG.

2.2 — A Componente Aleatória

A componente aleatória de um MLG considera que se dispõe de um vetor de observações $y = (y_1, \dots, y_n)^T$ como realização das variáveis aleatórias $Y = (Y_1, \dots, Y_n)^T$ independentes ou pelo menos não correlacionadas, com médias $\mu = (\mu_1, \dots, \mu_n)^T$, pertencente à família exponencial de distribuições. Esta família tem densidade de probabilidade na forma

$$\pi(y; \Theta_i, \phi_i) = e^{\phi_i [y\Theta_i - b(\Theta_i) + c(y, \phi_i)]} \quad (2.1)$$

$i = 1, \dots, n$

onde $b(\cdot)$ e $c(\cdot, \cdot)$ são funções conhecidas para cada observação, ϕ_i é um parâmetro de dispersão suposto conhecido e Θ_i é o parâmetro natural (ou canônico) que caracteriza a distribuição em (2.1).

Pode-se demonstrar, facilmente, que

$$E(Y) = \mu = \frac{d}{d\Theta} b(\Theta) \quad \text{e} \quad \text{Var}(Y) = \frac{V}{\phi}$$

sendo $V = \frac{d\mu}{d\Theta}$ denominada fundação de variância.

É fácil verificar que várias distribuições de probabilidade conhecidas e largamente utilizadas como a normal, Poisson, binomial, gama e normal inversa, são membros da família exponencial de distribuições.

Para a escolha de uma distribuição razoável para o erro, o analista deve examinar os dados, principalmente, quanto aos seguintes aspectos: assimetria, natureza contínua ou descontínua e intervalo de variação. Por exemplo, se o campo de variação dos dados é o conjunto dos reais, a distribuição escolhida deve ser a normal; para dados contínuos positivos apresentando assimetria com variâncias superiores às médias é bastante razoável supor a distribuição gama ou a normal inversa; a distribuição de Poisson é apropriada para dados discretos na forma de contagens, podendo ser, também, utilizada para dados contínuos, quando a variância for aproximadamente igual à média; a distribuição binomial serve para análise de dados na forma de proporções ou, ainda, para dados apresentando subdispersão.

2.3 — A Componente Sistemática

A componente sistemática de um MLG, $\eta = (\eta_1, \dots, \eta_n)^T$, também chamado de preditor linear, é uma função linear dos parâmetros desconhecidos $\beta = (\beta_1, \dots, \beta_p)^T$, representada por

$$\eta = X\beta$$

sendo X a matriz modelo de ordem $n \times p$, suposta de posto completo.

A escolha de um preditor linear adequado pode envolver técnicas de seleção de covariáveis, como, por exemplo, o método das covariáveis adicionais (Cordeiro, 1986, seção 4.8).

Para formação da matriz modelo, deve-se escolher um modelo intermediário entre o minimal que contém o menor número de termos necessários para o ajustamento e o modelo maximal que inclui o maior número de termos que o analista pretende considerar. Os termos destes modelos extremos são, geralmente, obtidos por informações *a priori* da estrutura dos dados.

2.4 — A Função de Ligação

A função de ligação deve ser compatível com a distribuição proposta para o erro e deve ser escolhida de tal forma a facilitar a interpretação do modelo.

No Modelo Linear Clássico, por exemplo, como os valores dos dados e preditores podem assumir qualquer valor real, a ligação identidade deve ser a escolhida. Entretanto, para os modelos gama, Poisson e normal inverso, a ligação identidade é menos atrativa, uma vez que não restringe os valores esperados ao intervalo $(0, \infty)$. Por outro lado, se efeitos sistemáticos multiplicativos contribuem para as médias dos dados, uma ligação logarítmica, que torna os efeitos aditivos contribuindo para os preditores lineares, pode ser a mais apropriada. Entretanto, a escolha da melhor função de ligação ainda é um problema em aberto.

Para cada tipo de distribuição existe uma ligação denominada canônica, que torna os efeitos sistemáticos aditivos na escala dada pela ligação. Embora estas ligações forneçam propriedades estatísticas de interesse para o modelo, tais como suficiência, facilidade de cálculo e unicidade das estimativas de máxima verossimilhança, não há razão para se trabalhar sempre com elas. Na prática, utiliza-se a ligação canônica quando não há indicativo de outra preferível.

Para os modelos normal, binomial, Poisson, gama e normal inverso, têm-se as respectivas ligações canônicas: identidade ($\eta = \mu$), logística ($\eta = \log[\mu/(1 - \mu)]$), logarítmica ($\eta = \log \mu$), recíproca ($\eta = \mu^{-1}$) e inversa do quadrado ($\eta = \mu^{-2}$).

2.5 — A Estimação dos Parâmetros

A estimação do vetor de parâmetros desconhecidos de um MLG, $\beta = (\beta_1, \dots, \beta_p)^T$ é feita utilizando-se o método da máxima verossimilhança, que consiste em encontrar a log-verossimilhança para o modelo $L(\beta)$ e, em seguida, resolver o sistema $U(\hat{\beta}) = 0$, onde $U(\beta)$ é conhecido como função escore ou função suporte, sendo dada por

$$U(\beta) = \frac{\partial L(\beta)}{\partial \beta} = \left(\frac{\partial L(\beta)}{\partial \beta_1}, \dots, \frac{\partial L(\beta)}{\partial \beta_p} \right)^T \quad (2.2)$$

com

$$\frac{\partial U(\beta)}{\partial \beta_r} = \sum_{i=1}^n \phi_i [y_i - \mu_i] W_i \frac{d\eta_i}{d\mu_i} x_{ir}, \quad r = 1, \dots, p \quad (2.3)$$

onde $w = (d\mu/d\eta)^2/V$ é denominada função peso.

Quando o erro é normal e a ligação, a identidade, tem-se o Modelo Clássico de Regressão como um caso particular de um MLG e o sistema (2.3) é formado por equações lineares. Nos demais casos, têm-se um sistema de equações não-lineares, e métodos numéricos iterativos, como o de Newton-Raphson e o de Fisher (conhecido, também, como método do escore para parâmetros), são necessários para obtenção da estimativa $\hat{\beta}$.

Nelder e Wedderburn (1972) demonstraram que a solução destas equações é equivalente a um procedimento iterativo de mínimos quadrados ponderados e o sistema GLIM utiliza, essencialmente, este procedimento.

2.6 — Verificação da Qualidade do Ajustamento

O teste de ajustamento de um MLG, com p parâmetros a partir de n observações, é feito através da estatística

$$S_p = 2(\hat{L}_n - \hat{L}_p) \quad (2.4)$$

onde S_p é chamado de desvio do modelo e $\hat{L}_n$ e $\hat{L}_p$ são os máximos da log-verossimilhança para os modelos saturado (número de observações igual ao número de parâmetros β 's) e em investigação ($p < n$), respectivamente.

O desvio é uma medida da distância entre os valores ajustados e observados ou, equivalentemente, do modelo em investigação ao saturado.

Embora pouco seja conhecido sobre a distribuição do desvio, na prática compara-se S_p com o valor crítico $\chi^2_{n-p; \alpha}$ da distribuição qui-quadrado a um nível de significância α . Se $S_p \leq \chi^2_{n-p; \alpha}$, o modelo é aceito e, em caso contrário, rejeitado ao dado nível α . Apenas para o modelo normal-linear S_p tem distribuição qui-quadrado exata. Para outros modelos existem, apenas, resultados assintóticos (Cordeiro, 1983).

2.7 — Análise do Desvio

A análise do desvio em um MLG tem como objetivo testar modelos encaixados, isto é: um modelo com p_1 parâmetros (M_{p_1}) *versus* outro maior com p_2 parâmetros (M_{p_2}), $p_1 < p_2$, ambos com mesma ligação e distribuição. Isto é equivalente a testar a hipótese nula

$$H_0 : \beta_{p_1+1} = \dots = \beta_{p_2} = 0.$$

Para isto calcula-se o desvio dos modelos M_{p_1} e M_{p_2} e compara-se a diferença ($S_{p_1} - S_{p_2}$) com o ponto crítico da distribuição qui-quadrado com $(p_2 - p_1)$ graus de liberdade a um nível de significância especificado α . Se $S_{p_1} - S_{p_2} > \chi^2_{p_2-p_1; \alpha}$, rejeita-se H_0 e o conjunto de parâmetros $\beta_{p_1+1}, \dots, \beta_{p_2}$ é significativo para dar maior poder de explicação ao modelo; em caso contrário, não se pode rejeitar H_0 e o modelo M_{p_1} pode ser o escolhido para explicar os dados.

A estatística ($S_{p_1} - S_{p_2}$) é interpretada como uma medida de variação dos dados, explicada pelos termos que estão em M_{p_2} e não estão em M_{p_1} , incluídos os efeitos dos termos em M_{p_1} e ignorando quaisquer efeitos dos termos que não estão em M_{p_2} .

Quando se deseja analisar vários modelos encaixados $M_{p_1}, M_{p_2}, \dots, M_{p_r}$, com dimensões respectivas $p_1 < p_2 < \dots < p_r$ e desvios $S_{p_1} > S_{p_2} > \dots > S_{p_r}$, tendo os modelos a mesma ligação e distribuição, geralmente, constrói-se uma tabela de análise do desvio, conhecida como tabela de "ANODEV". Aqui é feita uma análise de variância generalizada nos mesmos moldes das clássicas tabelas ANOVA.

2.8 — Técnicas de Diagnóstico

Por técnicas de diagnóstico entende-se a análise dos resíduos para detectar observações aberrantes e a análise da influência de observações sobre o ajustamento global do modelo.

Para a análise dos resíduos, várias estatísticas podem ser adotadas. Cox e Snell (1968) propuseram uma expressão dos resíduos generalizados correspondentes à variável aleatória Y_i da forma

$$R_i = h_i(y_i, \hat{\mu}_i) \quad (2.5)$$

onde h_i é uma função escolhida de modo a estabilizar a variância ou induzir simetria na

distribuição amostral de R_i , e y_i e $\hat{\mu}_i$ são os respectivos valores observados e ajustados, correspondentes a Y_i . Várias formas para a função $h(\cdot, \cdot)$ têm sido propostas.

O resíduo de Pearson é o mais simples e definido por $P_i = \phi_i^{1/2}(y_i - \hat{\mu}_i)/\hat{V}_i^{1/2}$. A desvantagem deste resíduo é que sua distribuição é, geralmente, bastante assimétrica para modelos não-normais.

Da distribuição assintótica de $\hat{\beta}$ com estrutura de covariância $(X^T W \Phi X)^{-1}$ pode-se concluir que $\text{Var}(\hat{\mu}_i)$ e $\text{Var}(Y_i - \hat{\mu}_i)$ são, aproximadamente, iguais a $V_i h_{ii}$ e $V_i (1 - h_{ii})$, respectivamente, onde h_{ii} é o i -ésimo elemento da diagonal da matriz $H = \Phi^{1/2} W^{1/2} X (X^T W \Phi X)^{-1} X^T W^{1/2} \Phi^{1/2}$. Esta matriz desempenha, na teoria dos MLGs, o mesmo papel da matriz *hat* do Modelo Clássico de Regressão. Se o parâmetro de dispersão é constante, este não entra na formação de H .

As considerações anteriores conduzem a definição do resíduo de Pearson studentizado com expressão

$$t_i = \Phi_i^{1/2}(y_i - \hat{\mu}_i)/\{\hat{V}_i(1 - \hat{h}_{ii})\}^{1/2} \quad (2.6)$$

onde todas as quantidades estimadas estão avaliadas em $\hat{\beta}$. O Resíduo (2.6) é mais informativo que P_i .

Outros tipos de resíduos, como o resíduo de Anscombe (1953) e aqueles definidos como componentes do desvio, têm interesses específicos.

O resíduo de Anscombe (1953) visa à normalização e estabilização da variância, sendo dado por

$$A_i = \Phi_i^{1/2} [N(y_i) - N(\hat{\mu}_i)]/N'(\hat{\mu}_i)\hat{V}_i^{1/2}$$

onde N é uma função escolhida de tal forma a aproximar a distribuição de R_i à normal e $N'(\mu) (V/\Phi)^{1/2}$ é uma aproximação para o desvio padrão de $N(Y)$.

O resíduo como componente do desvio, que tem o objetivo de normalizar R_i , é utilizado usualmente na forma

$$D_i = \delta_i \cdot d_i^{1/2}$$

sendo δ_i o sinal de $(y_i - \hat{\mu}_i)$ e d_i a i -ésima componente de

$$S_p = 2 \sum_{i=1}^n \phi_i \{y_i(\tilde{\Theta}_i - \hat{\Theta}_i) - b(\tilde{\Theta}_i) + b(\hat{\Theta}_i)\},$$

onde $\tilde{\Theta}_i = q(y_i)$, $\hat{\Theta}_i = q(\hat{\mu}_i)$, e $\Theta = q(\mu)$ é a função que relaciona o parâmetro natural com a média.

A vantagem do resíduo D_i é que não requer o conhecimento da função normalizadora.

Observa-se que para modelos bem ajustados as diferenças entre D_i e t_i são pequenas, enquanto que, de uma forma geral, os resultados com o uso de D_i e A_i são similares.

A idéia básica de influência é verificar a dependência do modelo estatístico sobre as várias observações ajustadas. Se com a eliminação de observações ocorrer uma pequena perturbação, em algumas observações, conduzir a uma mudança apreciável nas estimativas dos parâmetros β 's, estas observações poderão ser consideradas influentes. As medidas usuais de influência geralmente correspondem às variações das estimativas decorrentes da eliminação de variáveis.

Para exemplificar, a influência da observação Y_i sobre o modelo é obtida pela distância de $\hat{\beta}$ a $\hat{\beta}_{(i)}$, onde $\hat{\beta}_{(i)}$ é a estimativa de β se γ_i é eliminada do modelo. Nos modelos lineares generalizados podemos usar as seguintes medidas desta distância:

$$(\hat{\beta}_{(i)} - \hat{\beta})^T [\text{Cov}(\hat{\beta})]^{-1} (\hat{\beta}_{(i)} - \hat{\beta}) \quad (2.7a)$$

$$2 \{L(\hat{\beta}) - L(\hat{\beta}_{(i)})\} \quad (2.7b)$$

ambas estatísticas tendo distribuição assintótica qui-quadrado com um grau de liberdade. Pode-se demonstrar que estas medidas são, aproximadamente, iguais à estatística de Cook (1977) do Modelo Clássico de Regressão, dada por

$$C_i = \frac{(y_i - \hat{\mu}_i)^2 \hat{h}_{ii}}{\hat{V}_i (1 - \hat{h}_{ii})^2} = \left(\frac{\hat{h}_{ii}}{1 - \hat{h}_{ii}} \right) t_i^2 \quad (2.8)$$

Um gráfico de C_i versus i indicará as observações influentes no modelo adotado. Considerações adicionais podem ser vistas em Atkinson (1986).

2.9 — Seleção do Modelo

É aconselhável que na análise de um conjunto de dados se formule vários modelos plausíveis, alternando-se a ligação e o tipo de distribuição, adicionando-se ou retirando-se vetores colunas independentes de uma matriz modelo original ou, ainda, considerando transformações adequadas aos dados brutos.

A escolha da função de transformação para os dados não é tão crucial como nos modelos clássicos de regressão, pois nos MLGs as hipóteses de normalidade e constância da variância não são essenciais para a distribuição da variável resposta e,

ainda, pode-se achar uma estrutura aditiva aproximada de termos para representação da média da distribuição, usando uma ligação apropriada, diferente da escala de medição dos dados.

Uma forma simples de verificar a adequação da ligação é pela inclusão da covariável $\hat{\eta} \otimes \hat{\eta}$ no modelo, onde $\otimes$ é o produto direto, isto é, componente vezes componente. Se a redução causada no desvio for significativa, a ligação não é apropriada. Outra maneira seria construir o gráfico da variável dependente modificada y^* versus $\hat{\eta}$, que sendo aproximadamente linear indica que a ligação escolhida é adequada.

Pregibon (1980), usando as idéias da família de transformação para dados desenvolvidas por Box e Cox (1964), definiu uma família de ligações imprimindo transformações às médias, do tipo $f(\mu, \lambda) = (\mu^\lambda - 1)/\lambda$.

Para definição do conjunto de variáveis independentes a serem incluídas na estrutura linear, além do método das covariáveis adicionais, citado anteriormente, outros métodos de seleção, como o "STEPWISE" e o critério C_p de Mallows (1973), podem ser utilizados.

3 — UMA ANÁLISE DE DADOS REAIS VIA O GLIM

Analisa-se uma amostra composta por 50 (cinquenta) dados de lotes urbanos, situados nos bairros de Casa Forte, Torre e Iputinga, em Recife, PE, utilizados pelo extinto BNH para avaliação de diversas glebas, a serem adquiridas para implantação do Projeto Recife — Programa de Revitalização do Capibaribe — Subprograma Áreas Vazias.

Tenta-se, a partir da amostragem, construir um conjunto de modelos plausíveis para explicar a formação do valor de lotes nos bairros em pauta, utilizando o pacote GLIM.

Para a construção dos modelos, consideram-se como variável dependente os valores unitários dos terrenos coletados em Cz\$/m² (V), e como covariáveis as suas respectivas características físicas de testada efetiva (T), profundidade equivalente (P) e nível de urbanização (U). Leva-se em conta, também, a localização: Iputinga (R_1), Torre (R_2) e Casa Forte (R_3), e, ainda, a natu-

reza do evento (N) e a época da sua ocorrência (C), obtendo-se, inicialmente, um modelo do tipo:

$$V_i = \beta_0 + \beta_1 T_i + \beta_2 P_i + \beta_3 U_i + \beta_4 N_i + \beta_5 C_i + \beta_6 R_{1i} + \beta_7 R_{2i} + \beta_8 R_{3i} + \epsilon_i \quad (3.1)$$

Trata-se o nível de urbanização como uma variável qualitativa variando de 1 a 8, que cresce com o nível de beneficiamento existente na vizinhança imediata do terreno; a natureza do evento como uma variável *dummy*, sendo 0 (zero) para ofertas e 1 (um) para negociações; a contemporaneidade como uma variável quantitativa representada pelo número de meses ocorridos entre uma data base (Dez./79) até a de ocorrência do evento; e a localização, também, como uma variável do tipo 0-1.

Apresentam-se, a seguir, comentários adicionais aos resultados e conclusões obtidos da análise desses dados, cuja listagem encontra-se no APÊNDICE 1. No APÊNDICE 2 são apresentadas as principais informações sobre as diretivas do GLIM utilizadas.

Pela listagem observa-se que foram construídos 4 (quatro) modelos. Inicialmente (análise 1) tentou-se ajustar o modelo clássico de regressão que corresponde ao erro normal e a ligação identidade, sendo fortemente rejeitado, visto que o desvio apresentou-se extremamente grande; em seguida, tentou-se uma transformação adequada para a variável resposta, através do teste de Tukey (1949), verificando-se que a não-aditividade foi significativa ao nível de 5%. Calculou-se, então, o expoente da transformação potência que sugeriu a transformação logarítmica. Esta transformação pode, também, ser confirmada pelo método de Box e Cox (1964).

Constrói-se um segundo modelo (análise 2) com transformação logarítmica na variável resposta e verifica-se uma sensível redução no desvio para 6.432, sendo aceito ao nível de 0,5%, tendo em vista que $\chi^2_{42;0.05} = 20,71$ é superior a 6.432.

Substituindo-se as estimativas dos parâmetros no modelo de regressão e explicitando-se $\hat{V}_i$, encontra-se:

$$\hat{V}_i = 423,26 \cdot 1,013^{T_i} \cdot 0,999^{P_i} \cdot 0,835^{N_i} \cdot 1,057^{C_i} \cdot 1,140^{U_i} \cdot 0,648^{R_{1i}} \cdot 0,804^{R_{2i}} \quad (3.2)$$

Por este modelo, observa-se que, na região abrangida pela pesquisa, os valores de lotes crescem de forma diretamente proporcional à testada, ao nível de urbanização e ao tempo; decrescem com o aumento da profundidade equivalente e são mais valorizados no sentido Iputinga, Torre e Casa Forte. Verifica-se ainda que, em média, os terrenos são negociados por 83,50% do valor de oferta inicial, a taxa de valorização territorial no período pode ser estimada em 5,7% a.m., os valores unitários crescem de 14% a cada nível de urbanização e são em Iputinga e Torre 64,8% e 80,4% dos valores de terrenos similares em Casa Forte, respectivamente. Todos os resultados são bastante coerentes com a crença *a priori* que se tinha destes efeitos. A estimativa do parâmetro da covariável (R_{3i}) foi igual a zero, tendo em vista que este vetor coluna da matriz modelo é uma combinação linear de R_{1i} e R_{2i} . Neste caso, a estimativa foi absorvida pela média geral %GM.

O objetivo principal da análise 3 é estimar os expoentes dos fatores testada e profundidade, obedecendo aos modelos sugeridos pela NBR-5676.

Assim, impõe-se a transformação logarítmica a estas duas covariáveis e ajusta-se o modelo a seguir, observando-se resultados similares aos encontrados pelo modelo anterior.

$$\hat{V}_i = 304,29 \cdot T_i^{0,260} \cdot P_i^{-0,344} \cdot 0,814^{N_i} \cdot 1,055^{C_i} \cdot 1,141^{U_i} \cdot 0,650^{R_{1i}} \cdot 0,810^{R_{2i}} \quad (3.3)$$

Verifica-se, desta forma, expoentes $n_1 = 0,260$ para o fator testada e $n_1 = -0,344$ para o fator profundidade, na região pesquisada. Na prática, estes fatores são considerados, pela grande maioria dos avaliadores em todo o Brasil, iguais a 0,25 e $-0,50$, respectivamente. O fator multiplicativo ao valor de oferta inicial, visando a retirar a superestimativa que geralmente acompanha as ofertas, denominado fator fonte, foi adotado igual a 0,80 pelo BNH para as diversas avaliações que se processam nos bairros considerados. Pode-se notar que este fator é bastante coerente com o estimado por este modelo, que é igual a 0,814.

Na análise 4, faz-se a tentativa através do modelo gama, tendo em vista que a própria

natureza dos dados, todos situados no campo dos reais positivos, sugere esta opção.

Com o objetivo de encontrar uma ligação adequada, constroem-se gráficos da variável dependente modificada (y^*) *versus* preditores lineares estimados ($\hat{\eta}$), considerando-se a identidade, a logarítmica e a recíproca (canônica), sendo a logarítmica a que apresentou resultados mais favoráveis.

Desta forma, ajusta-se aos dados um modelo gama com ligação logarítmica, que apresenta resultados bem próximos dos encontrados nas análises 2 e 3, permanecendo os mesmos sinais para as estimativas. Os resíduos são, igualmente, pequenos, sendo que neste modelo a observação 17 parece estar mais bem ajustada.

Construíram-se ainda os gráficos dos resíduos de Anscombe *versus* os quantis da distribuição normal, indicando razoável aceitação da hipótese de erro gama para os dados.

O gráfico dos resíduos de Anscombe *versus* valores ajustados pelo modelo apresentaram pontos dispersos aleatoriamente, sem nenhum padrão definido, podendo-se concluir que os erros são distribuídos independentemente.

O processamento completo dos dados encontra-se com os autores, podendo ser fornecido aos interessados.

4 — CONCLUSÃO

Os resultados encontrados neste trabalho deixam evidente que a aplicação adequada do método comparativo deve estar fundamentada na Metodologia da Pesquisa Científica. Neste sentido, dispõe o avaliador da Teoria dos Modelos Lineares Generalizados, que, embora ainda pouco difundida no Brasil, é o ferramental estatístico mais poderoso da atualidade, disponível para análise de dados univariados. A construção de um MLG, porém, torna-se bastante facilitada com a utilização do sistema GLIM, que pode ser implementado em qualquer microcomputador de 16 *bits* com co-processador

aritmético (8087/80287) e pelo menos 350K RAM.

O tratamento dispensado a dados de lotes de 3 (três) bairros do Recife, ampliado para toda a cidade, forneceria a base para montagem de uma Planta Genérica de Valores Municipais, de grande importância para estipulação das cotas de Imposto Predial e Territorial Urbano. Este procedimento poderá ser aplicado para qualquer cidade do país.

Torna-se difícil, contudo, a escolha de um único modelo mais adequado aos dados analisados, pois a complexidade dos mesmos fica mais bem explicada pelo conjunto de modelos, obtidos nas análises 2, 3 e 4.

APÊNDICE 1

ANÁLISE DE DADOS DE LOTES DE TRÊS BAIRROS DO RECIFE

```
$ DATA 50 T P N C U R 1 R 2 R 3 V
$ UNITS 50
$ DINPUT 43
$ CAL B = V $
```

19	30	0	31	6	1	0	0	2983
17	29	1	22	7	1	0	0	2695
36	29	0	32	6	1	0	0	3831
10	20	0	26	4	1	0	0	2250
12	36	1	25	6	0	1	0	3588
11	31	1	20	7	0	1	0	1261
10	30	1	14	8	0	1	0	1587
15	20	0	31	4	0	1	0	2667
10	30	0	31	4	0	1	0	3333
13	36	0	31	6	0	1	0	4273
16	34	0	52	7	0	1	0	29994
9	52	1	27	8	0	0	1	4288
24	39	0	31	6	0	0	1	7478
15	82	0	31	8	0	0	1	10339
36	39	1	3	6	0	0	1	2422
16	135	1	32	8	0	0	1	6734
22	22	1	39	7	0	0	1	4019
40	233	1	50	8	0	0	1	16094
14	84	1	42	8	0	0	1	9267
16	33	0	54	8	0	0	1	32567
17	27	1	45	6	0	0	1	9918
17	30	1	29	8	0	0	1	7843
12	247	1	26	8	0	0	1	7093

17	21	0	55	5	0	0	1	29790
25	37	0	55	5	0	0	1	32258
24	30	0	44	5	1	0	0	4047
14	26	1	29	5	1	0	0	3710
12	40	0	44	5	1	0	0	6250
13	30	0	44	8	1	0	0	10970
14	28	0	43	4	1	0	0	5639
13	41	0	43	7	1	0	0	9259
10	25	0	44	7	1	0	0	10737
9	22	0	43	4	1	0	0	4067
15	32	0	43	5	1	0	0	5208
15	30	0	32	5	1	0	0	3333
15	64	0	32	5	1	0	0	2083
66	29	0	8	5	1	0	0	2712
15	25	0	31	5	1	0	0	5333
14	24	1	23	7	1	0	0	3535
25	80	1	20	7	1	0	0	2306
18	30	1	44	7	1	0	0	13789
10	28	1	43	6	1	0	0	10714
9	30	0	57	8	0	1	0	30476
14	30	0	57	3	0	1	0	28571
14	71	0	57	7	0	1	0	35211
9	25	0	57	4	0	1	0	14846
15	66	0	57	6	0	1	0	15344
16	25	1	49	7	0	1	0	8040
18	23	0	54	5	0	0	1	24154
14	30	0	30	5	1	0	0	3333

	ESTIMATE	S.E.	PARAMETER
1	9897.	5806.	%GM
2	-131.7	79.33	T
3	10.12	15.13	P
4	-132.5	110.4	C
5	398.2	1612.	N
6	-543.9	575.4	U
7	1945.	1943.	R1
8	-1672.	1664.	R2
9	ZERO	ALIASED	R3
10	0.5716E-04	0.8650E-05	FV2
SCALE PARAMETER TAKEN AS			0.1523E+08

A não-aditividade é significativa ao nível de 5%, pois a estimativa do parâmetro correspondente à covariável adicional FV2 mostrou-se significativa a este nível, com $0.5396E-04/0.8733E-05=6.18$, resultado superior ao ponto crítico da distribuição T com 40 graus T $(40;0.05)=2.33$, e alguma transformação na variável resposta é necessária.

\$C Cálculo do expoente da função potência
\$CAL %Q=1-2*0.00005396*%CU(V)/%NU \$PRINT %Q \$
-0,0840
\$C

Como %Q é próximo de zero, sugere a transformação LOG
Verificação da transformação LOG pelo método de BOX e COX
\$CAL LT = %LOG(V) :
%G = %EXP(%CU(LT)/%NU) \$
\$MACRO BOXCOX \$CAL
%D = %L * %G ** (%L - 1) \$
\$CAL B = (V ** %L - 1)/%D \$FIT.
\$ENDMAC
\$CAL %L = 1 \$FIT
T + P + N + U + C + R1 + R2 + R3
\$USE BOXCOX \$

CYCLE	DEVIANCE	DF
1	0.1290E+10	42
CURRENT DISPLAY INHIBITED		
CYCLE	DEVIANCE	DF
1	0.1290E+10	42
\$C Como o desvio coincide com o anterior, a macro está correta		
\$CAL %L=.5 \$USE BOXCOX \$CAL %L=.1 \$USE BOXCOX \$		
CURRENT DISPLAY INHIBITED		
CYCLE	DEVIANCE	DF
1	0.5010E+09	42

ANÁLISE 1

\$C ANÁLISE PELO MODELO NORMAL LINEAR \$
\$YVAR B \$FIT T+P+C+N+U+R1+R2+R3 \$

CYCLE	DEVIANCE	DF
1	0.1290E+10	42

\$C

Este modelo é fortemente rejeitado, uma vez que o desvio apresentou-se bastante grande

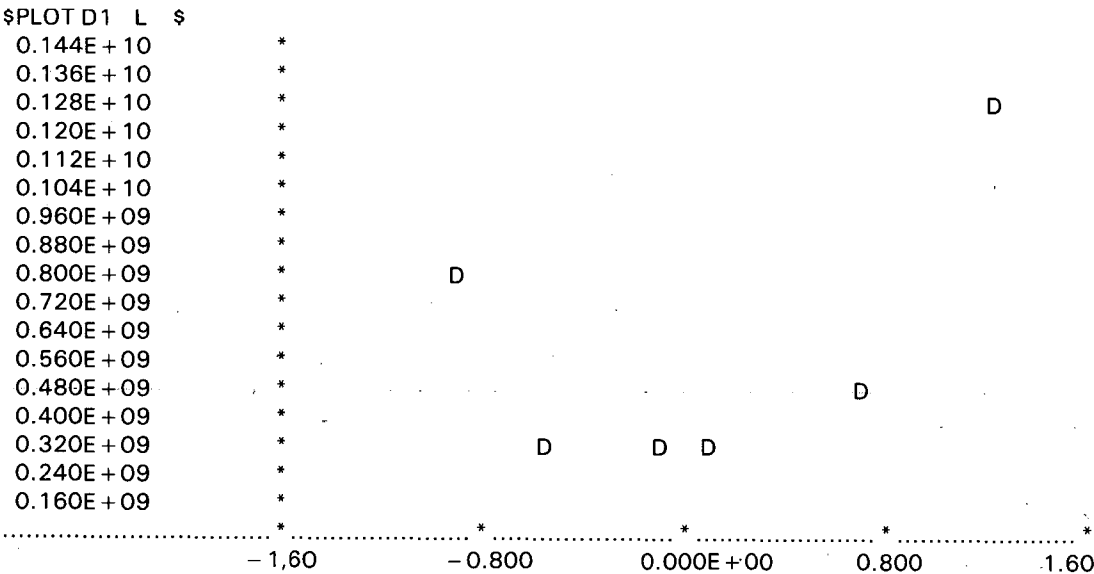
\$C Verificação da não-aditividade através do teste de TUKEY
\$CAL FV2 = %FV * %FV \$FIT + FV2 \$
DIS ME \$

\$C
CYCLE DEVIANCE DF
1 0.6244E+09 41
Y-VARIATE B
ERROR NORMAL LINK IDENTITY
LINEAR PREDICTOR
%GMT P C N U R1 R2 R3 FV2

— — — CURRENT DISPLAY INHIBITED
CYCLE DEVIANCE DF
1 0.3069E+09 42
\$CAL %L = -.1 \$USE BOXCOX \$CAL
%L = .5 \$USE BOXCOX \$
— — — CURRENT DISPLAY INHIBITED
CYCLE DEVIANCE DF
1 0.2815E+09 42
— — — CURRENT DISPLAY INHIBITED
CYCLE DEVIANCE DF
1 0.3463E+09 42
\$CAL %L = - 1 \$USE BOXCOX \$

— — — CURRENT DISPLAY INHIBITED
CYCLE DEVIANCE DF
1 0.7506E+09 42
\$VAR 6 L D1 \$CAL
L(1) = 1:D1(1) = 1288000000 \$
\$CAL L(2) = .5:D1(2) = 508200000
:L(3) = .1:D1(3) = 316300000 \$
\$CAL L(4) = -.1:D1(4) = 291900000
:L(5) = -.5:D1(5) = 359600000 \$
\$CAL L(6) = - 1:D1(6) = 773700000 \$
\$C

GRÁFICO DOS DESVIOS VERSUS VALORES DE L



\$C

Como o ponto de mínimo deste gráfico está próximo de zero, pode-se aceitar a transformação LOG como adequada

ANÁLISE 2

\$C ANÁLISE ATRAVÉS DO MODELO
NORMAL LINEAR COM
TRANSFORMAÇÃO LOG NA VARIÁVEL
DEPENDENTE.
\$CAL Z = %LOG(V) \$YVAR Z \$
— — — CURRENT DISPLAY INHIBITED
\$FIT T+P+N+C+U+R1+R2+R3 \$DIS
MERC \$
CYCLE DEVIANCE DF
1 6.432 42
Y-VARIATE Z
ERROR NORMAL LINK IDENTITY

LINEAR PREDICTOR
%GMT PNCUR1 R2 R3

	ESTIMATE	S.E.	PARAMETER
1	6.128	0.4329	%GM
2	0.1261E-01	0.6403E-02	T
3	-0.5282E-03	0.1446E-02	P
4	-0.1927	0.1511	N
5	0.5570E-01	0.5051E-02	C
6	0.1268	0.4976E-01	U
7	-0.4788	0.1531	R1
8	-0.2913	0.1659	R2
9	ZERO	ALIASED	R3
SCALE PARAMETER TAKEN AS 0.1531			

UNIT	OBSERVED	FITTED	RESIDUAL
1	8.001	8.360	-0.3593
2	7.899	7.768	0.1311
3	8.251	8.631	-0.3797
4	7.719	7.720	-0.1038E-02
5	8.185	7.929	0.2562
6	7.140	7.767	-0.6278
7	7.370	7.548	-0.1784
8	7.889	8.249	-0.3601
9	8.112	8.180	-0.6883E-01
10	8.360	8.469	-0.1086
11	10.31	9.804	0.5047
12	8.364	8.539	-0.1755
13	8.920	8.897	0.2266E-01
14	9.244	9.014	0.2293
15	7.792	7.296	0.4962
16	8.815	8.862	-0.4713E-01
17	8.299	9.261	-0.9618
18	9.686	10.12	-0.4294
19	9.134	9.421	-0.2866
20	10.39	10.33	-0.5705E-01
21	9.202	9.402	-0.2002
22	8.967	8.763	0.2044
23	8.867	8.418	0.4486
24	10.30	10.03	0.2736
25	10.38	10.12	0.2607
26	8.306	9.020	-0.7146
27	8.219	7.868	0.3506
28	8.740	8.864	-0.1234
29	9.303	9.262	-0.4095E-01
30	8.637	8.713	-0.7539E-01
31	9.133	9.074	-0.5966E-01
32	9.281	9.100	0.1814
33	8.311	8.653	-0.3423
34	8.558	8.850	-0.2922
35	8.112	8.238	-0.1268
36	7.642	8.221	-0.5789
37	7.905	7.545	0.3601
38	8.582	8.185	0.3963
39	8.170	7.789	0.3819
40	7.743	7.731	0.1263E-01
41	9.532	9.006	0.5260
42	9.279	8.723	0.5560
43	10.32	10.12	0.2015
44	10.26	9.552	0.7078
45	10.47	10.04	0.4313
46	9.605	9.619	-0.1323E-01
47	9.638	9.926	-0.2878
48	8.992	9.449	-0.4568
49	10.09	9.984	0.1080
50	8.112	8.114	-0.2830E-02

CORRELATIONS OF ESTIMATES

1	1.0000						
2	-0.5000	1.0000					
3	0.0653	-0.1500	1.0000				
4	-0.1196	0.1550	-0.1122	1.0000			
5	-0.5879	0.3216	-0.0825	0.4507	1.0000		
6	-0.6621	0.0844	-0.2595	-0.3873	-0.0331	1.0000	
7	-0.5526	0.1440	0.2212	0.1941	0.2582	0.1835	1.0000
8	-0.4358	0.2659	0.1879	0.1114	0.0757	0.1410	0.5973
	1	2	3	4	5	6	7
8	1.0000						
	8						

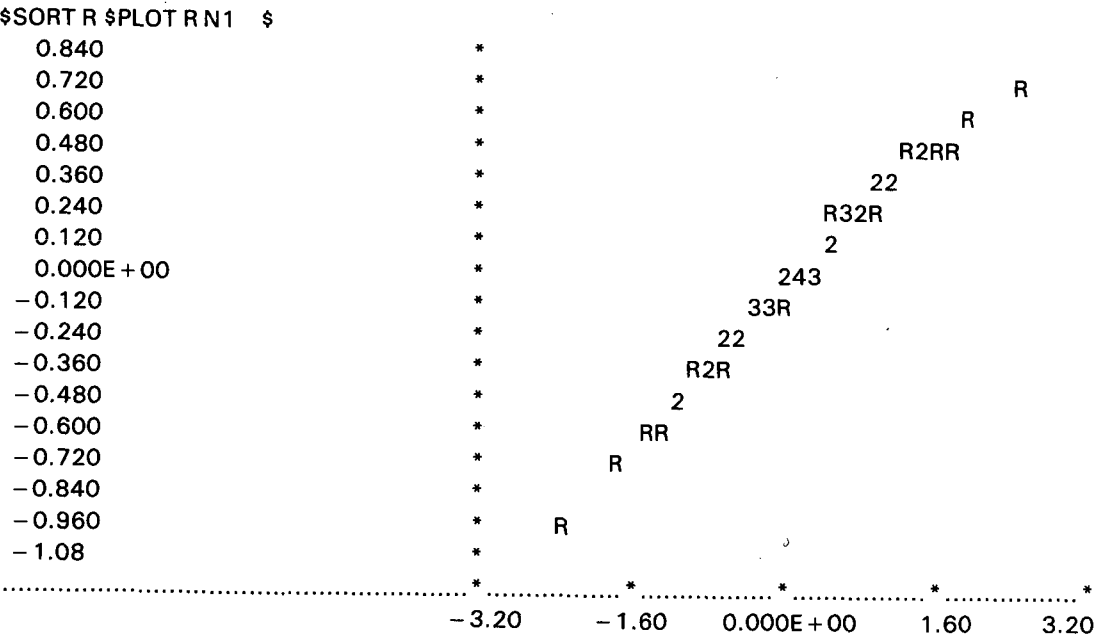
- \$C Pelos resultados obtidos neste ajustamento verifica-se que:
- (1) O modelo é aceito ao nível de 5%, tendo em vista que o desvio encontrado é inferior ao ponto crítico da distribuição qui-quadrado com 42 graus de liberdade, que é igual a 55,76.
 - (2) Todos os resíduos apresentam-se no intervalo $[-1; +1]$ e 88% entre -0.5 e $+0.5$, indicando condições favoráveis ao ajustamento. Apenas a observação 17 com resíduo -0.96 parece não estar bem explicada pelo modelo.
 - (3) A matriz das correlações não indica correlações fortes entre quaisquer estimativas.
- \$C

ANÁLISE GRÁFICA DOS RESÍDUOS

\$CAL $R = Z - \%FV : N1 = \%ND((\%GL(50,1) - .5)/50)$ \$

\$C

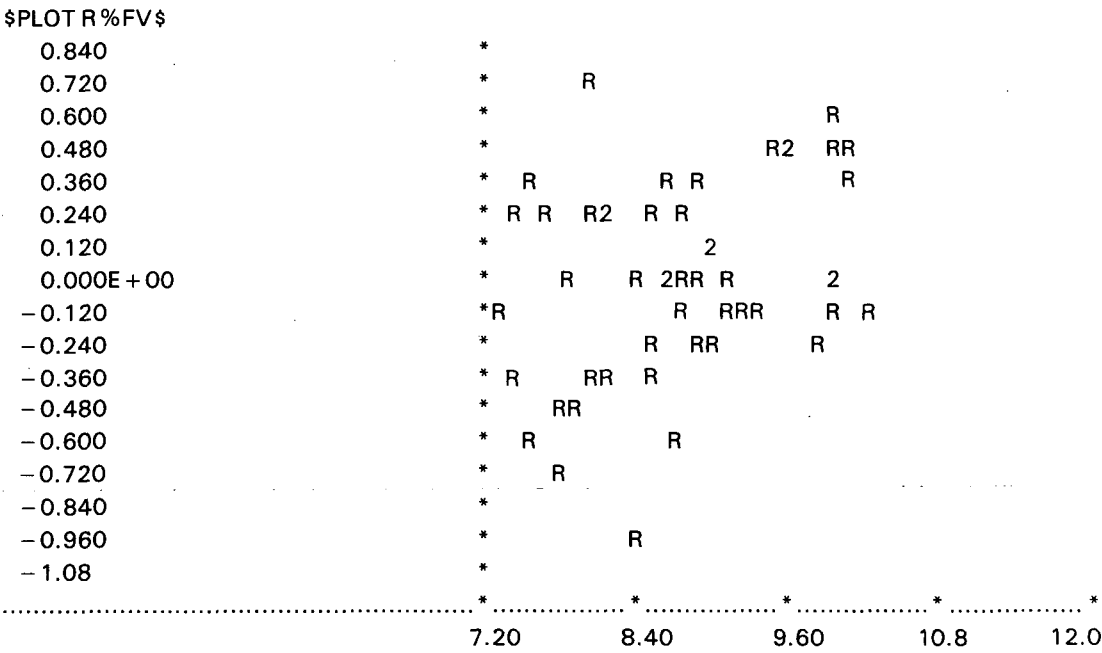
GRÁFICO NORMAL DOS RESÍDUOS



Como os pontos estão bastante próximos à bissetriz do primeiro quadrante, pode-se aceitar a distribuição $N(0,1)$ para os dados.

\$C

GRÁFICO DOS RESÍDUOS VERSUS VALORES AJUSTADOS



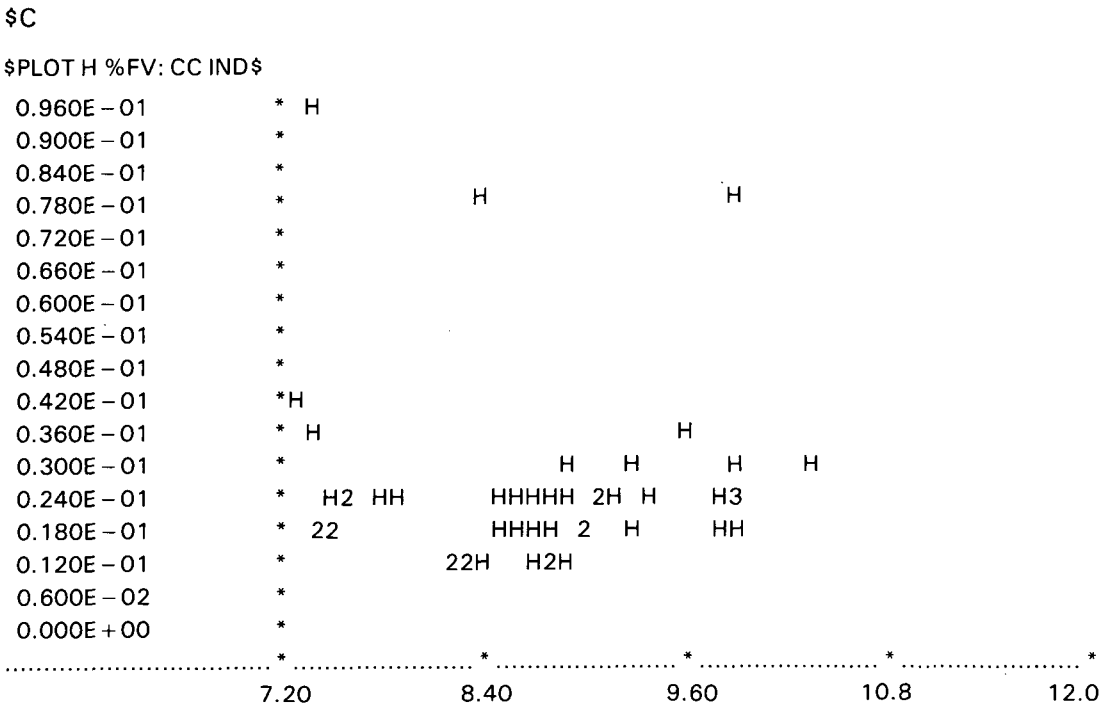
\$C

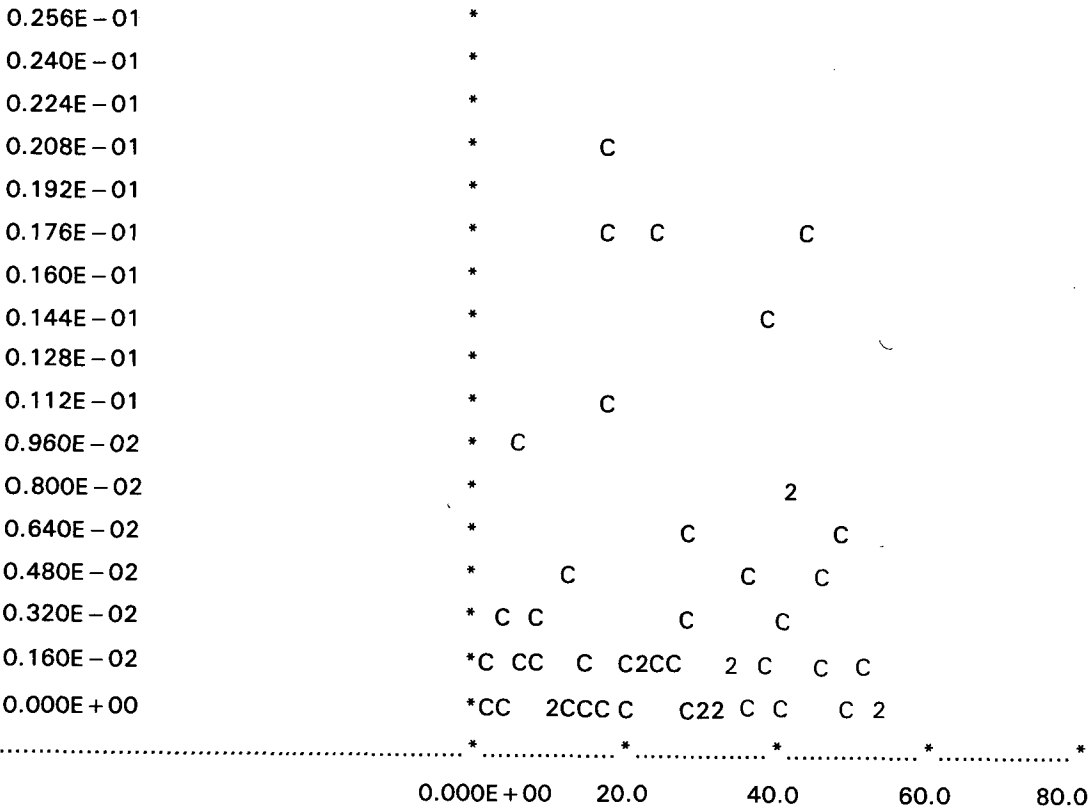
Como os pontos apresentam-se distribuídos aleatoriamente, pode-se aceitar a hipótese de independência e constância da variância dos dados.

\$EXTRACT %VL \$CAL VA = %GL(1,50):AA = VA:H = %VL*

% WT:TT = (Z - %FV)/(VA*(AA - H))*0.5\$

\$CAL CC = H*TT**2/(AA - H): IND = %GL(50,1) \$





§C

O gráfico dos elementos da diagonal de H versus %FV nota-se que nenhum elemento da matriz "HAT" é superior a P/N, isto é, 0.16. Logo, não há pontos influentes estatisticamente. A mesma conclusão pode, ser obtida do gráfico da estatística de COOK C versus o índice das observações, pois o máximo valor desta estatística é 0.02, sendo bem inferior a 1.

ANÁLISE 3

Uma análise alternativa utilizando para a testada e profundidade os modelos sugeridos pela NBR – 5676.

§CAL TN = %LOG(T):PN = %LOG(P) §
§FIT TN + PN + N + C + U + R1 + R2 + R3
§DIS ME §

CYCLE	DEVIANCE	DF
1	6.665	42

	ESTIMATE	S.E.	PARAMETER
1	5.839	0.7064	%GM
2	0.2362	0.1565	TN
3	-0.2901E-01	0.1211	PN
4	-0.2172	0.1521	N
5	0.5421E-01	0.4998E-02	C
6	0.1264	0.5266E-01	U
7	-0.4746	0.1571	R1
8	-0.2832	0.1729	R2
9	ZERO	ALIASED	R3
SCALE PARAMETER TAKEN AS 0.1587			

ANÁLISE 4

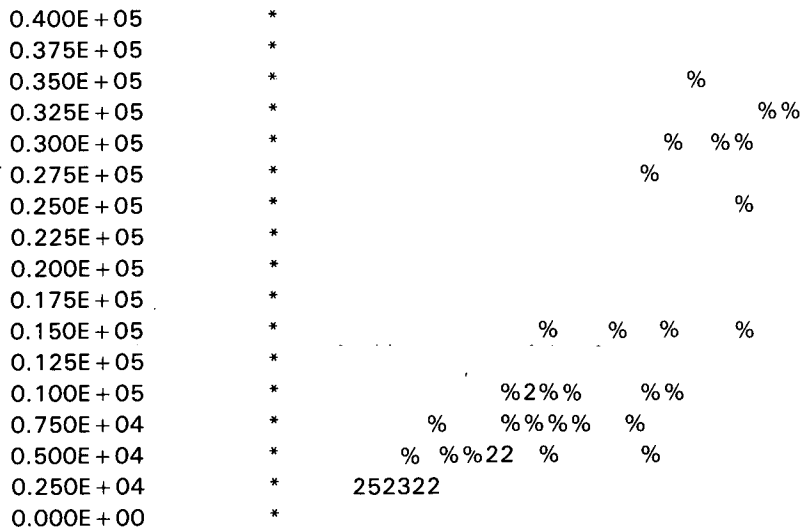
Y—VARIATE Z
ERROR NORMAL LINK IDENTITY

LINEAR PREDICTOR
%GM TN PN N C U R1 R2 R3

§C A análise com um modelo alternativo com erro gama

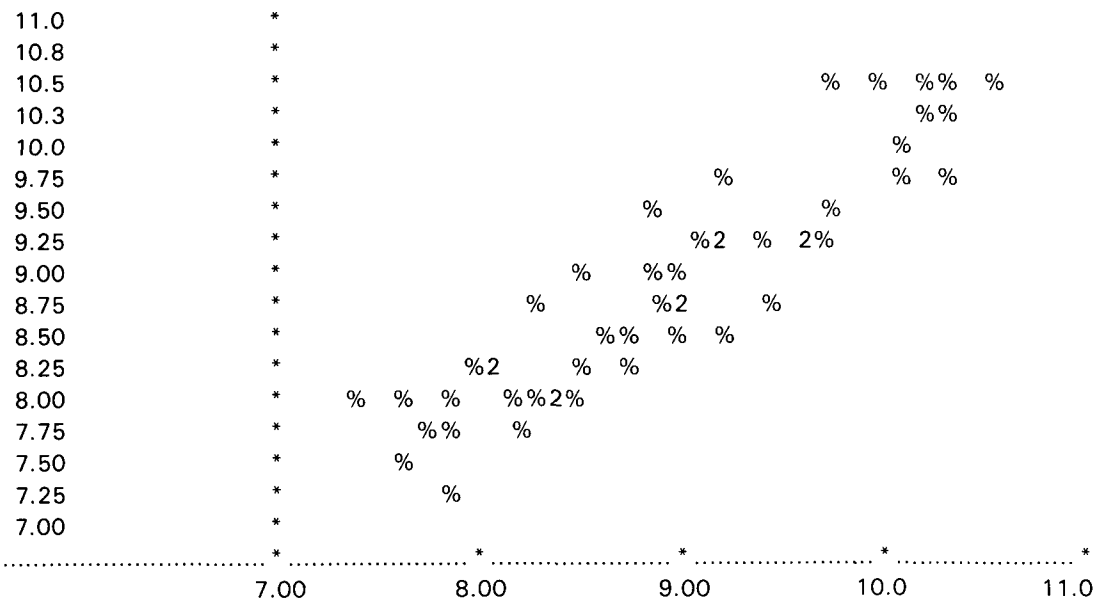
Escolha da função de ligação

CYCLE	DEVIANCE	DF
13	10.43	42

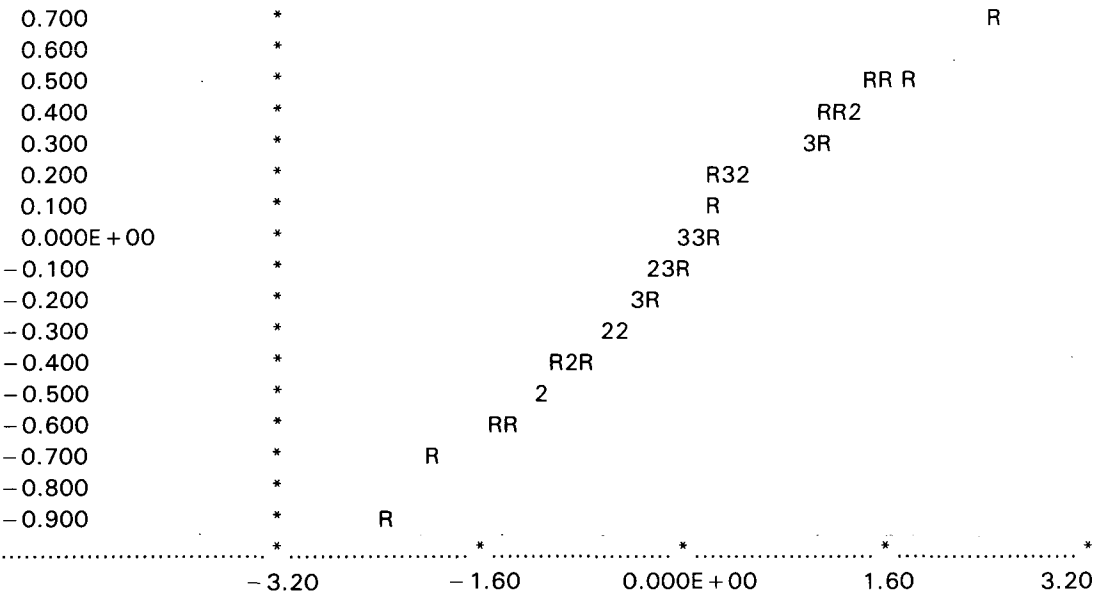


___ CURRENT DISPLAY INHIBITED

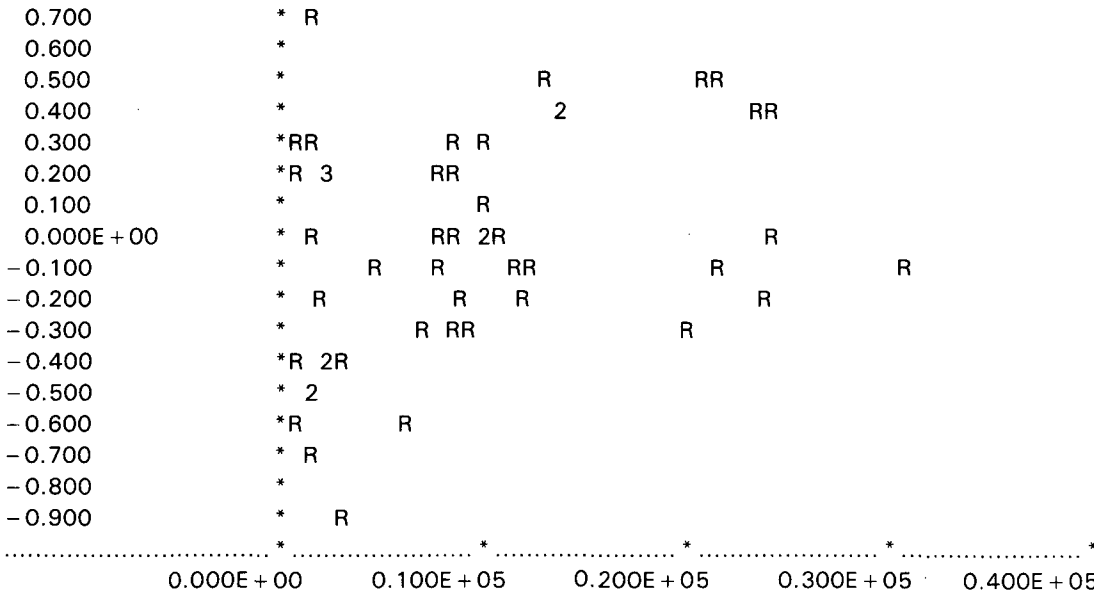
CYCLE	DEVIANCE	DF
4	6.109	42



\$SORT RA \$PLOT RA N1\$



\$PLOT RA %FV\$



\$C

\$EXTRACT %VL \$

\$CAL VA = %FV**2 \$

\$CAL AA = %GL(1,50):TT = (V - %FV)/(VA - (AA - H))**(1/2) \$

\$CAL CC = H*TT**2/(AA - H)

1ª) escrever o programa GLIM no sistema operador;

2ª) entrar no GLIM através do comando R GLIM; e

3ª) executar o programa usando:

- REINPUT [nº do programa]
- RETURN

Para a saída de todas as diretivas utilizadas, deve-se usar no início do programa GLIM, a diretiva \$ECHO .

Apresenta-se, a seguir, comentários resumidos sobre as diretivas utilizadas no processamento dos dados.

\$DATA 50 TPNCUR1 R2 R3 V

Rotula as variáveis cujos 50 valores serão lidos através da diretiva DINPUT.

\$UNITS 50

Define o número de dados igual a 50.

\$CAL B = V

Transforma o nome da variável V em B.

\$YVAR B

Define B como variável dependente.

\$FIT T + P + C + N + U + R1 + R2 + R3

Ajusta o modelo com as variáveis independentes T, P, C, N, U, R1, R2 e R3, fornecendo o número de iterações do processo de ajustamento, o desvio e o número de graus de liberdade.

\$CAL FV2 = %FV ★ % FV

Cria a variável FV2, onde %FV ("FIT VALUES") representa o vetor de valores ajustados.

\$FIT + FV2

Adiciona ao modelo anterior a covariável FV2.

\$DIS ME

Imprime resultados do último modelo ajustado, sendo pela opção M o modelo corrente e as estimativas dos parâmetros e erros-padrões pela opção E. São também bastante usadas as opções: R, que fornece os dados, os valores ajustados e os resíduos generalizados; e C, que imprime a matriz das correlações das estimativas dos parâmetros.

\$CAL %Q = 1 - 2 ★ 0,00005396 ★ %CU(V)/%NU

Cria um escalar %Q, onde %CU(V) e %NU são escalares do sistema que representam a soma acumulada das componentes de V e o tamanho da amostra, respectivamente.

\$PRINT %Q

Imprime o valor de Q.

\$MACRO BOXCOX

Causa o armazenamento das instruções, com o nome BOXCOX, para uso futuro, finalizando com a diretiva \$ENDMAC.

\$PLOT D1 L

Traça o gráfico do vetor D1 versus L.

\$CAL R = Z - %FV: N1 = %ND
((%GL(50,1) - 0,5)/50)

Calcula os resíduos de Pearson e em seguida os quantiles teóricos da distribuição normal, onde o símbolo : significa a repetição de \$CAL e % ND(X) representa a função inversa da integral da distribuição N(0,1) de $-\infty$ até X, %GL(K,N) gera valores de 1 até K em blocos de N.

\$SORT R

Ordena crescentemente o vetor R.

\$ERR G

Declara a distribuição gama para a variável resposta. Outras opções mais utilizadas são B(binomial), N(normal) e P(Poisson). No caso de omissão dessa diretiva o GLIM admite erro normal.

\$LINK I

Declara a função de ligação como a identidade. Outras opções mais utilizadas são: E(potência de expoente o número dado), L(log), R(recíproco) e S(raiz quadrada). No caso de omissão desta diretiva o GLIM admite a ligação canônica.

\$CYCLE 20

Limita a 20 o número máximo de iterações ao algoritmo de estimação. Fez-se ne-

cessária a utilização desta diretiva, tendo em vista que o algoritmo não convergiu nas 10 primeiras iterações (*display* natural do GLIM).

\$STOP

Termina a operação no GLIM e o controle retorna para o sistema operador.

BIBLIOGRAFIA

- ANSCOMBE, F. J. — Contribution to the Discussion of H. Hotelling's paper. *J. R. Statist. Soc., B.*, 15, 229-30, 1953.
- . The Examination and Analysis of Residuals. *Technometrics* 5, 141-60, 1963.
- ATKINSON, A. C. — Plots, Transformations and Regression. Oxford University Press, Oxford, 1986.
- BAKER, R. J.; NELDER, J. A. — The GLIM System, Release 3, Generalized Linear Interactive Modelling. Numerical Algorithms Group, Oxford, 1978.
- BOX, G. E. P.; COX, D. R. — An Analysis of Transformations. *J. R. Statist. Soc., B.*, 26, 211-52, 1964.
- COOK, R. D. — Detection of Influential Observations in Linear Regression. *Technometrics* 19, 15-18, 1977.
- CORDEIRO, G. M. — Improved Likelihood Ratio Statistics for Generalized Linear Models. *J. R. Statist. Soc., B.* 45 (3), 404-13, 1983.
- . Modelos Lineares Generalizados. Livro texto do VII SINAPE, Campinas, 1986.
- COX, D. R.; SNELL, R. J. — A General Definition of Residuals. *J. R. Statist. Soc. B.*, 30, 248-75, 1968.
- DANTAS, R. A. — Avaliação de Glebas Inseridas na Malha Urbana. (Tese de Mestrado) Departamento de Engenharia Civil — UFPE, 1987.
- DRAPER, N. R.; SMITH, H. — Applied Regression Analysis. 2nd. Edn., Wiley, New York, 1981.
- MALLOWS, C. L. — Some Comments on C_p . *Technometrics* 15, 661-75, 1973.
- NELDER, J. A.; WEDDERBURN, R. W. M. — Generalized Linear Models. *J. R. Statist. Soc. A.*, 135, 370-84, 1972.
- PREGIBON, D. — Goodness of Link Test for Generalized Modelling. *Appl. Statist.*, 29, 15-24, 1980.
- TUKEY, J. W. — On Degree of Freedom for Non-Additivity. *Biometrics*, 5, 232-42, 1949.
- WETHERILL, G. B. — Intermediate Statistical Methods. Chapman and Hall, London, 1981.
- WILKS, S. S. — The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.* 9, 60-62, 1938.

RESUMO

Desenvolve-se uma nova metodologia de pesquisa científica de avaliação de imóveis utilizando-se os modelos lineares generalizados. Faz-se uma análise de 50 lotes através do pacote GLIM incluindo o teste de Tukey e a transformação de Box e Cox.

ABSTRACT

A new scientific research methodology of appraisal of lands is developed by using the generalized linear models. An analysis of 50 lot values is made by the software GLIM incorporating the Tukey's test and the Box — Cox's transformation.

APLICAÇÃO DE ANÁLISE MULTIVARIADA A UM PROBLEMA DE DOSAGEM DE FÓSFORO PARA OVINOS

Lúcia Silva Kubrusly*
Dorinha M.S.S.V. Marcondes*
Cláudio T. Bornstein*

1 — FORMULAÇÃO DO PROBLEMA

O experimento consistiu em ministrar diversas dosagens de fósforo a um conjunto de 12 carneiros durante 13 semanas num total de 156 observações. Os carneiros foram divididos em três grupos de quatro animais recebendo, respectivamente, 0,2 e 5 g de fósforo por dia. O experimento foi realizado no Centro de Energia Nuclear na Agricultura — CENA em Piracicaba. Visando avaliar os resultados foram medidas, semanalmente, 15 variáveis que relacionamos na Tabela 1, junto com o coeficiente de variação (desvio padrão/média) das medidas colhidas.

O que ressalta, a primeira vista, é o valor elevado do coeficiente de variação da variável PUR. Com referência a PUR constatou-se que houve, efetivamente, falhas na coleta da amostra.

Na Tabela 2 reproduzimos a matriz de correlação. Com o objetivo de estabelecer relações entre as 15 variáveis, em questão, serão realizadas algumas análises de estrutura de dados cujos resultados serão apresentados nas Seções 2 e 3.

Inicialmente foi feita uma Análise de Variância — AV para verificar se o experimento estava ou não estabilizado, isto é, se o tempo era ou não um elemento significativo. Foi confirmada a hipótese de que não era significativo, o que permitiu a análise dos dados independentemente do tempo, considerando as 156 observações das 15 variáveis.

As técnicas usadas nesse trabalho para análise da estrutura dos dados são: a Análise das Componentes Principais — ACP e a Análise Fatorial — AF. Essas técnicas foram escolhidas porque permitem: primeiro, um exame das correlações entre as variáveis; segundo, a determinação das dimensões mais significativas para o conjunto de dados considerado; e, finalmente, a eliminação

* Os autores são, respectivamente, pesquisadora do LNCC — Laboratório Nacional de Computação Científica, Técnica do CENA/SP — Centro de Energia Nuclear na Agricultura e Professor da COPPE/UFRJ — Coordenação de Projetos de Pesquisas em Engenharia.

TABELA 1

VARIÁVEIS E COEFICIENTES DE VARIAÇÃO
PARA AVALIAÇÃO DOS RESULTADOS DE DOSAGEM DE FÓSFORO

Símbolo	Significado	Coefficientes de Variação
FENO	Ingestão de feno (g/dia)	0,28
PMICRO	Fósforo incorporado pelos microorganismos (mg)	0,54
PHRUM	pH do líquido do rúmen	0,05
CAP	Proporção Ca:P nas fezes	0,65
CAIN	Ca ingerido (g/dia)	0,16
CAOUT	% Ca fezes	0,22
POUT	% P fezes	0,51
PIN	Fósforo ingerido (g/dia)	0,60
PLAS	Fósforo no plasma (mg/100 ml)	0,21
MICRO	Microhematócrito (%)	0,15
HEMO	% hemoglobina	0,14
PRUM	Fósforo no rúmen (mg/100 ml)	0,42
PUR	Fósforo na urina (mg/dia)	2,09
PHUR	pH da urina	0,03
PESO	Peso dos animais (kg)	0,26

TABELA 2

MATRIZ DE CORRELAÇÕES

	FENO	PMICRO	PHRUM	CAP	CAIN	CAOUT	POUT	PIN	PLAS	MICRO	HEMO	PRUM	PUR	PHUR	PESO
FENO	1														
PMICRO	0,56	1													
PHRUM	0,37	0,40	1												
CAP	0,25	- 0,04	0,23	1											
CAIN	0,96	0,53	0,37	0,26	1										
CAOUT	0,42	0,11	0,03	0,18	0,39	1									
POUT	- 0,31	0,01	- 0,34	- 0,82	- 0,32	0,03	1								
PIN	- 0,08	0,35	- 0,09	- 0,71	- 0,09	- 0,05	0,77	1							
PLAS	- 0,17	0,07	- 0,16	- 0,48	- 0,20	0,04	0,55	0,53	1						
MICRO	0,62	0,30	0,15	0,16	0,57	0,40	0,19	- 0,15	0,06	1					
HEMO	0,27	- 0,01	- 0,21	0,04	0,22	0,30	0,01	- 0,15	0,08	0,70	1				
PRUM	- 0,26	0,10	- 0,32	- 0,64	- 0,28	- 0,03	0,79	0,71	0,62	- 0,08	0,04	1			
PUR	- 0,20	0,13	0,04	- 0,21	- 0,21	- 0,27	0,22	0,23	0,28	- 0,09	- 0,01	0,39	1		
PHUR	0,06	0,31	0,11	- 0,56	0,05	- 0,06	0,46	0,59	0,30	0,03	- 0,15	0,43	0,19	1	
PESO	0,76	0,69	0,63	0,36	0,77	0,33	- 0,41	- 0,01	- 0,16	0,49	0,02	- 0,30	- 0,11	0,12	1

das variáveis que após a análise são consideradas pouco significativas.

2 — ANÁLISE DE COMPONENTES PRINCIPAIS — ACP

A ACP parte de n variáveis originais $x_1, x_2, \dots, x_n$ procurando obter n componentes principais $y_1, y_2, \dots, y_n$ tais que os y_j sejam combinações lineares dos x_i , isto é:

$$y_j = \sum_{i=1}^n a_{ji} x_i \quad j = 1, \dots, n$$

tal que:

$$\text{var}(y_j) = \text{máxima}$$

$$\sum_{i=1}^n a_{ji}^2 = 1 \quad \text{Cor}(y_j, y_{j'}) = 0 \quad j \neq j'$$

Estas componentes são calculadas de modo que a primeira componente principal y_1 tenha variância máxima. Calculado y_1 determinamos y_2 procurando, novamente, maximizar a variância da segunda componente que não deverá ser correlacionada com y_1 . Também y_3 é obtido maximizando-se a variância, desde que não seja correlacionada com y_1 e y_2 . Assim, a variância de $y_1, y_2, \dots, y_n$ decresce e freqüentemente as $m \ll n$, componentes principais, contri-

buem com a parte mais significativa da variância total. Nesse caso, é possível desprezar as últimas componentes principais, reduzindo a dimensão do problema. Maiores detalhes sobre ACP poderão ser vistos em Herig (5), Kendall (6) e Morrison (8).

Na Tabela 3 consta o resultado da ACP com as quatro primeiras componentes correspondendo a mais de 77% da variância total. Os cálculos foram feitos utilizando-se o pacote Statistical Package of Social Science — SPSS (11).

Os coeficientes que aparecem na tabela representam a correlação entre a componente C_j e a variável x_i ($C_{ji} = a_{ji} \sqrt{\lambda_j}$, onde λ_j é a variância de C_j). Os que apresentam valor maior ou igual que 0,7 foram marcados com o sinal •, enquanto que aqueles que apresentam valor entre 0,5 e 0,7 foram marcados com o sinal o.

Observando os coeficientes da primeira componente maiores ou iguais a 0,5, em valor absoluto, podemos notar que existem dois grupos respectivamente dos coeficientes positivos e negativos:

FENO, CAP, CAIN, MICRO, PESO
POUT, PIN, PLAS, PRUM

Segundo Harman (4), o fator bipolar indica dois grupos de variáveis. Conforme veremos, mais adiante, eles correspondem aos

TABELA 3
ANÁLISE DE COMPONENTES PRINCIPAIS — ACP

	C1	C2	C3	C4
FENO	0,73 •	0,57 o	0,05	- 0,08
PMICRO	0,28	0,72 •	- 0,34	0,10
PHRUM	0,48	0,24	- 0,57 o	0,15
CAP	0,74 •	- 0,42	- 0,01	0,14
CAIN	0,73 •	0,53	0,02	- 0,11
CAOUT	0,36	0,32	0,46	- 0,35
POUT	- 0,80 •	0,42	0,16	- 0,19
PIN	- 0,61 o	0,65 o	- 0,16	- 0,17
PLAS	- 0,54 o	0,44	0,20	0,16
MICRO	0,52 o	0,48	0,53 o	0,24
HEMO	0,20	0,20	0,81 •	0,30
PRUM	- 0,74 •	0,47	0,15	0,08
PUR	- 0,36	0,17	- 0,19	0,78 •
PHUR	- 0,35	0,62 o	- 0,28	- 0,11
PESO	0,72 •	0,53 o	- 0,28	0,01
% variância total	33,7	23,4	12,9	7,3
Variância	5,05	3,51	1,93	1,09

conjuntos mais importantes das variáveis, mais precisamente o do cálcio e do fósforo. Eles fornecem as duas dimensões mais importantes do experimento.

3 — ANÁLISE FATORIAL — AF

A contrário da ACP, o modelo de Análise Fatorial — AF procura reproduzir, da melhor forma possível, a correlação entre as variáveis originais. Parte-se da hipótese de que cada uma das n variáveis originais $x_1, x_2, \dots, x_n$ é função de um certo número $m < n$ de fatores $f_1, f_2, \dots, f_m$ comuns a todas elas e de fatores específicos $u_1, u_2, \dots, u_n$ de modo que x_i pode ser reproduzido por x'_i , onde

$$x'_i = \sum_{j=1}^m a_{ij}f_j + d_i u_i \quad i = 1, \dots, n$$

Para a solução deste modelo são admitidas as seguintes hipóteses:

$$i) \sum_{j=1}^m a_{ij}^2 + d_i^2 = 1$$

ii) f_j, x'_i, u_i são variáveis padronizadas

$$iii) \text{Cor}(u_i, u_j) = 0 \quad \forall i \neq j'$$

$$\text{Cor}(f_i, u_j) = 0 \quad \forall i, j$$

$$\text{Cor}(f_i, f_j) = 0 \quad \forall j \neq j' \text{ (salvo nas rotações não ortogonais de uma solução inicial)}$$

A AF procura determinar os coeficientes a_{ij} e d_i , de modo que a matriz R' das correla-

ções das variáveis reproduzidas x'_i seja o mais semelhante possível com a matriz R das correlações das variáveis originais x_i . Assim, enquanto a ACP reproduz a variância total, a AF reproduz as correlações entre as variáveis. Por outro lado, a ACP gera fatores não correlacionados, o que não, obrigatoriamente, ocorre com a AF. Um maior aprofundamento da AF poderá ser visto em Harman (4), Kubrusly (7) e Rumel (10).

Na Tabela 4 constam os resultados da AF calculados com auxílio do SPSS. A determinação do valor de m , correspondente à dimensão do espaço comum, foi feita a partir da ACP, inicial que indica um número máximo de quatro fatores significativos correspondentes a uma variância maior ou igual a um. Optamos por manter, apenas, os três primeiros fatores porque não foi possível nenhuma interpretação para o quarto fator.

A solução apresentada foi obtida pelo método VARIMAX. Trata-se de uma transformação ortogonal de uma solução inicial da AF de tal forma que os fatores são não correlacionados e os coeficientes b_{ij} , que aparecem na Tabela 4, correspondem à correlação entre as variáveis e os respectivos fatores.

Novamente verificamos que é possível formar os dois grupos de variáveis:

FENO, PMICRO, PHRUM, CAIN, PESO

e

CAP, POUT, PIN, PLAS, PRUM, PHUR

TABELA 4
ANÁLISE DE FATORIAL — AF (VARIMAX)

	F1	F2	F3
FENO	- 0,10	0,76 ●	0,40
PMICRO	0,24	0,75 ●	0,06
PHRUM	- 0,21	0,63 ●	- 0,16
CAP	- 0,81 ●	0,17	0,09
CAIN	- 0,12	0,75 ●	0,34
CAOUT	0,00	0,19	0,40
POUT	0,91 ●	- 0,28	0,00
PIN	0,90 ●	0,13	- 0,15
PLAS	0,60 ●	- 0,09	0,15
MICRO	- 0,06	0,38	0,83 ●
HEMO	- 0,01	- 0,10	0,85 ●
PRUM	0,81 ●	- 0,17	0,08
PUR	0,25	0,01	- 0,02
PHUR	0,60 ●	0,27	0,11
PESO	- 0,16	0,92 ●	0,13

que correspondem, respectivamente, ao segundo e ao primeiro fator e que representam os conjuntos do cálcio e do fósforo, como veremos mais adiante.

4 — INTERPRETAÇÕES E CONCLUSÕES

Comparando os resultados da ACP e da AF, e selecionando as variáveis que apresentam coeficientes significativos nas duas análises, é possível identificar os seguintes conjuntos de variáveis:

FENO, CAIN, PESO
POUT, PIN, PLAS, PRUM
MICRO, HEMO

Caracterizemos o primeiro grupo como o do cálcio ingerido. O fato de FENO e CAIN pertencerem ao mesmo grupo, parece indicar que a ingestão de cálcio se deve, ao menos em parte, ao feno utilizado na alimentação do animal. A ingestão de feno refletiu-se, por sua vez, no peso dos carneiros, variável fundamental do ponto de vista econômico. A matriz de correlação da Tabela 2 já permitia prever que estas três variáveis estavam bastante correlacionadas.

O segundo grupo será caracterizado como o do fósforo ingerido. O fato de PIN e POUT pertencerem a um mesmo grupo (veja, também, correlação alta na Tabela 2) parece indicar que foi fornecida uma quantidade excessiva de fósforo que, não podendo ser absorvida, foi eliminada. A literatura mostra que existe correlação altamente positiva entre PIN e POUT (veja Cohen (2)). Efetivamente o principal caminho de excreção de fósforo em ruminantes é pelas fezes.

O fósforo é um elemento que estimula o crescimento dos microorganismos, aumentando a digestão de celulose (veja Anderson et alii (1) e Evans e Davis (3)). No entanto, o fato de tanto na ACP (veja C2 na Tabela 3) como na AF a variável PMICRO estar no grupo do cálcio, parece indicar que foram fornecidos aos animais níveis de fósforo acima daqueles capazes de estimular a atividade microbiana. Esta última foi, então, afetada por outros nutrientes, em parte, provindos do feno como, por exemplo, o cálcio.

O fato de PRUM pertencer ao grupo do fósforo ingerido parece óbvio e é confirmado pelas pesquisas feitas por Evans e Davis (3) Preston e Pfander (9). Preston e Pfander (9) observaram que o nível de fósforo do plasma de ovelhas estava diretamente relacionado ao teor desse elemento na dieta, confirmando também o fato de PLAS pertencer ao grupo do fósforo ingerido.

Um fato interessante ocorre com CAOUT. A Tabela 2 já indica que esta variável não é altamente correlacionada com nenhuma outra. Isto é confirmado pela ACP e pela AF pois CAOUT não pertence a nenhum grupo de variáveis, o que pode ser explicado se considerarmos que neste experimento não houve grande variação da quantidade de cálcio ingerido, pois esta variou muito mais em função da alimentação do que em função de outro fator qualquer. Como consequência houve assimilação da maior parte do cálcio ingerido.

Este fato permite uma conclusão interessante sobre a variável CAP que é diretamente proporcional a CAOUT e inversamente proporcional a POUT. Como CAOUT praticamente não é afetada pelo experimento, a variação do CAP dependeu fundamentalmente da variação do POUT. Efetivamente, a Tabela 2 mostra uma forte correlação negativa entre CAP e POUT/PIN, enquanto a correlação com CAOUT é pequena. Isto foi confirmado pela AF onde CAP pertence ao grupo do fósforo, apresentando um coeficiente negativo que indica a correlação negativa de CAP com as demais variáveis do grupo. Estes fatos contradizem, aparentemente, os resultados da ACP pois, lá, CAP pertence ao grupo do cálcio. Esta aparente contradição é resolvida se lembrarmos que a correlação negativa de CAP com as variáveis do seu grupo, implica, também, que CAP tenha o sinal oposto ao sinal das demais variáveis, dando, assim, erroneamente a impressão de pertencer ao grupo do cálcio na ACP.

Outra conclusão interessante da ACP e da AF permite confirmar algumas idéias iniciais acerca da variável PUR. Efetivamente, tanto na Tabela 3 como na Tabela 4, PUR não se encontra associada a nenhuma outra variável, o que parece confirmar, o fato de que neste caso houve falha na coleta de

amostras. Efetivamente, constatou-se que deveriam ter sido estabilizadas com ácido clorídrico. Como isto não foi feito houve precipitação de carbonatos, afetando os resultados. Na realidade o que ocorre é que quando há um excesso de fósforo, além da excreção fecal, há um aumento na excreção de P pela urina. Este aumento de P tende a elevar o pH da urina, pois o fósforo contribui para manter o lado básico. Isto talvez explique o fato de PHUR pertencer, fracamente, ao grupo do P na AF, embora não pertença a este grupo pelos resultados da ACP. Confirma também, a correlação positiva, embora não muito alta, existente entre PHUR e PIN na Tabela 2.

As variáveis MICRO e HEMO parecem ser variáveis secundárias, uma vez que há pouca evidência que permita associá-las às duas dimensões fundamentais do experimento. Tanto na AF como na ACP elas constituíram um grupo a parte, associado ao terceiro fator/componente. Na ACP, MICRO, no entanto, aparece também associada ao grupo do cálcio. Examinando a matriz de correlações verificamos que, efetivamente, existe uma correlação positiva entre MICRO e FENO/CAIN. MICRO mede a proporção entre células brancas e vermelhas. Ora, a ingestão de feno implica em ingestão de nutrientes que vão afetar o metabolismo do animal, afetando esta proporção. Este fato pode explicar os valores encontrados para a correlação entre MICRO e FENO.

A forte correlação entre MICRO e HEMO pode ser, talvez, melhor entendida se lembrarmos que a hemoglobina está presente nas células vermelhas. Provavelmente, foi a forte correlação positiva entre MICRO e HEMO — associada a uma baixa correlação entre HEMO e as variáveis do grupo do cálcio — que “puxou” MICRO para fora deste gru-

po, fazendo com que MICRO e HEMO constituíssem um grupo a parte.

Analisando a Tabela 4, verificamos que a variável PHRUM está associada ao segundo fator, representativo do grupo do cálcio, apesar de se apresentar com um coeficiente não muito elevado. Isto talvez possa ser explicado pelo fato de alimentos fibrosos como o feno, que são degradados lentamente, tenderem a aumentar o pH do rúmen.

PHRUM, na ACP, não foi associada ao grupo do cálcio fazendo, no entanto, parte da terceira componente junto com MICRO e HEMO. Estas três variáveis parecem fazer parte de um grupo cuja associação ao grupo do cálcio é fraca ou secundária.

Finalizando, podemos notar, através das conclusões e interpretações feitas, como a ACP e a AF permitem condensar e sistematizar a informação que se acha dispersa na matriz de correlação. A AF permitiu melhor separação das variáveis em grupos. A ACP, pelo fato de tentar maximizar a variância da primeira componente, juntou variáveis de diversos grupos prejudicando a discriminação.

É também importante que conclusões e interpretações, a partir da ACP e da AF, sejam feitas por uma equipe de pesquisadores, onde exista gente com um profundo conhecimento do problema específico que se está tratando. Foi isto que permitiu explicar a aparente contradição entre resultados da AF e da ACP no que diz respeito à variável CAP. Também a relevância de CAOUT, associada à quinta componente na ACP e a nenhum fator na AF, só pode ser reconhecida se lembrarmos que ela permite uma importante conclusão sobre a assimilação do cálcio. Uma aplicação cega de análise multivariada, talvez tendesse a eliminar esta variável.

BIBLIOGRAFIA

- 1 — ANDERSON, R.; et alii. _____. A laboratory technique for measuring phosphorus availability of feed supplement fed to ruminants. *J. Anim. Sci.*, 15 (2), 489-95, 1956.
- 2 — COHEN, R. D. H., _____. Phosphorus nutrition of beef cattle. The use of faecal and blood phosphorus for the estimation of phosphorus intake. *Aust. J. of Exp. Agric. and Animal Husb.*, 14 (71), 709-15, 1974.
- 3 — EVANS, J. L.; DAVIS, G. K. _____. Effects of phosphorus sulphur and molybdenum supplementation of a steer ration on cellulose digestion and on the rumen liquid concentrations of phosphorus and sulphur. *J. Anim., Sci.*, 20 (1), 200, 1961.

- 4 – HARMAN, H. H. _____. Modern factor Analysis, the University of Chicago Press, 1975.
- 5 – HERIG, L. C. T. _____. O modelo das componentes principais. Tese de M. Sc COPPE/UFRJ, 1979.
- 6 – KENDALL, M. _____. Multivariate analysis. Charles Griffin & Co. Ltd., 1975.
- 7 – KUBRUSLY, L. S. _____. O modelo de análise fatorial. Tese de M. Sc., COPPE/UFRJ, 1981.
- 8 – MORRISON, D. F. _____. Multivariate statistical methods. Mc Graw—Hill, 1981.
- 9 – PRESTON, R. L.; PFANDER, W. H. _____. Phosphorus metabolism in lambs fed varying phosphorus intake. *J. Nutr.* 83 (4), 369-78, 1964.
- 10 – RUMMEL, R. J. _____. Applied factor analysis. Northwestern University Press, 1970.
- 11 – SPSS — Statistical package of social science — v. 8, 1983.

RESUMO

Neste trabalho examinamos com auxílio de análise multivariada, mais especificamente análise fatorial e componentes principais, um experimento que consistiu em ministrar diversas dosagens de fósforo a um grupo de 12 carneiros. Os resultados deste experimento foram medidos através da observação de 15 variáveis. As técnicas utilizadas permitiram reconhecer os grupos de variáveis mais importantes, bem como as dimensões mais relevantes do experimento.

ABSTRACT

Using multivariate analysis, particularly, factor analysis and principal components we examine an experiment which consisted in studying the effect of phosphorus supplementation of sheep ration. The result of the experiment was measured examining 15 variables for a group of 12 sheep. The used techniques made it possible to recognize the most important variables and dimensions of the experiment.

PESQUISA AGROPECUÁRIA DO PARANÁ-SAFRA 1986/87*

(PROGRAMA DE APERFEIÇOAMENTO DAS ESTATÍSTICAS AGROPECUÁRIAS)

Charles Curt-Mueller **

Gil Silva ***

Álvaro Gonzales Villalobos ****

1 — INTRODUÇÃO

Num país de vocação agrícola como o Brasil, no qual se deseja que sua agricultura seja expandida e modernizada, é fundamental que a informação produzida, seja caracterizada pela agilidade, eficiência e confiabilidade, de modo a orientar adequadamente a política governamental para o setor.

A metodologia proposta no presente trabalho objetiva, basicamente, melhorar a produção das estatísticas agropecuárias brasileiras, através da operacionalização de um sistema de coleta de dados, baseado em métodos probabilísticos, no qual são combinados dados obtidos através de levantamento de campo, com dados obtidos através de sensoriamento remoto, via satélite de rastreamento de recursos natu-

rais. Em seu desenvolvimento participaram, conjuntamente, o IBGE e o INPE, sendo aplicada pela primeira vez, em escala maior, no ano de 1987, no Estado do Paraná.

O presente documento descreve os trabalhos executados, desde a etapa inicial da construção do painel, para seleção das amostras de áreas, até a obtenção dos resultados das estimativas.

2 — OBJETIVO

O objetivo principal da pesquisa é o de implementar um sistema operacional de informações agropecuárias, baseado em métodos probabilísticos, que permita ao IBGE produzir, periodicamente, estatísticas necessárias à avaliação da situação e ao planejamento da agricultura e da pecuária.

* Projeto financiado pelo IPEA

Recebido em 16 de junho de 1988

** Presidente da Fundação Instituto Brasileiro de Geografia e Estatística — IBGE e coordenador geral do projeto

*** Responsável pela coordenação técnica na Fundação Instituto Brasileiro de Geografia e Estatística — IBGE

**** Responsável pela coordenação técnica no Instituto de Pesquisas Espaciais — INPE

Os autores desejam expressar seus agradecimentos a todos que colaboraram para que fossem alcançados os objetivos desejados, principalmente aos diversos membros do IBGE e do INPE, atuantes nas equipes multidisciplinares do projeto, no ex-DEECA, atual DEAGRO/DPE, no DERN/DGC, no ex-DRN/BA, atual DRG/BA, na ex-DITRE/DF, na DEGE/PR e na DI, assim como aos técnicos do projeto SIAG, do Departamento de Sensoriamento Remoto, do INPE e, finalmente, aos avaliadores da Revista Brasileira de Estatística pelas sugestões e valiosos comentários apresentados.

As características essenciais do sistema são as seguintes:

- Desenho probabilístico estratificado e interpenetrante de uma amostra de áreas para servir de base a todos os procedimentos de estimação;

- Construção do painel de amostragem de áreas — formado por estratos do uso do solo, substratos geográficos, unidades primárias de amostragem, e segmentos — utilizando técnicas de interpretação de imagens de satélite (imagens do sensor TM/LANDSAT) e de aerofotografias;

- Seleção de uma amostra probabilística de segmentos, isto é, de áreas nas quais foi realizada a pesquisa no Estado do Paraná, para levantar dados sobre: características dos estabelecimentos agropecuários, área plantada, a ser plantada e colhida, rendimento médio esperado e obtido e datas de plantio e colheita das culturas, características da pecuária bovina, e informações sobre estocagem;

- Aplicação de estimadores diretos (estimador fechado e ponderado) para obter estimativas das variáveis da pesquisa, baseadas unicamente nos dados da amostra de áreas;

- Aplicação de estimadores de regressão, utilizando técnicas de análise digital de dados de satélite — processados com algoritmos estatísticos de *Clustering* — como informação suplementar aos dados da amostra de segmentos, para melhorar a previsão dos estimadores diretos da área das culturas; e

- Aplicação de estimadores de painéis múltiplos, em combinação com o modelo de amostragem, para melhorar a precisão do estimador ponderado para variáveis que apresentam um grande percentual da produção concentrado em um pequeno número de estabelecimentos agropecuários. É o caso, por exemplo, dos totais da pecuária.

amostragem, e a existência de condições básicas adequadas (mapeamento, cobertura total aerofotogramétrica em escala 1:25.000 correspondente a 1980, e uma ampla rede de assistência técnica), foi decidido iniciar a implementação do sistema de informações agropecuárias no Estado do Paraná. Foram as seguintes as etapas da pesquisa:

3.1 — Estudo-piloto numa área do estado

Escolheu-se como área do estudo piloto a região de Cascavel, situada no oeste do Paraná, quadrante 223.077, limitada pelas coordenadas 53°40' a 52° e 23°30'S a 25°50'S com uma superfície de 35.000 km² (185 X 185 km). Nessa área foram selecionados, usando-se o fotoíndice, 40 segmentos de aproximadamente 2 km, distribuídos nos estratos do uso da terra.

Os objetivos principais do estudo-piloto foram os seguintes:

- Verificar a operacionalidade dos procedimentos e instrumentos de coleta. Testar o questionário, a folha de listagem e o uso da fotografia aérea para delimitar os talhões de cada segmento e estimar as suas áreas;

- Verificar a operacionalidade do uso dos métodos do "segmento fechado" e "segmento ponderado", do ponto de vista dos informantes e entrevistadores, para todas as variáveis do questionário;

- Verificar a viabilidade da utilização do sistema gráfico INTERGRAPH para o registro dos limites dos estabelecimentos e dos talhões dos segmentos selecionados;

- Verificar os métodos de processamento de dados; e

- Verificar diferentes aspectos da metodologia para integrar dados de campo com dados digitais de satélite.

3.2 — Desenho da amostra

3.2.1 — Tamanho da amostra

Uma série de considerações, estatísticas referentes aos coeficientes de variação (CV) das variáveis agropecuárias de interesse, custo e prazo para a obtenção de resultados finais, e, em especial, a experiência

3 — METODOLOGIA UTILIZADA NA PESQUISA

Dada a importância da agricultura e da pecuária no Estado do Paraná, a experiência anterior em uma série de pesquisas por

nas amostras estaduais americanas do JES, indicaram que com uma amostra de 450 segmentos se poderia obter CVs de 5 a 15% para as principais variáveis da pesquisa. Portanto, até poder contar com dados de um primeiro levantamento, considerou-se como adequada uma amostra de $n=450$ segmentos para o Estado do Paraná.

3.2.2 — Estratificação da amostra

Para se contar com um sistema eficiente de pesquisas estatísticas agropecuárias, ou seja, para melhorar a precisão dos estimadores neste tipo de pesquisa, a experiência tem comprovado a conveniência de se empregar uma amostra estratificada segundo o uso atual da terra. Deste modo, tratou-se de dividir a área do estado em estratos relativamente homogêneos em relação ao uso da terra.

3.2.2.1 — Definição dos estratos utilizados no Estado do Paraná

Como passo prévio foi necessário contar com uma sistematização da informação estatística e cartográfica da área a ser coberta. Na definição dos estratos do uso da terra foram utilizadas, principalmente, as imagens de satélite LANDSAT/TM, canais 3 e 4, na escala 1:250.000, indicadas a seguir:

IMAGENS DE SATÉLITE UTILIZADAS

Órbita/ ponto	Data	Órbita/ ponto	Data
220/77	22/07/84	223/76	01/01/85
220/78	22/07/84	223/77	19/01/85
221/76	01/10/84	223/78	19/01/85
221/77	01/10/84	224/76	01/01/85
221/78	01/10/84	224/77	10/01/85
222/76	09/09/85	224/78	10/01/85
222/77	22/09/85		
222/78	22/09/85		

Estas cenas LANDSAT (correspondentes a áreas de 185 X 185 km) podem ser visualizadas no mapa-índice para LANDSAT 4 e no mapa de localização das imagens de satélite no Paraná (Figuras 1 e 2). Utilizaram-se imagens em preto e branco, já

que, na data de estratificação (maio 1986), dos 122 quadrantes coloridos — canais 3, 4 e 5 — processados, a metade foi rejeitada devido ao realce de cor pouco satisfatório obtido.

A escala escolhida (1:250.000) permitiu identificar áreas de até 5 km² sem perder a visão sinótica necessária para delimitar os estratos.

Na definição dos estratos foram utilizados também dados do Censo Agropecuário de 1980, mapas do uso do solo, e mapas da distribuição das culturas.

A definição dos estratos foi a seguinte:

ESTRATO	DEFINIÇÃO SEGUNDO O USO DA TERRA
I	Áreas com mais de 80% de área cultivada
II	Entre 50% e 79% de área cultivada
III	Entre 15% e 49% de área cultivada, exceto pastagens
IV	Entre 15% e 49% de área cultivada, e pastagens
V	Com menos de 15% de área cultivada
VI	Áreas não agrícolas (parques, reservas florestais, montanhas etc.)

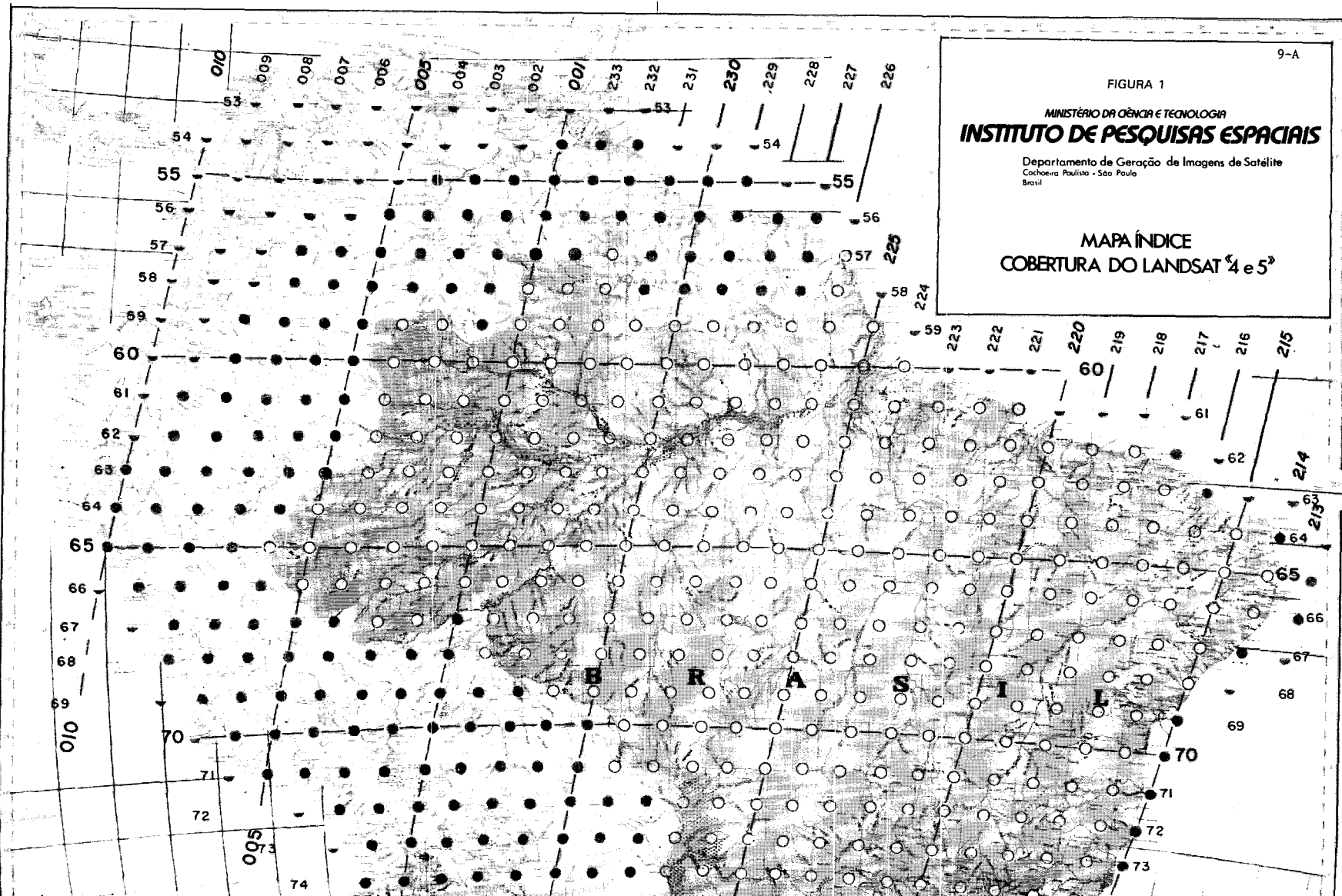
As áreas urbanas, que para a pesquisa, são aquelas com mais de 1 km², identificáveis nas imagens TM em escala 1:100.000, não foram incluídas nos estratos. Também não se incluíram os rios, reservatórios, canais e outras superfícies cobertas de água. Esses dois grupos completaram a área total do estado.

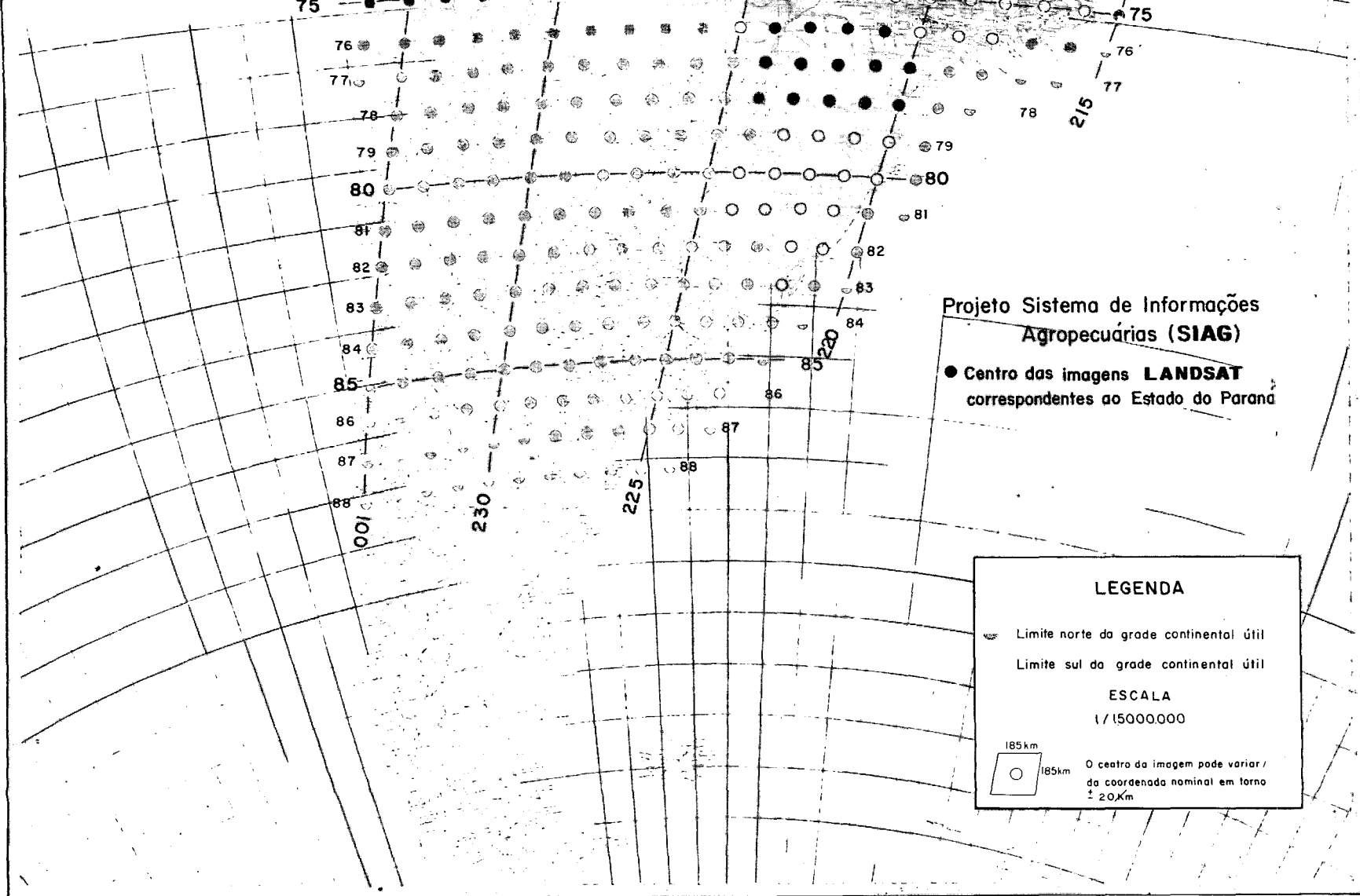
3.2.2.2 — Delimitação dos estratos

A delimitação original dos estratos, que se convencionou chamar estratos preliminares (EPs), foi efetuada em transparências sobre imagens LANDSAT/TM, canais 3 e 4, na escala 1:250.000, aplicando procedimentos de interpretação visual.

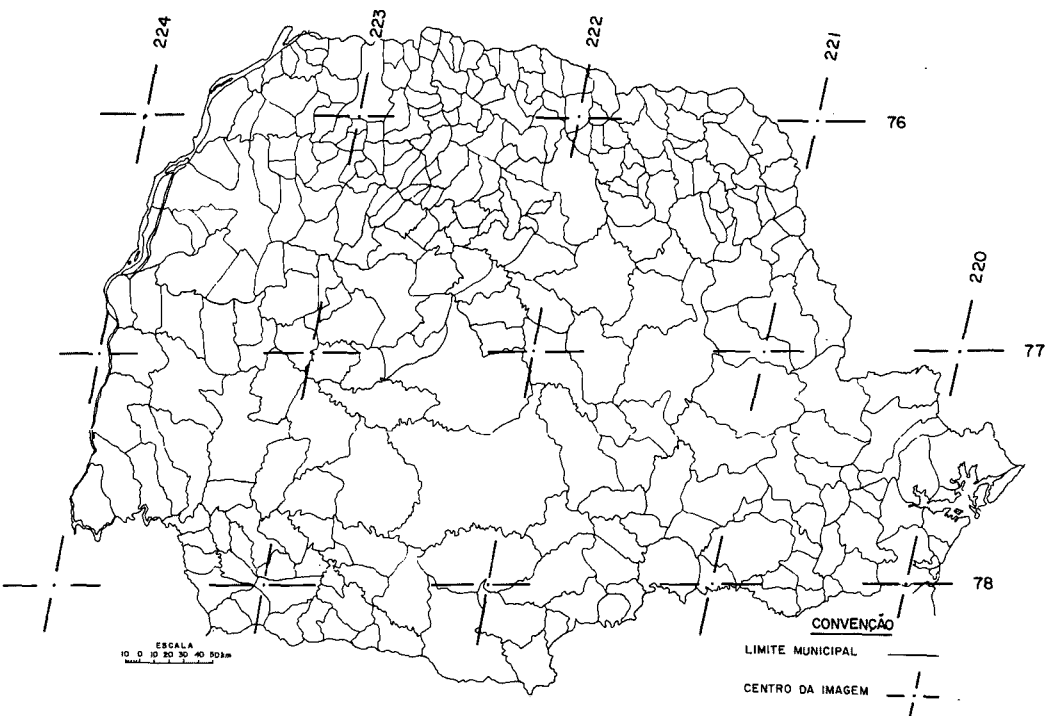
Na construção dos EPs não foram separadas áreas de menos de 5 km², e considerou-se, para definir cada EP, uma área mínima de 1% da área total do estado.

A informação contida nas 12 transparências geradas com os EPs foi digitalizada





-FIGURA 2-
DIVISÃO ADMINISTRATIVA E LOCALIZAÇÃO DAS IMAGENS DE SATÉLITE NO ESTADO DO PARANÁ



e arquivada no equipamento de computação INTERGRAPH, disponível no IBGE, em Salvador (Bahia). Também foi digitalizada e arquivada no INTERGRAPH uma base cartográfica da totalidade do estado. Usando as cartas topográficas (escala 1:250.000) do Projeto RADAM (ver Figura 3) incluíram-se os limites das cidades, rios, córregos, lagos, estradas e outros limites físicos permanentes. Finalmente, foram arquivados também os limites dos 311 municípios — subdivisões políticas — do estado. Em seguida, o INTERGRAPH gerou 93 *overlays* (transparências, na escala 1:100.000), que incluía a base cartográfica mencionada,

os limites municipais e os limites dos estratos preliminares, além de uma tabulação com o cálculo da área dos EPs.

Estes *overlays* foram utilizados, sobre mesas de luz, conjuntamente com imagens LANDSAT/TM, canais 3 e 4, na mesma escala 1:100.000, e com as cartas topográficas disponíveis — escalas 1:100.000 e 1:50.000, para ajustar os limites dos EPs a limites físicos permanentes (estradas, rios, córregos etc.), definindo portanto os estratos definitivos.

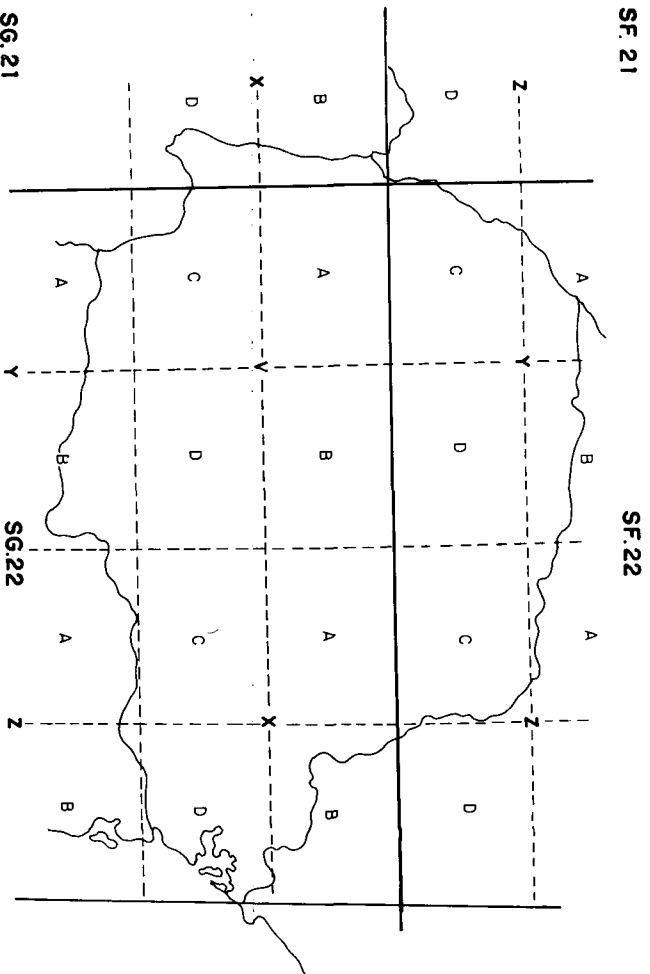
As imagens LANDSAT/TM (canais 3 e 4, escala 1:100.000) utilizadas foram as seguintes:

IMAGENS DE SATÉLITE UTILIZADAS

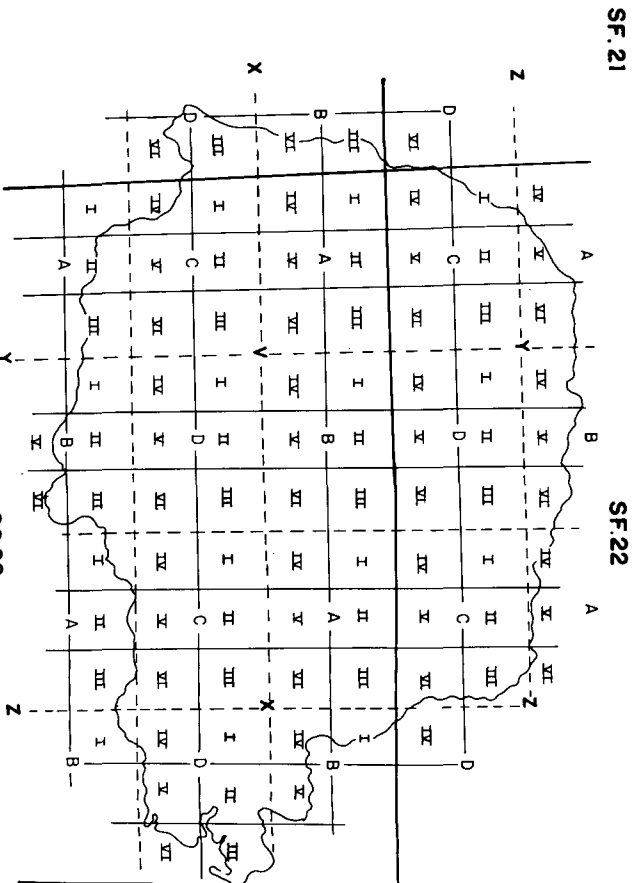
Órbita/ponto			Quadrante-data		
Órbita/ponto			Quadrante-data		
220/77	C,D	22/07/84	222/77	A,B,C,D	22/09/84
220/78	A,B,C,D	22/07/84	222/78	A,B,C,D	22/09/84
220/78	B	22/05/85	223/76	A,B,C,D	01/01/85
221/76	A,B,C,D	01/10/84	223/77	A,B,C,D	19/01/85
221/77	A,B,C,D	01/10/84	223/78	A,B,C,D	19/01/85
221/78	A,B,C	01/10/84	224/76	B, D	01/01/85
221/78	D	27/07/84	224/77	B, D	10/01/85
222/76	A,B,C,D	09/09/85	224/78	B	10/01/85

FIGURA 3
ESTADO DO PARANÁ

DIVISÃO DAS CARTAS NA ESCALA = 1 = 250.000



DIVISÃO DAS CARTAS NA ESCALA = 1 = 100.000



3.2.3 — Construção das unidades primárias de amostragem

O desenho estatístico da amostra subdividiu os estratos em Unidades Primárias de Amostragem — UPAs, e estas em segmentos. O que se pretendia era selecionar uma amostra probabilística de segmentos. As UPAs são unidades intermediárias estabelecidas com a finalidade de evitar particionar todo o estado em segmentos — o que seria muito caro e trabalhoso.

As UPAs são áreas contínuas, com limites físicos permanentes, claramente identificáveis, e que contêm um número aproximadamente igual de segmentos.

De posse da base cartográfica digitalizada, mencionada anteriormente, e utilizando-se imagens de satélite TM/LANDSAT, canais 3 e 4, correspondentes à totalidade do Estado do Paraná, foram delimitadas 6.926 UPAs, cobrindo todo o estado (Figura 4). Essa tarefa, bastante complexa, foi executada por aproximadamente 30 técnicos no período de 2 meses.

3.2.3.1 — Atribuição de medidas de tamanho às UPAs. Alocação da amostra

O desenho probabilístico da amostra requeria a atribuição de medidas de tamanho às UPAs. Estabeleceu-se como medida do tamanho de uma UPA o número inteiro mais próximo do quociente entre a sua área e a área do segmento-padrão correspondente ao estrato.

A alocação dos segmentos da amostra nos estratos foi determinada em função da medida de tamanho dos estratos, da sua definição de homogeneidade interna em termos do número de segmentos, com semelhante utilização do solo do tamanho dos segmentos, e do número e tipo de variáveis em consideração (Tabela 1).

3.2.3.2 — Subestratificação geográfica — amostras interpenetrantes

A subestratificação geográfica consistiu num segundo nível de estratificação, introduzido no desenho, de modo a aumentar a eficiência do modelo, utilizando a proximidade geográfica como variável adicional de estratificação.

Constituiu-se um painel para seleção de amostras interpenetrantes ou réplicas. Em outras palavras, a amostra de um determinado estrato do uso do solo foi constituída por réplicas, que são amostras aleatórias independentes com um segmento por subestrato. O número de réplicas por estrato é um múltiplo de 5, de modo a permitir a substituição anual de 20% dos segmentos da amostra. O desenho interpenetrante do painel facilita, em particular, a rotatividade da amostra, realizada para evitar erros decorrentes do cansaço do informante.

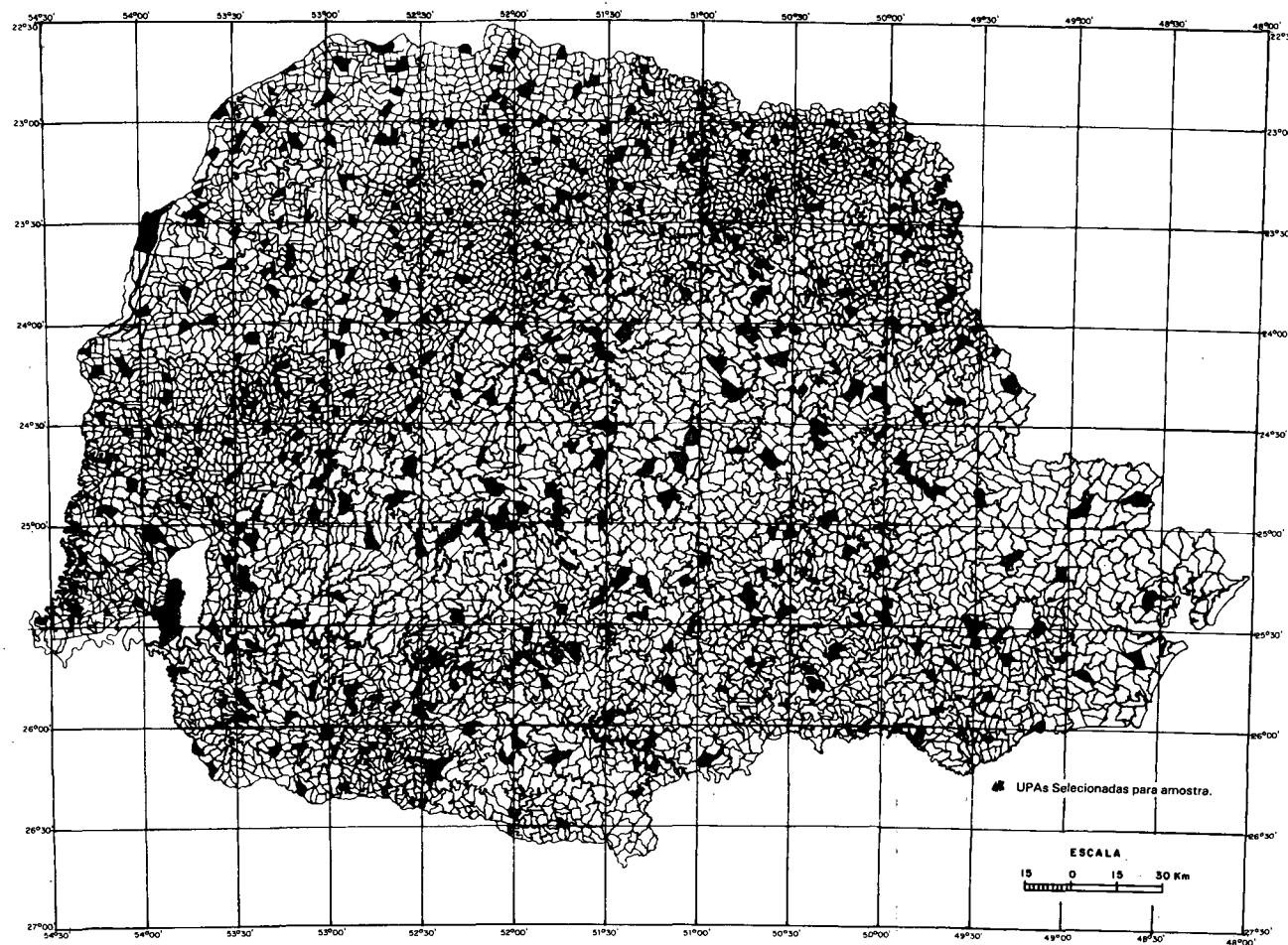
Dentro de cada estrato de uso do solo, os substratos geográficos foram formados por UPAs, ou partes de UPAs, com características agropecuárias semelhantes, que incluíram igual número de segmentos exceto no último substrato do estrato, que contou com um número ligeiramente diferente de segmentos, nos casos em que não foi possível construir todos os substratos de igual tamanho.

3.2.3.3 — Seleção das amostras interpenetrantes

Dentro de cada substrato geográfico foram selecionados sistematicamente r_n segmentos, com igual probabilidade e sem reposição (procedimento de seleção conhecido como EPSEM), formando assim as r_n réplicas do estrato. Usando as medidas de tamanho acumuladas (na ordem estabelecida) das UPAs, determinaram-se as UPAs selecionadas em cada substrato, e o número de segmentos da amostra em cada UPA. Cada UPA selecionada foi particionada no número de segmentos indicado pela sua medida de tamanho. Procedimento análogo foi utilizado para selecionar a amostra em todos os substratos de todos os estratos do uso do solo do estado. Observe-se que, a rigor, as UPAs constituem uma etapa transparente no processo de seleção. São utilizadas principalmente para evitar a tarefa exaustiva de construir todos os segmentos do estado; os segmentos foram delimitados unicamente nas UPAs que foram selecionadas.

Para o primeiro levantamento no Paraná foram selecionadas 5 amostras interpenetrantes, ou réplicas. Deste modo, sabendo-se que $L_n = n_n/5$ — número de substratos

FIGURA 4
MAPA DO ESTADO DO PARANÁ COM AS UNIDADES PRIMÁRIAS DE AMOSTRAGEM
(UPAs)



Mapa elaborado pela divisão de informática do IBGE-DRN/BA pelo Sistema Intergraph.

TABELA 1

ÁREA DOS ESTRATOS, NÚMERO DE SEGMENTOS NA POPULAÇÃO E NA AMOSTRA E INTERVALO DE SELEÇÃO — 1987

ESTRATO	NÚMERO DE SEGMENTOS NA POPULAÇÃO E NA AMOSTRA								
	Número de UPAs	Área (km ²)	Tamanho do segmento	Número de segmentos (N _h)	n _h	%	Número de subestratos (L _h)	Tamanho do subestrato (N _{hi})	Intervalo de seleção
TOTAL	6 926	194 767	—	99 540	450	—	90	—	—
1	1 774	29 814	1	29 808	135	30	27	1104/1104	220,8
2	1 158	19 062	1	19 120	85	18,9	17	1125/1120	225,0
3	1 890	52 501	2	26 246	120	26,7	24	1094/1084	218,8
4	801	23 915	2	11 916	55	12,2	11	1083/1086	216,6
5	548	29 963	4	7 513	35	7,8	7	1073/1075	214,6
6	755	39 512	8	4 937	20	4,4	4	1234/1235	246,8

geográficos do h-ésimo estrato do uso do solo —, os valores n_h decorrentes da alocação da amostra nos estratos foram ajustados para que fossem múltiplos de 5.

3.2.3.4 — Procedimento de estimação direta e cálculo de erros de amostragem

A partir dos dados de campo foram utilizados dois tipos de estimadores diretos para estimar o total do estado de uma determinada variável: o estimador fechado ($\hat{Y}_c$) e o estimador ponderado ($\hat{Y}_w$).

A escolha destes dois estimadores se deve ao tipo de unidade de informação correspondente a uma determinada variável, já que os limites dos segmentos não coincidem necessariamente com os dos estabelecimentos agropecuários.

A unidade de informação para o estimador fechado é uma parte ou subdivisão do segmento, correspondente a um produtor, ou mesmo a uma área não agropecuária do segmento, ou seja, a parte (ou totalidade) de um estabelecimento contida no segmento, ou uma área não agropecuária do segmento.

O estimador ponderado requer dados sobre a totalidade de cada estabelecimento incluído total ou parcialmente no segmento e, também, sobre a área contida no segmento em cada um desses estabelecimentos, que

é utilizada para ponderar os dados da totalidade do estabelecimento. Portanto, o estimador ponderado é aplicável no caso das variáveis nas quais o estabelecimento é a unidade de informação. Na pesquisa do Paraná utilizou-se o estimador ponderado para as estimativas do rebanho bovino e o fecho para as demais variáveis.

3.2.3.4.1 — Estimador fechado ($\hat{Y}_c$)

Considere-se a seguinte notação:

L — número de estratos do uso do solo no estado

h — subíndice que indica o estrato ($h = 1, \dots, L$)

i — subíndice indicador do subestrato ($i = 1, 2, \dots, L_h$)

j — subíndice indicador do segmento

k — subíndice indicador de um estabelecimento agropecuário

N — número de segmentos do estado

N_h — número de segmentos do h-ésimo estrato

N_{hi} — número de segmentos no i-ésimo subestrato do h-ésimo estrato

$$N = \sum_{h=1}^L N_h = \sum_{h=1}^L \sum_{i=1}^{L_h} N_{hi}$$

n — número de segmentos na amostra

n_h — número de segmentos da amostra no h-ésimo estrato

n_{hi} — número de segmentos da amostra no i -ésimo subestrato do h -ésimo estrato (réplicas).

$$n = \sum_{h=1}^L n_h = \sum_{h=1}^L \sum_{i=1}^{L_h} n_{hi}$$

$$L_h = \frac{n_h}{n_{hi}} = \frac{N_h}{N_{hi}}$$

$f_{hi} = \frac{n_{hi}}{N_{hi}}$ — fração de amostragem correspondente ao i -ésimo subestrato do h -ésimo estrato

m_{hij} — número de estabelecimentos (ou partes de) contidos no j -ésimo segmento, do i -ésimo subestrato, do h -ésimo estrato

y_{hijk} — valor da variável no k -ésimo estabelecimento, no j -ésimo segmento, do i -ésimo subestrato do h -ésimo estrato

$$t_{hij} = \begin{cases} y_{hij} & \text{se } m_{hij} > 0 \\ 0 & \text{se } m_{hij} = 0 \end{cases}$$

$$\hat{Y}_c = \sum_{h=1}^L \sum_{i=1}^{L_h} f_{hi}^{-1} \sum_{j=1}^{n_{hi}} t_{hij}$$

ou

$$\hat{Y}_c = \sum_{h=1}^L \sum_{i=1}^{L_h} N_{hi} \left(\frac{1}{n_{hi}} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{m_{hij}} Y_{hijk} \right)$$

3.2.3.4.2 — Estimador ponderado

Seja:

x_{hijk} — valor da variável correspondente ao total do k -ésimo estabelecimento do j -ésimo segmento, do i -ésimo subestrato, do h -ésimo estrato

P_{hijk} — peso do k -ésimo estabelecimento, do j -ésimo segmento, do i -ésimo subestrato, do h -ésimo estrato. É o quociente da área do k -ésimo estabelecimento, dentro do j -ésimo segmento, dividido pela sua área total

$$W_{hijk} = P_{hijk} \cdot x_{hijk}$$

$$t'_{hij} = \begin{cases} W_{hij} & \text{se } m_{hij} > 0 \\ 0 & \text{se } m_{hij} = 0 \end{cases}$$

$$\bar{t}_{hi} = \frac{1}{n_{hi}} t'_{hi} = \frac{1}{n_{hi}} \sum_{i=1}^{n_{hi}} t'_{hij}$$

$$\hat{Y}_w = \sum_{h=1}^L \sum_{i=1}^{L_h} N_{hi} \bar{t}_{hi} = \sum_{h=1}^L \sum_{i=1}^{L_h} f_{hi}^{-1} \sum_{j=1}^{n_{hi}} t'_{hij}$$

$$\hat{Y}_w = \sum_{h=1}^L \sum_{i=1}^{L_h} f_{hi}^{-1} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}$$

$$\hat{Y}_w = \sum_{h=1}^L \sum_{i=1}^{L_h} \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{m_{hij}} P_{hijk} \cdot X_{hijk}$$

3.2.3.4.3 — Variâncias dos estimadores

3.2.3.4.3.1 — Estimador fechado

$$\hat{Y}_c = \sum_{h=1}^L \hat{Y}_{c,h}$$

$$\hat{Y}_{c,h} = \sum_{i=1}^{L_h} N_{hi} \left(\frac{1}{n_{hi}} \sum_{j=1}^{n_{hi}} t_{hij} \right) = \sum_{i=1}^{L_h} N_{hi} \bar{t}_{hi}$$

$$V(\hat{Y}_{c,h}) = N_h^2 V(\hat{Y}_{c,hi})$$

$$= N_h^2 \left[\sum_{i=1}^{L_h} \left(\frac{N_{hi}}{N_h} \right)^2 \cdot \hat{s}_{hi}^2 \cdot \frac{1}{n_{hi}} (1 - f_{hi}) \right]$$

$$= N_h^2 \left[\sum_{i=1}^{L_h} \frac{1}{L_h^2} \cdot \hat{s}_{hi}^2 \cdot \frac{1}{n_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}} \right) \right]$$

$$= \frac{N_h^2}{L_h^2} \sum_{i=1}^{L_h} \hat{s}_{hi}^2 \cdot \left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}} \right)$$

onde

$$\hat{s}_{hi}^2 = \frac{1}{2(n_{hi} - 1)} \sum_{j=1}^{n_{hi}} (t_{hij} - t_{hij+1})^2$$

$$\hat{s}_{hi}^2 = \frac{1}{2(n_{hi} - 1)} \sum_{j=1}^{n_{hi}} (t_{hij} - t_{hij+1})^2$$

$$\hat{s}_{hi}^2 = \frac{1}{2(n_{hi} - 1)} \sum_{j=1}^{n_{hi}} (t_{hij} - t_{hij+1})^2$$

3.2.3.4.3.2 — Estimador ponderado

$$\hat{Y}_w = \sum_{h=1}^L \hat{Y}_{w,h}$$

$$\hat{Y}_{w,h} = \sum_{i=1}^{L_h} N_{hi} \bar{t}_{hi}$$

$$V(\hat{Y}_{w,h}) = \frac{N_h^2}{L_h^2} \sum_{i=1}^{L_h} \hat{s}_{hi}^2 \left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}} \right)$$

onde

$$\hat{s}_{hi}^2 = \frac{1}{2(n_{hi} - 1)} \sum_{j=1}^{n_{hi}} (t'_{hij} - t'_{hij+1})^2$$

$$V(\hat{Y}_w) = \sum_{h=1}^L \frac{N_h^2}{L_h^2} \sum_{i=1}^{L_h} \left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}} \right) \cdot$$

$$\cdot \frac{1}{2(n_{hi} - 1)} \sum_{j=1}^{n_{hi}} (t'_{hij} - t'_{hij+1})^2$$

3.2.3.4.4 — Técnicas de painéis múltiplos

Para as variáveis que apresentam uma distribuição em que um grande percentual da produção está concentrado num número relativamente pequeno de estabelecimentos agropecuários, é conveniente a utilização da técnica de painéis múltiplos, a fim de melhorar a precisão da estimação direta, a partir dos dados de uma amostra probabilística de áreas. No caso do sistema de informações agropecuárias, a técnica de painéis múltiplos foi utilizada para as variáveis da pecuária bovina. Com base no cadastro do Censo Agropecuário de 1985, foi elaborada uma listagem atualizada de um pequeno número de estabelecimentos correspondentes a uma elevada percentagem do total da variável em consideração, que se convencionou chamar de *listagem de estabelecimentos especiais*. O estimador de painéis múltiplos é um estimador que combina as estimativas de uma amostra — ou da totalidade — da listagem de estabelecimentos especiais, com as provenientes da amostra de áreas.

O estimador de painéis múltiplos é da forma:

$$\hat{Y}_p = \hat{Y}_e + \hat{Y}_{e,w}$$

Onde, para uma determinada variável, $\hat{Y}_e$ indica o estimador correspondente à listagem de estabelecimentos especiais, $\mathcal{L}$ e $\hat{Y}_{e,w}$ indica o estimador ponderado correspondente

aos estabelecimentos da amostra de áreas não contidos em $\mathcal{L}$. O primeiro passo utilizado na aplicação dos métodos de painéis múltiplos foi comparar todos os estabelecimentos correspondentes aos segmentos da amostra, com os de $\mathcal{L}$; sendo os não-pertencentes a $\mathcal{L}$, então, classificados no domínio de não-interseção, $\bar{\mathcal{L}}$. Admitindo que no conjunto $\mathcal{L}$ de estabelecimentos especiais seleciona-se uma amostra estratificada, então:

$$\hat{Y}_e = \sum_{r=1}^R N_r' n_r'^{-1} \sum_{u=1}^{n_r'} z_{ru}$$

onde R indica o número de estratos de $\mathcal{L}$; N_r' o número de estabelecimentos no r -ésimo estrato ($r = 1, \dots, R$) de $\mathcal{L}$; n_r' indica o número de estabelecimentos da amostra, no r -ésimo estrato; e z_{ru} é o valor da variável para o u -ésimo estabelecimento da amostra no r -ésimo estrato de $\mathcal{L}$.

O estimador $\hat{Y}_{e,w}$ é definido pela igualdade

$$\hat{Y}_{e,w} = \sum_{h=1}^L \sum_{i=1}^{L_h} \sum_{j=1}^{n_{hi}} \frac{N_{hi}}{n_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}^* - \sum_{h=1}^L \sum_{i=1}^{L_h} \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{n_{hi}} W_{hij}^* = \sum_{h=1}^L \sum_{i=1}^{L_h} N_{hi} \cdot \bar{W}_{hij}^*$$

onde

$$W_{hijk}^* = \begin{cases} W_{hijk} & \text{se o } k\text{-ésimo estabelecimento do } j\text{-ésimo segmento, do } i\text{-ésimo substrato, do } h\text{-ésimo estrato do uso do solo é classificado em } \mathcal{L}. \\ 0 & \text{se o estabelecimento não é classificado em } \mathcal{L} \end{cases}$$

Como o estimador $\hat{Y}_{e,w}$ utiliza o procedimento do segmento ponderado, o estimador da variância de $\hat{Y}_{e,w}$ é igual a

$$\hat{V}(\hat{Y}_{e,w}) = \sum_{h=1}^L \frac{N_h^2}{L_h^2} \sum_{i=1}^{L_h} \left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}} \right) \cdot \frac{1}{n_{hi} - 1} \sum_{j=1}^{n_{hi}} (W_{hij}^* - W_{hi}^*)^2$$

No cálculo do $\hat{V}(\hat{Y}_p)$, os estimadores $\hat{Y}_i$ e $\hat{Y}_{i,w}$ consideram-se não-independentes.

No Estado do Paraná, identificou-se um painel de 259 estabelecimentos especiais, com mais de 2.000 cabeças de bovinos segundo o Censo de 1985, incluindo 884.708 cabeças, ou seja, aproximadamente 40% do total do estado.

4 — A COLETA DA INFORMAÇÃO

4.1 — Treinamento e seleção do pessoal para os trabalhos de campo

Tendo em vista a complexidade das tarefas dos supervisores e entrevistadores nessa pesquisa e a elevada exigência de qualidade na metodologia que se adotou, foi dada muita atenção ao treinamento do pessoal de campo. Como eram necessários experiência em trabalhos de coleta de dados e um bom conhecimento das áreas cobertas empregaram-se, nos levantamentos, agentes de rede de coleta do IBGE no Paraná. Foram escolhidos, por indicação do Delegado do IBGE no Estado, 55 desses agentes, que participaram da fase de treinamento e depois foram submetidos a um teste de seleção. Desse teste resultou uma equipe de campo composta de 10 supervisores e 30 entrevistadores, à qual coube executar o levantamento dos 450 segmentos da amostra.

A fase de treinamento teve início em novembro de 1986, com o treinamento dos técnicos escolhidos para atuar como instrutores dos candidatos a supervisores e entrevistadores. O treinamento destes últimos iniciou-se com uma etapa teórica, entre 24 e 28/11/86, ao final da qual foram selecionados, mediante testes, 42 dos treinandos. Estes passaram à segunda etapa, desenvolvida entre 1º e 05/12/86.

A etapa prática do treinamento consistiu na ida dos treinandos a campo, em grupos de quatro, para realizar as tarefas para as quais estavam sendo preparados, sob a supervisão de instrutores. Nesta fase deu-se ênfase não só à aplicação de conhecimentos técnicos dos treinandos, como também ao desenvolvimento de habilidades e atitu-

des específicas para a tarefa a ser desempenhada.

Ao final do treinamento, foram escolhidos 10 supervisores do trabalho de campo, que receberam orientação específica sobre os trabalhos que lhes caberia executar. A cada supervisor coube chefiar uma equipe de três entrevistadores, na tarefa de realizar os levantamentos da amostra localizada em cada uma das 10 áreas em que, para fins do trabalho de campo, o Paraná foi dividido.

4.2 — Os trabalhos de campo

Para a fase de coleta de dados, criou-se no estado uma unidade de coordenação, composta de um coordenador, de supervisores e de entrevistadores.

O período de coleta de informação teve a duração de 35 dias, entre 19 de janeiro e 24 de fevereiro de 1987. A escolha da época do levantamento foi feita com base no calendário agrícola do Paraná. Era necessário que o levantamento fosse feito o mais cedo possível no ano agrícola — afinal de contas, a metodologia visa a fornecer subsídios para uma previsão de safras; entretanto, não se podia iniciar os trabalhos antes que, para as principais lavouras, o plantio tivesse sido concluído e as plantas tivessem nascido, sendo identificáveis em imagens de satélite.

A organização dos trabalhos de campo — as facilidades para o envio e remessa de materiais, para comunicações, transporte e, em geral, a logística necessária às tarefas do coordenador, supervisores e entrevistadores — teve como base a Delegacia do IBGE no Paraná, que conta com pessoal e instalações em todo o estado.

Os principais instrumentos utilizados nos trabalhos de campo da pesquisa foram:

- Fotografias aéreas em escala 1:10.000 e 1:25.000 com o segmento assinalado;
- Mapa rodoviário, ou semelhante, em escala adequada, com o segmento assinalado;
- Folha de Listagem (formulário que serve para relacionar todos os estabelecimentos contidos, totalmente, ou em parte, no segmento selecionado) (Anexo 1). O seu preenchimento permite identificar os estabelecimentos agropecuários aos quais foram aplicados questionários; e
- Questionário (formulário destinado à coleta de informações correspondentes a cada

um dos estabelecimentos relacionados na Folha de Listagem) (Anexo 2).

O questionário abrangeu as seguintes partes:

- Características gerais do estabelecimento;
- Área plantada, a ser plantada e colhida, rendimento médio esperado e obtido e datas de plantio e colheita de cada cultura;
- Características da pecuária bovina; e
- Dados de estoque de produtos armazenados na propriedade.

Na operação de coletar as informações referentes a um segmento da amostra, cabia à equipe de levantamento executar as seguintes tarefas:

- Identificar em campo a área do segmento. Em áreas remotas essa tarefa nem sempre era fácil, mas pôde ser executada a contento;
- Enumerar os estabelecimentos agropecuários contidos, no seu todo ou em parte, no segmento, preenchendo a *Folha de Listagem*;
- Assinalar no *overlay* que cobria a fotografia aérea, escala 1:10.000, do segmento, todos os estabelecimentos nele incluídos;
- Preencher um questionário para cada estabelecimento contido, no seu todo ou em parte, no segmento. O questionário era aplicado junto ao responsável pelo estabelecimento e cabia ao entrevistador obter informações, talhão a talhão, sobre as culturas neles contidas e sobre outros usos que estavam sendo feitos do solo. Completava, também, os demais quesitos do questionário;
- Assinalar no *overlay* sobre a fotografia aérea, escala 1:10.000, todos os talhões do segmento, de forma a poder identificar claramente o uso que estava sendo feito do solo em cada um deles; e
- Medir cada talhão a fim de controlar a qualidade da informação obtida dos entrevistadores. Para tal, mostrou-se útil o uso de uma grade de pontos, que era aplicada sobre o *overlay* da fotografia.

4.3 — A apuração dos resultados

O material coletado em campo foi transportado à sede do IBGE no Rio de Janeiro, onde sofreu o seguinte processamento:

- Digitação e crítica manual e automática;
- Ponderação direta dos resultados, com os estimadores correspondentes;
- Tabulação dos resultados ponderados e cálculo dos coeficientes de variação, da variância e de outras estatísticas;
- Emissão de tabelas finais;
- Preparo de arquivo de dados auxiliares a serem usados como informação complementar na etapa de interpretação de dados de imagens de satélite; e
- Remessa das fotografias dos segmentos e outros materiais ao INPE, em São José dos Campos (SP), para a fase de interpretação de dados digitais de satélite.

5 — RESULTADOS — ESTIMATIVAS PELA EXPANSÃO DOS DADOS DA AMOSTRA DA PESQUISA DO PARANÁ

Os principais resultados da Pesquisa Agropecuária do Paraná, de 1987, encontram-se nas Tabelas de 2 a 7, a seguir. Essas tabelas contêm não só as estimativas obtidas a partir da amostra da pesquisa e os respectivos coeficientes de variação, como também, para fins de comparação, dados correspondentes, obtidos por outras fontes (como o Levantamento Sistemático da Produção Agrícola — LSPA, e o Censo Agropecuário).

Antes de iniciar o exame dos resultados, convém explicar o significado dos coeficientes de variação associados aos valores estimados. Como se sabe, as estimativas da pesquisa foram calculadas com base nos dados de uma amostra; para sua interpretação correta, é necessário que se recorde que a amostra cobre uma parte pequena da população, não sendo razoável esperar que ela seja um retrato exato desta. Assim, só por puro acaso o valor estimado de uma variável será exatamente igual à sua magnitude na população. A partir das estimativas obtidas pela amostragem, entretanto, intervalos de confiança, contendo com uma certa probabilidade o verdadeiro valor da população, podem ser determinados. A amplitude desses intervalos pode ser calculada através dos coeficientes de variação (CV)

estimados, que nada mais são do que a raiz quadrada das variâncias relativas computadas da amostra.

Aceitando-se que as estimativas de interesse se distribuam segundo uma variável aleatória normal, pode-se estabelecer os limites de confiança, para cada um dos resultados, com um determinado grau de probabilidade. Exemplificando, suponha-se que uma certa estimativa T apresentou um coeficiente de variação de $V\%$. Isto significa que se pode afirmar que o valor da população estará compreendido no intervalo $T \pm V.T$ com 68% de probabilidade, ou, ainda, que o valor verdadeiro estará compreendido no intervalo $T \pm 1,96 \times V.T$ com 95% de probabilidade, e assim por diante.

Verifica-se, portanto, que, ao se aumentar o grau de confiança, aumenta-se a amplitude do intervalo e também que, quanto maior o coeficiente de variação, maior será o intervalo de confiança para uma dada probabilidade. Portanto, dependendo da magnitude do coeficiente de variação, o resultado estimado pode apresentar muito, pouco ou nenhum significado.

A seguir, apresenta-se uma análise dos resultados da Pesquisa Agropecuária do Paraná, tendo em vista, principalmente, avaliar sua eficiência.

5.1 — Estimativas para as lavouras de ciclo curto

A Tabela 2, a seguir, contém as estimativas das áreas plantadas com as principais lavouras de ciclo curto do Paraná, na safra 1986/87. Foram incluídas as lavouras de algodão, do arroz, da batata-inglesa (1.^a e 2.^a safras), do feijão (1.^a e 2.^a safras), do milho (safra principal e safrinha) e da soja. A tabela apresenta, também, as estimativas da área desses produtos, obtidas pelo LSPA, em junho de 1987, bem como a variação percentual entre os dados obtidos pela expansão da amostra do Paraná e os do LSPA.

Tiveram coeficientes de variação relativamente reduzidos as estimativas para o milho — safra principal (5,2%), a soja (6,2%), o feijão 1.^a safra (9,0%) e o arroz (9,5%). Em comparação com o LSPA de junho, a área estimada para o milho (2.809.364 ha) excede em apenas 5,7% à do LSPA (2.658.000 ha); a da soja é de 16,3% maior (1.998.576 ha contra 1.718.000 ha);

a do feijão 1.^a safra é de 3,0% inferior (675.411 ha contra 696.399 ha); e a do arroz é 26,7% menor (148.724 ha contra 202.923 ha).

Qualificando comparações do tipo das realizadas acima, é importante lembrar que as estimativas do LSPA no Paraná são consideradas de boa qualidade para um levantamento subjetivo. Para a sua obtenção, a rede do IBGE no estado conta com uma colaboração bastante ativa de instituições como a Secretaria de Estado da Agricultura — SEAG/PR, a Empresa de Assistência Técnica Rural — EMATER, a Companhia de Financiamento da Produção — CFP, a Organização de Cooperativas do Estado do Paraná, o Banco do Brasil, a Empresa Paranaense de Classificação de Produtos — CLASPAR, entre outras, quase todas com acompanhamentos cuidadosos e freqüentes da situação e das tendências da safra agrícola do Paraná. A despeito dessa colaboração, porém, as estimativas do LSPA referentes a produtos como o arroz e o feijão, por exemplo, cultivados que são, predominantemente, por pequenos agricultores, sem financiamento e sem assistência técnica oficial, tendem a ser mais precárias. Já as relativas às lavouras de algodão e de soja, por exemplo, são bastante acuradas, pois existem vários registros, acompanhamentos oficiais e formas de verificação que permitem estabelecer as áreas com elas plantadas com uma confiabilidade razoável. Para a lavoura do milho e da batata-inglesa, a confiabilidade dos dados do LSPA é maior que no caso do feijão, mas menor que no caso da soja e do algodão. De uma forma geral, porém, acredita-se que os dados da pesquisa subjetiva no Paraná dificilmente estão muito afastados da realidade; eles se constituem, portanto, em indicadores úteis para a avaliação dos resultados da primeira pesquisa com a nova metodologia, mas não se deve inferir que, existindo um afastamento substancial dos valores estimados pela expansão da amostra em relação aos do LSPA, é a pesquisa probabilística que está errada. Um juízo destes só pode ser feito com base no exame dos coeficientes de variação.

Portanto, podem ser consideradas bastante plausíveis as estimativas da pesquisa

TABELA 2

LAVOURAS DE CICLO CURTO, ESTIMATIVA DA ÁREA PLANTADA, COEFICIENTES DE VARIAÇÃO E ESTIMATIVAS DO LSPA — SAFRA DE VERÃO DE 1987

LAVOURA	ESTIMATIVA DA ÁREA PLANTADA			
	Área estimada (ha) (1)	Coeficiente de variação	Área estimada LSPA-junho (ha) (2)	Variação (%) (1/2)
Algodão	563 008	13,2	386 000	45,9
Arroz	148 724	9,5	202 923	- 26,7
Batata-inglesa (1ª safra)	17 069	47,3	29 000	- 41,1
Batata-inglesa (2ª safra)	18 916	48,6	21 000	9,9
Feijão (1ª safra)	675 411	9,0	696 399	- 3,0
Feijão (2ª safra)	54 717	29,6	31 000	76,5
Milho (safra principal)	2 809 364	5,2	2 658 000	5,7
Milho (safrinha)	165 764	17,9	230 000	- 27,9
Soja	1 998 576	6,2	1 718 000	16,3

referentes ao milho — safra normal, ao feijão 1ª safra e mesmo ao arroz (dada a precariedade do dado do LSPA para o produto). No caso da soja, como a estimativa do LSPA é bastante confiável, tomado isoladamente o valor obtido pela pesquisa, fica claro que ele superestima consideravelmente a área real cultivada com a leguminosa. Todavia, considerando que uma estimativa probabilística, como as da pesquisa do Paraná, representa o ponto central de um intervalo, cuja magnitude é determinada pelo coeficiente de variação, pode-se afirmar, com 95% de probabilidade, que a área efetivamente cultivada com a soja, na safra de 1986/87, se situa entre 1.756 mil e 2.241 mil hectares. Observe-se que o limite inferior desse intervalo é compatível com o dado de junho do LSPA referente à soja; incidentalmente, nos últimos meses o LSPA vem reajustando para cima sua avaliação da área cultivada com a leguminosa.

Se as estimativas da pesquisa referentes a lavouras amplamente cultivadas no espaço territorial do Paraná se revelaram de boa qualidade, o mesmo não pode ser dito em relação às culturas plantadas em áreas restritas do estado. Numa posição intermediária, temos os casos do algodão (cultivado quase exclusivamente acima do paralelo de 24°) e do milho-safrinha (plantado basicamente no sudeste do Paraná), com coefi-

cientes de variação de 13,2% e de 17,9%, respectivamente (Tabela 2); os casos extremos são os do feijão 2ª safra (plantado principalmente nas áreas do Paraná menos sujeitas a frios precoces) e o da batata-inglesa 1ª e 2ª safras (cultivadas em área bastante restrita do estado) com coeficientes de variação muito altos (29,6%, 47,3% e 48,6%, respectivamente). Está claro que para esses produtos a amostra da pesquisa não se mostrou adequada.

Para ilustrar o problema que surge quando um produto é cultivado em área muito restrita do espaço coberto pela amostra, consideremos o caso da batata-inglesa. O seu cultivo localizado fez com que, na amostra, somente 9 dos 450 segmentos (2% do total) apresentassem estabelecimentos com o cultivo de batata-inglesa 1ª safra; e que também apenas 9 (2% do total) tivessem produtores de batata-inglesa 2ª safra. Algo semelhante, mas de forma menos extrema, ocorreu nos casos do milho (safrinha) e do feijão 2ª safra. O problema está sendo estudado, à procura de formas de melhorar as estimativas referentes a esses produtos nos futuros levantamentos.

Apenas para registrar, a estimativa da pesquisa da área cultivada com a batata-inglesa 1ª safra é 41,1% menor que a do LSPA de junho; a da batata-inglesa 2ª safra é 9,9% maior; a do feijão 2ª safra é 76,5%.

superior e a do milho (safrinha) é 27,9% inferior. No caso do milho (safrinha) e no do feijão 2.^a safra, há o problema de que nem sempre o LSPA consegue discriminar precisamente o que é 1.^a e o que é 2.^a safra (ou o que é safra principal e o que é safrinha). Para fins de comparação, portanto, torna-se mais adequado considerar o total das duas safras. Procedendo-se desta forma, a diferença entre a estimativa da pesquisa e a do LSPA é de apenas 0,3% no caso do feijão (730.128 ha contra 727.399 ha) e de 3,0% no do milho (2.975.128 ha contra 2.888.000 ha).

Em suma, para as lavouras de ciclo curto, a metodologia adotada na pesquisa do Paraná se revelou bastante adequada. Ela permitiu estimar as áreas de um grupo bastante importante de lavouras, amplamente cultivado no estado, com um CV razoável. Quando definitivamente implantada e funcionando com a velocidade e a eficácia esperadas, ela fornecerá, cedo no período da safra, elementos importantes para a sua avaliação e acompanhamento. Juntamente com informações de outras origens, as estimativas permitirão determinar, com uma boa margem de segurança, as magnitudes das safras de soja, de milho, de feijão das águas e do arroz.

No caso de algumas lavouras cultivadas em áreas mais ou menos restritas, ainda existem problemas: os coeficientes de variação das estimativas são muito elevados, reduzindo o seu significado. Está-se enviando esforços para tentar contornar esses

problemas e melhorar as estimativas da área plantada com esses produtos, já no levantamento de 1988.

5.2 — Estimativas referentes à cana-de-açúcar, ao café e a mandioca

A Tabela 3 apresenta as estimativas das áreas cultivadas, com os respectivos coeficientes de variação de três lavouras de ciclo intermediário ou longo: a cana-de-açúcar, a mandioca e o café. Essas estimativas são, também, comparadas com dados do LSPA. Como se pode ver ali, a área plantada com a cana-de-açúcar foi calculada, a partir da amostra da pesquisa, em 160.846 ha, 7,6% inferior à do levantamento do LSPA; a área plantada com pés de café em idade produtiva da pesquisa totalizou 545.725 ha, superior em 20,8% aos 451.759 ha calculados pelo IBC em junho de 1987; e a área com mandioca, da pesquisa, foi de 79.056 ha, ou 9,1% inferior à do LSPA (87.000 ha).

A discrepância entre as estimativas da pesquisa e as do LSPA não é muito acentuada, especialmente no caso da cana e da mandioca. Entretanto, os coeficientes de variação (CV) obtidos para a cana e para a mandioca foram substanciais: 29,5% para a primeira, e 18,8% para a segunda. No caso do café, o CV se mostrou mais reduzido, embora ainda elevado (12,4%).

O curioso é que o produto com menor CV (o café) foi o que mais se afastou do dado

TABELA 3

ESTIMATIVA DA ÁREA PLANTADA, COEFICIENTES DE VARIAÇÃO E ESTIMATIVAS DO LSPA, SEGUNDO AS LAVOURAS DE CANA-DE-AÇÚCAR, CAFÉ E MANDIOCA — SAFRA DE 1987

LAVOURA	ESTIMATIVA DA ÁREA PLANTADA			
	Área estimada (ha) 1	Coeficiente de variação (%)	Estimativa do LSPA (1) 2	Variação (%) 1/2
Cana-de-açúcar	160 846	29,5	174 000	- 7,6
Café (2)	545 725	12,4	451 759	20,8
Mandioca	79 056	18,0	87 000	- 9,1

(1) Estimativas de junho para a cana-de-açúcar e a mandioca, e de junho de 1987 para o café. (2) Área plantada com pés de café em idade produtiva.

do LSPA e o com o maior CV se aproximou muito da informação do LSPA. Em ambos os casos, a qualidade das estimativas do LSPA é bastante boa; a lavoura do café é acompanhada pelo IBC com metodologia probabilística e existem várias formas de obter informações sobre a cana-de-açúcar. É só no caso da mandioca que a informação do LSPA não pode ser considerada muito confiável. De qualquer maneira, para os três produtos, os dados do LSPA se situam dentro do intervalo de confiança que se obtém das estimativas da pesquisa, se o nível de probabilidade for de 95%.

A principal razão para os altos coeficientes da variação nos casos da cana-de-açúcar e da mandioca parece estar, também, no seu cultivo mais ou menos localizado. No momento estão sendo estudadas formas de melhor adequar a amostra aos levantamentos desses produtos. Entretanto, dadas as disponibilidades de estimativas alternativas de boa qualidade para o café e para a cana-de-açúcar e a necessidade que existe de se dispor, no ano agrícola, de avaliações confiáveis sobre a magnitude das safras de ciclo curto — será dada prioridade à solução dos problemas relativos às estimativas da área plantada daqueles produtos.

5.3 — Estimativas referentes a outras utilizações da terra

A Tabela 4, a seguir, apresenta as estimativas da pesquisa, da área total dos estabelecimentos agropecuários do Paraná e dos usos não estritamente agrícolas de terra nos estabelecimentos — áreas em pastagens, em matas e florestas, em descanso e inaproveitáveis — bem como os respectivos coeficientes de variação. Como se pode ver ali, a área total em estabelecimentos agropecuários no Paraná foi estimada pela pesquisa, em 17.680.796 ha, com um CV de apenas 1,9%. A título de comparação, a área em estabelecimentos do Censo Agropecuário de 1985 é de 17.495.810 ha, muito próxima da obtida pela expansão da amostra.

No que diz respeito aos usos não estritamente agrícolas da terra nos estabelecimentos, a pesquisa obteve uma estimativa de 3.433.233 ha para a área em matas e flo-

restas (com um CV de 7,4%), de 6.760.562 ha para a área em pastagens (com um CV de 4,2%), de 1.050.975 ha para as terras em descanso (com um CV de 11,0%) e de 469.782 ha para as terras inaproveitáveis (com um CV de 10,1%).

Não existem dados de outras fontes para estabelecer uma comparação com essas estimativas. O Censo Agropecuário de 1985 ainda não apurou as áreas totais em matas e florestas, em pastagens, em descanso e inaproveitáveis, e tomar os dados de 1980 não seria apropriado, pois eles foram obtidos 6 (seis) anos antes da época da pesquisa do Paraná. Entretanto, a Tabela 4 fornece as proporções de cada forma de uso da terra em relação à área total em estabelecimentos, da pesquisa e do Censo de 1980. Observe-se que não há uma discrepância acentuada entre essas proporções, em cada caso. Com o avanço da pecuária no Paraná, é provável que a proporção da área total em pastagens tenha aumentado mais recentemente, aproximando-se do indicado pela pesquisa, e a metodologia da coleta de dados (com verificações em campo) nos dá certa segurança de que a proporção das terras inaproveitáveis da pesquisa se aproxima mais da realidade que a do Censo de 1980 (obtida com base na opinião do produtor).

5.4 — Estimativas dos efetivos de bovinos

A Tabela 5 apresenta as estimativas, e os respectivos coeficientes de variação, dos efetivos de bovinos do Paraná (total e por categorias principais) em janeiro de 1987. Como se pode ver ali, o número total de bovinos foi estimado em 9.649.534 animais, com um coeficiente de variação de 5,0%. Em outras palavras, pode-se afirmar que, com uma probabilidade de 95%, o efetivo de bovinos do Paraná se situava, no início de 1987, entre 8.704 mil e 10.595 mil animais.

A título de comparação, o número de bovinos levantados pelo Censo Agropecuário de 1985 totalizou 8.538.179 animais. Se adotarmos como hipótese um crescimento de 3% no rebanho bovino do estado em 1986 (ele pode ter sido maior, pois este foi um ano de forte retenção de gado pelos pro-

TABELA 4

ESTIMATIVA DA ÁREA TOTAL DOS ESTABELECIMENTOS AGROPECUÁRIOS, DA ÁREA EM MATAS E FLORESTAS, DA ÁREA EM PASTAGENS, DAS TERRAS EM DESCANSO E DAS TERRAS INAPROVEITÁVEIS NOS ESTABELECIMENTOS — 1987

ITENS	ESTIMATIVA DA ÁREA TOTAL				
	Áreas estimadas (ha)	Coeficiente de variação (%)	Área total Censo de 1985 (ha)	Proporção da área total estimada aos vários usos (%)	
				Estimada para pesquisa	Censo de 1980
Área total dos estabelecimentos	17 680 796	1,9	17 495 810	—	—
Área em matas e florestas	3 433 233	7,4	...	19,4	15,9
Área em pastagens ..	6 760 562	4,2	...	38,2	33,7
Terras em descanso	1 050 975	11,0	...	5,9	4,3
Terras inaproveitáveis	469 782	10,1	...	2,7	5,3

TABELA 5

ESTIMATIVA DO REBANHO BOVINO E COEFICIENTE DE VARIAÇÃO TOTAL E POR CATEGORIA DE ANIMAIS — 31-12-1986

CATEGORIA DE BOVINO	ESTIMATIVA DO REBANHO BOVINO				
	Número de cabeças	Coeficiente de variação (%)	Total de Censo de 1985	Participação relativa das categorias	
				Pesquisa	Censo de 1980
TOTAL	9 649 534	5,0	8 538 179	—	—
Menores de 2 anos ...	3 421 779	5,3	...	35,5	40,7
Maiores de 2 anos					
Vacas	3 775 394	5,9	...	39,1	34,8
Novilhas	845 829	9,0	...	8,8	7,8
Touros	215 114	4,8	...	2,2	2,2
Bois e garrotes	1 395 560	11,7	...	14,5	14,5

dutores), em 31/12/86 o número de animais chegaria a cerca de 8.794 mil cabeças, situando-se dentro do intervalo da estimativa.

A pesquisa estimou, também, que em 31/12/86 o rebanho bovino do Paraná era composto de 3.421.779 animais com menos de 2 anos de idade (com um CV de

5,3%), de 3.775.394 vacas (com um CV de 5,9%), de 845.829 novilhas (com um CV de 9,0%), de 215.114 touros (com um CV de 4,8%) e de 1.395.560 bois e garrotes (com um CV de 11,7%).¹

O Censo de 1985 não estabeleceu ainda a magnitude dos animais de cada categoria. Entretanto, a tabela nos permite comparar a

¹ A soma dos animais das várias categorias excede um pouco a estimativa do total de bovinos. Isto é natural, pois o montante de cada categoria é estimado separadamente. Só por coincidência essa soma seria igual ao total estimado pela pesquisa.

participação percentual de cada grupo no total estimado pela pesquisa, com a mesma participação obtida a partir dos dados do Censo de 1980. Como se pode ver, não há grande discrepância entre tais percentuais; ela é maior nos casos dos animais com menos de dois anos (35,5% para a pesquisa e 40,7% para o Censo de 1980) e das vacas (39,1% para a pesquisa e 34,8% para o Censo de 1980). Recorde-se, entretanto, que esses percentuais tendem a se modificar no tempo, dependendo da situação e da evolução dos mercados da carne e do leite.

5.5 — Estimativa da produção armazenada em estabelecimentos agropecuários

Com o objetivo de avaliar a conveniência de empregar na pesquisa de estoques de produtos agrícolas, em desenvolvimento pelo IBGE, dados das quantidades armazenadas em estabelecimentos agropecuários, a pesquisa do Paraná incluiu quesitos sobre o assunto. A Tabela 6, a seguir, apresenta as estimativas obtidas mediante a expansão da amostra, referentes às quantidades armazenadas em 31/12/86, dos principais produtos agrícolas, bem como, na mesma data, da capacidade total e da capacidade útil de armazenagem nos estabelecimentos agropecuários do estado; apresenta, também, os respectivos coeficientes de variação.

Como se pode ver ali, não foram encontrados estoques de algodão em estabelecimentos agropecuários. Para o arroz em casca, a estimativa foi de um estoque de 16.227,5 toneladas da safra anterior, e 2.390,6 toneladas da safra 1986/87. Para o café em coco, a estimativa foi um estoque de 63.526,7 toneladas de safras anteriores e de 4.720,2 toneladas da safra 1986/87. No caso do feijão preto, os estoques estimados totalizaram 3.396,0 toneladas de safras anteriores e 4.975,4 toneladas da safra 1986/87; no do feijão em cores, os totais foram de 3.982,9 e 22.763,0 toneladas, respectivamente. Para o milho, os estoques estimados atingiram 168.968,8 toneladas (safras anteriores) e 3.949,7 mil toneladas (safra 1986/87). Os estoques de soja somaram apenas 331,2 toneladas (de

safras anteriores) e 91,2 (da safra atual); os de trigo totalizaram 326,3 toneladas (evidentemente, da safra anterior).

Infelizmente, não existem dados que permitam avaliar tais resultados. Eles são compatíveis com as quantidades produzidas no estado, dos produtos cobertos na pesquisa, mas os coeficientes de variação elevados das estimativas indicam que existem problemas. Os menores CVs obtidos — para o milho, safras anteriores (15,0%) e para o arroz, safras anteriores (20,7%) — já são substanciais; os maiores — para a soja (nos dois casos) e para o trigo — ultrapassam o nível de 100%. Está claro que a amostra da pesquisa apresenta problemas para a estimativa de quantidades estocadas de produtos agrícolas. Está-se estudando a questão no sentido de adaptar a amostra aos objetivos da pesquisa de armazenagem.

As estimativas da capacidade total e da capacidade útil de armazenagem (Tabela 6) foram de 22.601.355 m³ e de 18.153.455 m³, respectivamente, com um CV de 20,8% nos dois casos. Segundo o Censo Agropecuário de 1980, a capacidade de armazenagem de grãos do Paraná em 31-12-80 era de 19.570.613 m³. Assim, parece razoável a estimativa de capacidade obtida pela pesquisa.

5.6 — As estatísticas de rendimento médio da pesquisa

A Tabela 7, a seguir, apresenta um conjunto de estatísticas de rendimento médio, calculadas com base nos dados da amostra, bem como os dados correspondentes de rendimento médio do LSPA, referentes aos principais produtos investigados na pesquisa.

Antes de discutir as informações da Tabela 7, cumpre explicitar a natureza das estimativas de rendimento médio da pesquisa. Em primeiro lugar, elas traduzem avaliações dos produtores e não o resultado de medições objetivas. Depois, para muitos produtos elas refletem as expectativas destes. Na ocasião do levantamento de campo, a colheita já estava bastante adiantada nos casos do feijão 1ª safra e da batata-inglesa 1ª safra, e o produtor tinha base concreta para

TABELA 6

ESTIMATIVAS E COEFICIENTES DE VARIAÇÃO DA PRODUÇÃO ARMAZENADA NOS ESTABELECIMENTOS AGROPECUÁRIOS 1986-87

PRODUTO	PRODUÇÃO ARMAZENADA			
	Estimativas		Coeficientes de variação (%)	
	Safras anteriores (t)	Safra 86/87 (t)	Safras anteriores	Safra 86/87
Algodão (em pluma).....	—	—	—	—
Algodão (em caroço).....	—	—	—	—
Arroz (em casca).....	16 227,5	2 390,6	20,7	43,0
Arroz (beneficiado).....	154,3	—	56,9	—
Café (em coco).....	63 526,7	4 720,2	38,8	84,3
Café (em grão).....	808,9	67,5	67,6	111,5
Feijão (em grão/preto).....	3 396,0	4 975,4	44,1	27,7
Feijão (em grão/cores).....	3 982,9	22 763,0	28,1	30,9
Milho (em grão).....	168 968,8	3 949,7	15,0	39,3
Soja (em grão).....	331,2	91,2	111,6	111,5
Trigo (em grão).....	236,3	—	105,5	—

OUTRAS ESTIMATIVAS	PRODUÇÃO ARMAZENADA	
	Volume (m ³)	Coeficiente de variação (%)
Capacidade total de armazenamento.....	22 601 355	20,8
Capacidade útil de armazenamento.....	18 153 455	20,8

avaliar rendimentos; para o arroz, o milho e a soja, por exemplo, na maioria dos casos, a colheita ainda iria se realizar, mas a evolução das lavouras dava aos agricultores elementos de avaliação; entretanto, para o feijão 2.ª safra e a batata-inglesa 2.ª safra, o plantio ainda estava no seu início e as avaliações de rendimento médio dos agricultores não passavam de palpite.

Trata-se, portanto, de avaliações subjetivas; elas têm uma vantagem em relação às estimativas do LSPA (também subjetivas), pois refletem a opinião de quem mais próximo está das lavouras, o produtor.

Em seguida avaliam-se, para os principais produtos, os resultados da pesquisa em comparação com os do LSPA.

5.6.1 — Algodão

As estatísticas de rendimento médio da pesquisa foram obtidas com base nas observações de 410 talhões. A menor expecta-

tativa de rendimento médio encontrada foi de 930 kg/ha e a maior, de 2.851 kg/ha, o que representa uma considerável dispersão. Entretanto, a média, a mediana e a moda dos rendimentos são quase coincidentes (1.816 kg/ha, 1.860 kg/ha e 1.860 kg/ha, respectivamente), indicando que a média se constitui em estatística adequada para refletir as expectativas dos produtores do Paraná, da quantidade produzida por hectare de algodão, em janeiro/fevereiro de 1986.

Comparando a média calculada a partir dos dados da pesquisa com a estimativa de rendimento do LSPA (abril de 1987) para o algodão, observe-se que esta é 8,2% inferior. Essa diferença deve ser avaliada com cautela: ela pode indicar que, em abril, o LSPA subestimava a produtividade do algodão, ou refletir um otimismo exagerado da parte dos agricultores do Paraná no início de 1987. Convém ter em mente o fato de que o rendimento da lavoura do algodão é

TABELA 7

ESTATÍSTICAS SOBRE O RENDIMENTO MÉDIO (kg/ha) DOS PRODUTOS DA PESQUISA E O RENDIMENTO MÉDIO (kg/ha) UTILIZADO PELO LSPA – ABRIL/1987

Produto	RENDIMENTO MÉDIO DOS PRODUTOS DA PESQUISA E DO LSPA									
	Número de talhões informados 1	Limite inferior (kg/ha) 2	Limite superior (kg/ha) 3	Média (kg/ha) 4	Mediana (kg/ha) 5	Moda (kg/ha) 6	Rendimento LSPA Abril/87 7	Variação percentual (%)		
								7/4 8	7/5 9	7/6 10
Algodão.....	410	930	2 851	1 816	1 860	1 860	1 667	-8,2	-10,3	-10,3
Arroz.....	556	494	2 789	1 441	1 364	1 488	1 739	20,6	27,4	16,8
Batata-inglesa (águas).....	29	8 678	30 992	17 653	19 835	19 835	16 379	-7,2	-17,4	-17,4
Batata-inglesa (secas).....	29	2 182	35 950	16 006	17 355	24 793	10 000	-37,5	-42,3	-59,6
Feijão (águas)	866	174	1 165	531	496	496	540	1,6	8,8	8,8
Feijão (secas).....	70	96	1 614	646	620	992	529	-18,1	-14,6	-46,6
Mandioca.....	300	10 703	29 298	18 607	20 000	20 661	20 000	7,4	-	-3,1
Milho	2 342	717	4 242	2 355	2 479	2 479	2 612	10,9	5,3	5,3
Milho (safrinha) ..	212	496	3 471	1 973	2 479	2 479	1 800	-8,7	-9,2	-27,3
Soja	545	1 364	2 156	2 156	1 983	1 983	2 100	-2,6	-0,3	5,9

bastante sensível às mudanças climáticas. Os dados de encerramento da safra de 1987 é que determinarão qual dessas duas hipóteses é a mais correta.

5.6.2 — Arroz

Os dados de rendimento médio do arroz, na pesquisa, foram obtidos dos levantamentos de 556 talhões. O menor rendimento médio obtido foi de 496 kg/ha e o maior, de 2.789 kg/ha. Como no caso do algodão, a média (1.441 kg/ha), mediana (1.368 kg/ha) e a moda (1.488 kg/ha) estão muito próximas.

A estimativa de abril do LSPA (1.739 kg/ha) é 20,6% mais elevada que a média da pesquisa. É interessante observar que esta se aproxima do rendimento médio do LSPA referente à safra de 1986 (1.471 kg/ha); é possível que os resultados médios desta safra tenham afetado as expectativas dos produtores, no início do ano, em relação à safra de 1987.

5.6.3 — Batata-inglesa

Vimos que um dos problemas da pesquisa foi o número reduzido de segmentos da amostra sempre que os produtos são cultivados apenas em áreas restritas do Paraná.

O caso mais extremo foi o da batata-inglesa.

Esse problema se refletiu, também, nas estatísticas de rendimento médio; tanto para a batata-inglesa da 1.ª safra como da 2.ª safra, o número total de talhões da amostra foi de apenas 29. Como à época da pesquisa nem havia sido completado o plantio da batata-inglesa 2.ª safra, será feita uma análise apenas dos resultados concernentes à 1.ª safra do produto.

Na amostra, os rendimentos médios da batata-inglesa 1.ª safra se situaram entre os extremos de 8.678 kg/ha e de 30.992 kg/ha. A média foi de 17.653 kg/ha, ou seja, 7,8% acima da estimativa do LSPA (16.379 kg/ha). Entretanto, a mediana e a moda (19.835 kg/ha nos dois casos) se apresentaram bastante superiores à média.

Na verdade, o tamanho reduzido da amostra limita em muito a validade das estimativas de rendimento da batata-inglesa da pesquisa.

5.6.4 — Feijão

Serão analisadas apenas as estatísticas referentes ao rendimento do feijão da 1.ª safra. Na ocasião da pesquisa, o plantio do feijão da 2.ª safra ainda não havia sido con-

cluído e foi reduzido o número de talhões com o produto na amostra (70).

Para o feijão das águas (1.ª safra), a pesquisa obteve informações de 866 talhões. Os rendimentos médios declarados variaram entre um mínimo de 174 kg/ha e um máximo de 1.165 kg/ha. A média ficou em 531 kg/ha, próxima da estimativa de abril do LSPA (540 kg/ha), e não muito distante da mediana e da moda (496 kg/ha).

Considerando que, em janeiro/fevereiro de 1987, a colheita do feijão das águas estava no seu auge, os dados da pesquisa refletiram de perto o rendimento efetivo do produto. Ao que tudo indica, o LSPA de abril captou um valor para o rendimento médio próximo ao que efetivamente se verificou.

5.6.5 — Milho

Como nos casos da batata-inglesa e do feijão, a análise do rendimento do milho se refere apenas à sua safra principal. À época da pesquisa, o milho da "safrinha" ainda não estava todo plantado, e o número de talhões dessa cultura na amostra (212) não foi muito elevado.

A amostra da pesquisa captou informações de rendimento médio de 2.342 talhões; obteve dados desse atributo que variaram entre 717 kg/ha e 4.242 kg/ha. A média e mediana e a moda do rendimento estiveram bastante próximas; elas foram, respectivamente, 2.355 kg/ha, 2.479 kg/ha e 2.479 kg/ha. A estimativa de rendimento médio do LSPA (abril) foi de 2.612 kg/ha, 10,9% superior à média obtida na pesquisa.

Não se encontrou explicação para essa discrepância entre as estimativas da pesquisa e do LSPA. Ela pode ter resultado do caráter subjetivo desse último levantamento, mas seria necessário um exame mais detalhado da evolução, em 1987, da safra de milho, para que tivéssemos condições de afirmar que é este mesmo o problema.

5.6.6 — Soja

A amostra do Paraná registrou o cultivo da soja em 545 talhões. Os valores extremos do rendimento esperado pelos que o cultivaram foram de 1.364 kg/ha e de 2.851 kg/ha. A média do rendimento

da amostra atingiu 2.156 kg/ha, valor muito próximo ao avaliado pelo LSPA (2.100 kg/ha). Não houve, também, acentuada discrepância entre a média e a mediana (2.107 kg/ha) e a moda (1.983 kg/ha). Ao que tudo indica, no caso da soja, as expectativas dos produtores se refletiram, de forma considerável, na avaliação dos entendidos que participam dos levantamentos do LSPA.

6 — PERSPECTIVA ATUAL E EVOLUÇÃO DOS TRABALHOS FUTUROS

O painel de amostras de área é completo no sentido estatístico, pois todas as suas unidades de investigação (estabelecimentos agropecuários) possuem uma probabilidade de seleção conhecida. É, também, o mais indicado para pesquisas de múltiplos propósitos, na área da agropecuária, podendo ser, ainda, utilizado em pesquisas sócio-econômicas na área rural, em que a definição da unidade de investigação é mais difícil de estabelecer. Como aspecto negativo, deve ser ressaltada a sua ineficiência para estimar variáveis que possuem a característica de serem raras, isto é, exploradas por poucos produtores, ou concentradas em pequenas regiões geográficas.

Embora os investimentos, em termos de custo de implantação e tempo, sejam altos, o tempo médio de vida útil de um painel é alto, usualmente, chegando a mais de 15 anos de utilização.

Os resultados da pesquisa do Paraná analisados, à luz da metodologia estatística, indicam a possibilidade de utilização de algumas técnicas que podem ser consideradas como soluções alternativas, para os principais problemas de estimativas da pesquisa. As principais são:

a) Para reduzir os coeficientes de variação das estimativas referentes à batata-inglesa, existem duas possibilidades: a de criar um estrato especial para a batata-inglesa e modificar o desenho da amostra visando a melhorar as estimativas referentes a esse produto; ou montar um cadastro dos grandes produtores, empregando, para a batata-inglesa, um painel múltiplo, num procedi-

mento semelhante ao adotado na estimativa de bovinos da pesquisa. Nesta decisão, devem ser levados em conta não só a possível eficiência como os custos de cada uma delas.

b) No caso do algodão, uma redução de coeficientes de variação pode ser tentada mediante a construção de uma lista de grandes produtores e o seu emprego em conjunção com o painel de amostras de área constituindo, aqui também, um painel múltiplo;

c) Para o caso dos suínos, a solução mais simples é utilizar um cadastro de grandes suinocultores, elaborado com base nos resultados do Censo Agropecuário, a serem incluídos na amostra com certeza. O procedimento é semelhante ao adotado no levantamento de bovinos;

d) Empregar o método do estimador ponderado para tentar melhorar a estimativa das quantidades armazenadas em estabelecimentos agropecuários, dos principais produtos agrícolas; e

e) Estudar as diferentes possibilidades de alocação da amostra. Apesar de se reconhecer que uma alocação ótima para uma variável pode não ser a mais adequada para as demais (o que na maioria das vezes acontece), um estudo de diferentes alocações pode ser uma solução alternativa.

Em suma, verifica-se que o emprego de estimadores de expansão direta, com os dados de campo da amostra probabilística de segmentos, permitiu estimar as áreas das principais lavouras do Paraná, cultivadas na safra de verão de 1987. A precisão desses estimadores pode ser melhorada usando-se um estimador de regressão obtido com base nas informações da análise de dados digitais de satélite.

A técnica para integrar os dados digitais de satélite com os dados de campo, proveniente da amostra probabilística, foi desenvolvida pelo Departamento de Agricultura dos Estados Unidos — USDA que, a partir de 1980, aplica regularmente para estimar a área de algumas culturas em vários estados. O INPE, que já realizou diversos experimentos, em escala reduzida, com o objetivo de adaptar a metodologia às condições do Brasil, está se empenhando em tornar essa metodologia aplicável ao novo sistema de previsão de safras agrícolas do IBGE.

É importante ressaltar que se trata de metodologia sofisticada, requerendo pessoal técnico de alto nível, equipamentos e programas de computadores especiais. Para a pesquisa do Paraná de 1987, o sistema de classificação digital está sendo testado em relação a uma parte do estado, com o objetivo de verificar o funcionamento do *hardware* e do *software* criados para tal e de preparar uma aplicação mais ampla no futuro.

Para realizar seus testes, o INPE estabeleceu um distrito de análise digital. Após receber informações de cobertura de nuvens de todas as passagens do TM/LANDSAT entre 15 de janeiro e 16 de março (a época do levantamento de campo) os técnicos do INPE decidiram empregar, nesta fase, somente a passagem TM-LANDSAT do dia 10 de fevereiro, para a órbita 223, pontos 76, 77 e 78. Os dados das cenas do LANDSAT constituíram o distrito de análise (DA-223) formado para os testes.

Feito isto, os talhões dos segmentos da amostra contidos na área do DA-223 foram localizados nas respectivas imagens LANDSAT e as informações de campo serviram de base para estabelecer a resposta espectral relacionada a cada tipo de cultura, nos diferentes estágios de desenvolvimento encontrados à época da pesquisa. Essas respostas espectrais estão sendo usadas para estabelecer uma relação entre a classificação digital e os dados de campo dos segmentos da amostra. Elas permitem associar cada *pixel* a um dado tipo de cultura; essa cultura estabelece uma certa assinatura espectral que, se mantida de forma coerente no espaço, servirá de base às estimativas das áreas com ela cultivadas — tanto dentro como fora dos segmentos da pesquisa.

É esta a base da técnica da regressão que vem sendo desenvolvida. Se houver uma correlação elevada entre as informações coletadas em campo sobre uma dada lavoura e a identificação, estabelecida pela classificação digital, dessa mesma lavoura nos segmentos da amostra, é muito grande a probabilidade de que a classificação digital, feita sobre todo o distrito de análise (sobre a área total pesquisada), forneça uma estimativa mais eficiente da área com a lavoura nele contida. Acontecendo isto, a precisão

dos estimadores será consideravelmente ampliada.

Um artigo futuro, específico sobre o assunto, trará maiores detalhes sobre esta metodologia.

Vale ressaltar que, não possuindo experiência prévia com o assunto, os técnicos do INPE e do IBGE envolvidos no projeto de previsão de safras subestimaram tanto as dificuldades que encontrariam em sua implantação, como o prazo necessário para se obterem resultados, uma vez disponíveis o material coletado em campo e os dados do LANDSAT. Na visita que técnicos do projeto efetuaram aos Estados Unidos, foi possível constatar que, a despeito da experiência, dos recursos e do equipamento de que a área de sensoriamento remoto do USDA dispõe, a previsão de safras com o emprego da classificação automática de dados de satélite ainda se encontra em estágio experimental; ademais, ainda leva de três a seis meses para que surjam os resultados finais referentes a um dado estado. O USDA se baseia essencialmente em estimativas

obtidas a partir de amostras resultantes de painéis de área e de cadastros múltiplos para realizar suas estimativas de safras. A metodologia de sensoriamento remoto é usada, complementarmente, para aprimorar essas estimativas em áreas selecionadas (os oito principais estados agrícolas do País).

A maior compreensão da exata extensão da tarefa que enfrentavam fez com que o INPE e o IBGE reavaliassem o cronograma dos trabalhos e a magnitude do objetivo a ser atingido no seu primeiro experimento de aplicação da classificação dos dados digitais de satélite para levantamento de safras. Trata-se, entretanto, de experimento crítico; a experiência com ele adquirida permitirá que, no futuro, se cubram áreas mais extensas em prazos mais curtos. Ele possibilitará, também, avaliar a adequação dos equipamentos usados no processo de classificação digital e estabelecer um programa de complementação do sistema de processamento disponível no INPE e, eventualmente, de formação de um sistema no IBGE.

SECRETARIA DE PLANEJAMENTO DA PRESIDÊNCIA DA REPÚBLICA FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA — IBGE DIRETORIA DE AGROPECUÁRIA, RECURSOS NATURAIS E GEOGRAFIA SUPERINTENDÊNCIA DE ESTATÍSTICAS AGROPECUÁRIAS Departamento de Estatísticas Contínuas Agropecuárias			CÓDIGOS		PERÍODO DE COLETA		NÚMERO DO ESTRATO		NÚMERO DA UP	
CONDICÃO DE RESIDÊNCIA DO RESPONSÁVEL			DE USO DA TERRA SEGUNDO ÁREAS COM OUTRA(S) EXPLORAÇÃO(ões) OU OCUPAÇÃO(ões)		INÍCIO	TÉRMINO	SEGMENTO		Nº DA FOLHA	
1 - Residente 2 - Não residente			1 - Criação de Outros Animais 2 - Piscicultura 3 - Edificações 4 - Cursos d'água 5 - Outras		/ /	/ /	NÚMERO	ÁREA (ha)	Nº DA FOTOGRAFIA AÉREA	/
PESQUISA AGROPECUÁRIA DE 1987 - PARANÁ									MUNICÍPIO	
CONVÊNIO IBGE/INPE										
PA.2 - FOLHA DE LISTAGEM										
VERIFICAÇÃO DA ÁREA DENTRO DO SEGMENTO			IDENTIFICAÇÃO DO RESPONSÁVEL		INFORMAÇÕES SOBRE O TIPO DE EXPLORAÇÃO AGROPECUÁRIA NA ÁREA ASSINALADA			ÁREA COM OUTRA(S) EXPLORAÇÃO(ões) OU OCUPAÇÃO(ões)		
Identificação da Área	Código de Residência do Responsável	Existe(m) Outra(s) Pessoa(s) Produzindo Independentemente, Dentro da Área Assinalada? (em caso afirmativo, preencha na coluna 4, quantas linhas forem necessárias)	Nome Completo	Endereço (via, rua ou estrada)	Desenvolveu Alguma Atividade Agrícola e/ou de Pecuária Bovina?	Registre a Área Dentro do Segmento e o Número do Questionário. A Seguir, Preencha o Questionário e Encerre a Entrevista		Registre a Área Dentro do Segmento		Registre o Código de Uso da Terra
						Área	Unidade de superfície nome Equivalência em m²	Número do questionário	Área	
1	2	3	4	5	6	7	8	9	10	11
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					
		Sim ☐ Não ☐			Sim ☐ → Não ☐ vá para a coluna 9					

BIBLIOGRAFIA

- CLARK, M. S.; Geuder, J. *Use of digital Landsat data to stratify area sampling frames for agricultural surveys*; Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture; Washington, March 1986 — SRS Staff Report.
- COCHRAN, W. G. *Sampling Techniques*, Third Edition, John Wiley, 1977.
- COOK, P. W. *A continuing development of an application for Landsat data: 1983 DCLC Wind wheat acreage estimates for four states*. Proceedings of the 18th. International Symposium on Remote Sensing of Environment; Paris, 1984.
- ESTATÍSTICAS Agropecuárias, IBGE, p. 1-27.
- GEUDER, J. *An evaluation of SRS area sampling frame designs — Ordering count units and creating paper strata*. Statistical Reporting Service, U.S. Department of Agriculture; Washington, 1983.
- *Paper stratification in SRS area sampling frames*; Statistical Research Division, SRS, U.S. Department of Agriculture; Report 79, Washington, 1984.
- GLEASON, C. *The auxiliary use of Landsat data in estimating crop acreages: results of the 1975 Illinois crop-acreage experiment*; SRS, U.S. Department of Agriculture, 1977.
- HANSEN, M. H.; Hurwitz, W. N.; Madow, W. G. *Sample Survey Methods and Theory*, v. I/II, John Wiley, 1953.
- HANUSCHAK, G. A.; Allen, R. D.; Wigton, W. H. *Integration of Landsat data into the crop estimation program of USDA's Statistical Reporting Service 1972 — 1982*; 1982 Machine Processing of Remotely Sensed Data Symposium, Purdue University; Indiana, 1982.
- HOLKO, M. L.; Sigman, R. S. *The Role of Landsat Data In Improving U.S. Crop Statistics*; Proceedings of the 18th. International Symposium on Remote Sensing of Environment; v. I, p. 307-320; Paris, 1984.
- HOUSEMAN, E. E. *Area Sampling Frame in Agriculture*, Statistical Reporting Service, United States Department of Agriculture, Washington, DC. 1975.
- HUDDLESTON, H. F. *A training course in sampling concepts for agricultural surveys*; Statistical Reporting Service, Publ. 21, U.S. Department of Agriculture; Washington, 1980.
- KISH, L. *Survey Sampling*, John Wiley, 1965.
- NEALON, P. *Review of the multiple area frame estimators*. Statistical Research Division, SRS n° 80, U.S. Department of Agriculture; Washington, 1984.
- RUSSELL, R.; Huddleston, H. F. *Surveys of crop production in Jamaica*. United Nations Seminar, San José, Costa Rica, 1980.
- WIGTON, W. H. *Production of agricultural statistics using satellite data*. Manuscrito, p. 1-105.

RESUMO

Com a finalidade de implementar um sistema objetivo e eficiente de produção de dados agropecuários, decidiu-se aplicar metodologia de pesquisa, no Estado do Paraná, semelhante à utilizada pelo United States Department of Agriculture-USA, em sua June Enumerative Survey, baseada em amostragem probabilística de áreas (segmentos).

O trabalho de campo, que consistiu na aplicação de 2.721 questionários a produtores de estabelecimentos agropecuários, foi realizado por 30 entrevistadores e 10 supervisores. As informações coletadas eram referentes a características gerais dos estabelecimentos, armazenagem e estocagem de grãos, pecuária bovina, áreas plantadas, a ser plantada e colhida, e rendimento médio para os principais produtos.

Para a maioria das estimativas das variáveis em estudo, foi utilizado um estimador do tipo fechado, sendo que, no caso da pecuária bovina, utilizou-se um estimador ponderado associado à técnica de painéis múltiplos já que se possuía um cadastro dos 259 maiores produtores do estado, segundo dados do Censo Agropecuário de 1985.

A pesquisa foi baseada na investigação de uma amostra interpenetrante de 450 segmentos de área selecionada de um modelo estratificado com um único estágio de seleção.

Na construção do painel de amostragem, que se constitui na subdivisão da área de investigação em 6 estratos de uso do solo, com limites permanentes perfeitamente identificáveis, foram utilizados da-

dos do Censo Agropecuário de 1980, cartas topográficas, interpretação visual de imagens de satélite (TM/LANDSAT canais 3 e 4, correspondentes ao verão de 1985), além de uma base cartográfica digitalizada, possibilitando o cálculo da área dos estratos e das UPAs — Unidades Primárias de Amostragem.

As UPAs foram ordenadas utilizando-se técnicas de análise de conglomerados, associando-se o critério da contigüidade geográfica, de modo a se introduzir no modelo uma subestratificação geográfica (zoneamento de áreas). A amostra de 450 segmentos foi, então, selecionada e plotada num fotomosaico, escala 1:25.000, do Estado do Paraná. Finalmente, os limites dos segmentos foram transferidos para aerofotografias, escala 1:25.000 e 1:10.000, correspondentes ao ano de 1980, que se constituiu no material básico para a coleta de dados no campo.

ABSTRACT

In order to produce timely agricultural data based on probability survey methods and taking into account the characteristics of the State of Paraná (around 200.000 sq. km), it was decided to apply area survey sampling methods following, mainly, the methodology used in the USDA June Enumerative Survey.

The questionnaire applied to each one of the 2.721 agricultural establishments, partially or totally included in the sampled segments, was used by 30 interviewers and 10 supervisors to collect ground data, in particular, on general characteristics of the establishments, grain and stocks, cattle and calf inventories, planted, to be planted and harvested area of the main crops, and corresponding yields, in order to forecast production.

A closed estimator was used for most of the variables of the study. The questionnaire was also applied to a 1985 Census list frame of 259 cattle ranches, in order to use a multiple frame estimator for cattle and calf inventories.

The survey was based on the investigation of a sample of 450 area segments, selected from an one-stage stratified interpenetrating design.

The construction of the sampling frame, a subdivision of the state — with identifiable permanent physical boundaries — into six land-use strata and 6.926 primary sampling units, was accomplished by using 1980 Census data, visual interpretation of satellite images (LANDSAT/TM, channels 3 and 4, Summer, 1985), topographic maps and a cartographic base of the state that was digitized, producing overlays and tabulations, with the area of the strata and of the PSU's.

The PSU's were ordered — using clustering techniques and geographical contiguity — to achieve further geographical stratification, and then 450 selected PSU's and segments were located into a 1:25.000, 1980 photomosaic available for the state. Finally the boundaries of the sample segments were transferred to corresponding 1980 photos, which were used for the field work.

ESTUDO DAS FERRAMENTAS PARA APURAÇÃO DE DADOS

Dulce Maria Rocha Barbosa*

Reina Maria Hanono*

1 — INTRODUÇÃO

Neste trabalho apresentamos uma análise das ferramentas existentes no mercado para aplicação em sistemas de apuração de dados, buscando atenuar os problemas de *software* encontrados no desenvolvimento destes sistemas. Para cada uma destas ferramentas é feita uma análise citando suas vantagens e desvantagens para uso genérico em qualquer tipo de pesquisa.

O IBGE se caracteriza pela realização de inúmeras pesquisas, destacando-se os censos (econômicos e sociais) e as pesquisas contínuas (mensais e anuais). Estas pesquisas passam pelas fases de Definição (Âmbito da Pesquisa, Planos de Crítica e Disseminação), Coleta, Transcrição, Crítica, Análise dos Dados e Tabulação dos mesmos para divulgação.

Para facilitar a execução das fases mencionadas, foram desenvolvidas ferramentas, na DI, que proporcionassem uma padronização de procedimentos visando a acelerar o desenvolvimento dos projetos nesta Diretoria.

A opção adotada objetivava solucionar os problemas encontrados em processos isolados do Censo de 80, de uma forma eficiente, sem levar em conta os aspectos de metodologia de desenvolvimento, integração, documentação e portabilidade.

O que é observado no transcorrer destas fases é que o Ciclo de Crítica para verificação dos erros de transcrição e consistência interna dos dados, envolvendo correções, tanto dos erros de digitação quanto das informações iniciais dos informantes, normalmente gera um volume de programação que consome semanas, até o teste final. As atividades realizadas nesta fase podem ser resumidas por:

- a) *Especificar plano de crítica* — estabelecimento, pelo usuário, dos procedimentos necessários para execução da tarefa de crítica.
- b) *Analisar procedimentos* — estudo, pelos analistas, da viabilidade de implementação dos procedimentos estabelecidos e especificação destes para o programador.
- c) *Codificar especificação* — codificação do(s) programa(s) correspondente(s) ao(s) procedimento(s) especificado(s).

* As autoras são Analistas de Sistema da Fundação Instituto Brasileiro de Geografia e Estatística — IBGE.
R. bras. Estat., Rio de Janeiro, 49(191):85-100, jan./jun. 1988.

d) *Verificar procedimentos* — teste dos programas gerados.

Constatamos, através de experiências e informações levantadas junto às equipes desenvolvedoras destes tipos de sistemas, a ocorrência de alguns problemas motivados por:

- a) inexistência de um padrão comum para especificação dos Planos de Crítica;
- b) inexistência de um apoio automatizado que dê suporte a estas especificações;
- c) inexistência de uma ferramenta que auxilie a comunicação entre os técnicos responsáveis pelas pesquisas e as equipes desenvolvedoras;
- d) inexistência de uma ferramenta automatizada que garanta a implementação de um código fiel aos procedimentos preestabelecidos.

Visando resolver estes problemas, iniciou-se um estudo para especificação dos requisitos de uma ferramenta que permitisse dotar os usuários de um auxílio automatizado para especificação dos procedimentos de crítica, assim como a geração automática dos programas correspondentes. Estes requisitos podem ser resumidos como:

- a) permitir seu uso por usuários, analistas e programadores;
- b) permitir maior segurança para especificação das tarefas, pela comunidade de usuários, analistas e programadores;
- c) permitir um auxílio que promova uma interação entre a especificação de uma tarefa e o desenvolvimento de sua implementação.

Uma vez estabelecidos os requisitos, vimos a necessidade de uma linguagem de programação voltada a agilizar as etapas de crítica e imputação (correção automática). Assim, foi feito um estudo comparativo entre algumas ferramentas de apuração de dados estatísticos.

A escolha das ferramentas teve em conta a difusão das mesmas nos grandes centros estatísticos, grupos de computação estatística e acadêmicos, e sua utilização em ambientes *mainframe*, até hoje utilizados em nossa Instituição para o processamento de grande volume de dados.

As ferramentas estudadas foram: CONCOR, UNEDIT, ATLAS, SISQUE, AERO, CAN-EDIT, OSIRIS IV e P-STAT,

utilizando-se como objetos de comparação as características de especificação, os recursos oferecidos, os tipos de arquivos processados e as características de processamento.

A seguir, é apresentada uma análise de cada uma das ferramentas citando suas vantagens e desvantagens para uso genérico em qualquer tipo de pesquisa em relação aos requisitos estabelecidos.

Logo após estas análises são apresentadas três tabelas comparativas visualizando os aspectos de processamento, funcionamento e recursos.

2 — CONCOR

Concor data editing package é um sistema desenvolvido pelos membros do CELADE, no ano de 1974, para consistência e correção de dados.

É formado por um conjunto integrado de programas que podem identificar e corrigir dados inválidos ou inconsistentes que estejam sendo preparados para tabulação e análise.

Foi desenvolvido especialmente para tratamento de dados censitários sendo um dos primeiros *softwares* desenvolvidos para este objetivo.

Requisitos para o arquivo de entrada:

- a) o arquivo tem de ser seqüencial;
- b) todos os registros têm de ter tamanho fixo;
- c) tem de existir uma única identificação de questionário, formada por dois campos de identificação, podendo ser no caso dos censos demográficos, o código da região geográfica, e número de domicílio;
- d) cada registro tem que ter um código de tipo de registro.

Através da linguagem CONCOR, o usuário descreve o arquivo a ser processado, as edições (críticas e correções) e os tipos de estatísticas a serem efetuadas.

Esta linguagem consiste de três divisões:

- a) *Dictionary division* — contém os comandos que definem e descrevem os arquivos de entrada, saída, os itens de dados, e as variáveis a serem usadas nas operações de edição do arquivo;

b) *Execution division* — contém os comandos para edição que se desejam sobre os dados, e os tipos de estatísticas que deverão ser efetuadas;

c) *Report division* — contém comandos para descrição de relatórios a serem impressos a partir das estatísticas realizadas sobre os dados do arquivo.

Nestas três divisões deve ser utilizado um formato-padrão para a entrada das especificações.

A linguagem CONCOR é procedural e tem dois tipos de comandos:

A — Comandos não executáveis (instruções para o compilador);

B — Comandos executáveis, que podem ser divididos nas seguintes categorias:

a) comandos para recodificação e conversão de dados;

b) comandos aritméticos;

c) comandos de verificação;

d) comandos de controle de fluxo de processamento;

e) comandos de alocação dinâmica de matrizes;

f) comandos de geração de registros de saída.

As informações destas três divisões são: a entrada para o sistema CONCOR que as analisa, e, caso corretas, gera um programa COBOL correspondente, e este, quando em execução, lê um registro do arquivo de entrada, processa-o de acordo com as especificações fornecidas pelo usuário no programa CONCOR e produz um arquivo "Limpo".

2.1 — Capacidade de processamento

O CONCOR possui as seguintes funções de edição de dados:

a) checagem de estrutura — verifica se todos os registros que formam um questionário estão presentes, e se não existem registros extras para um questionário;

b) checagem de limite de valores — determina se o valor de uma variável está compatível com os limites de valores estipulados para ela;

c) checagem de consistências — verifica se dois ou mais valores de variáveis, do mesmo registro ou não, estão consistentes de

acordo com as condições permitidas especificadas pelo usuário;

d) correção automática — permite que, além da identificação de um erro numa variável, esta possa ser corrigida de acordo com as especificações do usuário. O método de correção, através de matriz *hot-deck*, também é permitido;

e) arquivo de saída — permite a criação de um arquivo contendo os dados corrigidos, sendo seu formato idêntico ao do arquivo de entrada;

f) arquivos derivados — o CONCOR permite a criação de um arquivo que poderá conter qualquer combinação dos dados do arquivo de entrada, ou variáveis derivadas, permitindo, assim, a criação de um arquivo contendo dados recodificados ou reformata-dos;

g) relatórios de erros — o CONCOR gera estatísticas sobre os testes de condições executados, e o número de imputações feitas. Os relatórios destas estatísticas são gerados de acordo com as descrições fornecidas pelo usuário.

2.2 — Relatórios de Crítica

O CONCOR pode produzir quatro tipos de relatórios:

a) número e percentagem de erros encontrados e correções efetuadas pelo CONCOR;

b) número e percentagem de erros encontrados e correções efetuadas para cada item de dados editados;

c) listagem de todos ou alguns itens de dados de um registro ou questionário;

d) número e percentagem dos valores que foram usados como correções para itens de dados particulares.

As informações para cada um destes relatórios são de completo controle do usuário, sendo a formatação dos relatórios feita pelo CONCOR, podendo ser feita quebra ou seleção por área geográfica.

2.3 — Vantagens e Desvantagens

Para esta ferramenta, em relação aos objetivos propostos, podemos citar como vantagens e desvantagens os seguintes itens:

VANTAGENS:

- a) atribuição automática de valores a variáveis quando da ocorrência de incoerências com suas especificações;
- b) acesso a registros diferentes através de um índice de ocorrência para comparação de variáveis inter-registros;
- c) correção automática através de comandos de atribuição, podendo-se utilizar ou não uma matriz de imputação;
- d) geração automática de relatórios de estatísticas de ocorrências de erros e correções.

DESVANTAGENS:

- a) só permite arquivos sequenciais;
- b) todos os registros têm que ter o mesmo tamanho, sendo o tamanho máximo permitido de 999 bytes;
- c) o teste de validade para variáveis categorizadas tem de ser programado pelo usuário;
- d) só permite processamento por quebra de até dois níveis.

3 — UNEDIT

O UNEDIT é um *software* para crítica e correção de Dados Censitários a serem preparados para análise e tabulação. Este sistema, embora possua algumas opções de correção automática, tem como seu maior objetivo a pesquisa de erro que precede à correção manual, e foi desenvolvido tendo como objetivo principal a aplicação em Censos Demográficos.

O sistema UNEDIT consiste de dois módulos. O primeiro examina as especificações do programa escrito pelo programador em linguagem UNEDIT, checando os erros sintáticos e lógicos. O segundo módulo lê o arquivo de entrada e identifica os registros contendo os erros de acordo com as especificações fornecidas pelo usuário.

As estatísticas dos erros, por tipo de erro e nome do campo errado, são impressas para cada área geográfica separada.

3.1 — Características de especificação do usuário

O usuário especifica seu programa utilizando cinco tipos de formulários. Três deles

para o dicionário, um para especificação dos cálculos (aritméticos ou lógicos), a serem efetuados, e um para especificação de críticas e imputações.

3.1.1 — Especificação de dicionário

Formulário 1

Descrição de arquivos — contém comandos para definir e descrever os arquivos de entrada e saída, a forma de processamento (registro a registro ou um grupo de registros a cada vez) e o tamanho e composição da chave do registro.

Formulário 2

Descrição de registros — contém comandos para definir os registros do arquivo de entrada e suas variáveis. Admite como número máximo a definição de nove tipos de registros, e somente permite definir variáveis zonadas (numéricas ou não). Para o caso em que exista mais de uma ocorrência de um determinado tipo de registro num mesmo questionário permite a especificação de uma variável para identificar cada ocorrência, variável esta que será utilizada para fazer checagem de estrutura e/ou de consistência de valores de variáveis inter-registros.

Formulário 3

Especificação de códigos — especifica por item de dado os códigos válidos e valores *default* para os casos de erro (código inválido e branco), a serem utilizados nas operações. Este formulário é opcional.

3.1.2 — Especificação de cálculos prévios à edição

Formulário 4

Especificação de cálculos prévios à edição — é utilizado para atribuição de resultados de cálculos ou comparações a novas variáveis ou variáveis já existentes, definidas no Formulário 2, a serem utilizadas no Formulário 5. Permite, também, criar uma nova variável, a partir de uma variável já definida no Formulário 1 para uma ocorrência identificada de um tipo de registro, o que facilitará a checagem inter-

-registros (exemplo: para o caso de um questionário que contenha mais de um tipo de registro de pessoa, podemos criar a variável IDADE-CHEFE a partir da variável IDADE do registro de pessoa correspondente ao chefe). Este formulário é opcional.

3.1.3 — Especificação de inconsistências

Formulário 5

Especificação de inconsistências — especifica combinações de valores de variáveis que são considerados inconsistentes (erro) e, também, é utilizado para imputação de valores a variáveis erradas.

3.2 — Vantagens e Desvantagens

Para esta ferramenta podem-se destacar os seguintes itens como vantagens e desvantagens:

VANTAGENS:

- a) checagem automática da estrutura de um questionário, acusando ausência de tipo de registro no questionário;
- b) atribuição automática de valores a variáveis quando da ocorrência de inconsistência com suas especificações;
- c) verificação automática da validade de uma variável categorizada;
- d) identificação de um registro num grupo de registros repetidos do mesmo tipo;
- e) geração automática de relatórios de estatísticas de erros e relatório de crítica.

DESVANTAGENS:

- a) sistema de correção automática muito limitado, consistindo, apenas, de comandos de atribuição;
- b) só permite a definição de nove tipos de registros;
- c) o uso de vários tipos de formulários e a especificação da linguagem em campos formatados tornam difícil o aprendizado e a compreensão de um programa nesta linguagem.

4 — ATLAS

O ATLAS é um sistema para geração de programas de Crítica, Imputação e Tabulação de Dados na linguagem PL/I.

O processamento deste sistema é constituído de duas etapas básicas: Descrição dos Dados e Descrição dos Resultados.

Na primeira, o usuário descreve o arquivo de entrada, seu tipo, os registros, as variáveis e as categorias existentes para cada uma das variáveis categorizadas. Além disso, descreve as variáveis auxiliares a serem utilizadas no programa, bem como os demais arquivos a serem usados no processamento.

Na segunda parte, o usuário define o seu programa utilizando as estruturas permitidas pela linguagem que são formadas pelo uso de processos automatizados, pré-programados, cada um tendo uma função específica, e um conjunto de instruções inerentes a cada processo.

4.1 — Características de desenho

O ATLAS suporta as seguintes funções de edição de dados:

- a) Correção automática (através de atribuição de valores e/ou através do processo IMPUTACÃO podendo usar o modelo de matriz *HOT-DECK*).
- b) Crítica — geração automática do relatório de crítica opcionalmente com as estatísticas dos erros emitidos, ou gravação do arquivo de crítica;
- c) Geração automática da atualização bem como do relatório de ocorrências da atualização;
- d) Geração do arquivo de saída, arquivo de entrada corrigido ou arquivos formatados pelo usuário;
- e) Geração automática de rotinas de pesquisa de códigos;
- f) Geração de relatórios de frequências, tabulações e relatórios formatados pelo usuário.

4.2 — Características de especificação do usuário

A linguagem de especificação consiste de dois módulos:

Dicionário

Contém comandos para definir e descrever os arquivos de entrada e saída, os registros, as variáveis do registro de entrada (que poderão ser matrizes), assim como as

variáveis de trabalho a serem usadas no programa e a descrição dos arquivos auxiliares.

Programa

Contém os processos e procedimentos para a Crítica, Imputação e Tabulação dos Dados, assim como para a emissão de relatórios.

Nos dois módulos deverá ser utilizado um formato-padrão para a entrada das especificações, Formulário 3 para a descrição dos dados e Formulário 4 para a descrição dos resultados.

A informação contida nestas duas divisões constitui a entrada para o sistema ATLAS. O sistema faz uma análise sintática dos comandos e se não houver erros gera um programa-fonte em PL/I. Este é compilado e linkeditado e poderá ser testado antes de entrar em produção.

O sistema oferece a vantagem de se poder tratar o módulo-dicionário, separadamente, fazendo uma análise deste, permitindo, assim, que vários programas utilizem o mesmo dicionário.

4.3 Tipos de comandos

O ATLAS possui os seguintes processos automatizados:

- a) processos de início e término do programa;
- b) processos de leitura dos arquivos de entrada;
- c) processo para controle de execução;
- d) processos para criação de dados;
- e) processos para acompanhamento;
- f) processo para crítica;
- g) processo para atualização;
- h) processos para emissão de relatórios;
- i) processos para gravação de arquivos;
- j) processos de tabulação;

4.4 — Capacidades de processamento

O ATLAS permite processar qualquer tipo de arquivo, com registros de tamanho até 32 767 bytes, e admite para as variáveis qualquer tipo de formato e redefinição.

Possibilita o processamento dos registros do arquivo de entrada de forma seqüencial, assim como um grupo de registros armaze-

nados em memória, a cada vez, em função das variáveis especificadas para a quebra.

O ATLAS permite, também, o tipo de processamento *balanced-line* entre o arquivo de entrada e os arquivos paralelos, declarados como arquivos de referência.

A linguagem deste sistema permite as facilidades de chamadas de rotinas externas, geração de rotinas de pesquisas de códigos e passagem de parâmetros externos para o programa.

Este sistema permite que a listagem impressa pelo processo Crítica sirva para que o usuário realize as correções manuais necessárias na própria listagem, e esta sirva como o formulário para a entrada de dados. Por sua vez, o arquivo gravado na entrada de dados poderá entrar diretamente no processo de Atualização, definido nesta linguagem, que fará automaticamente as atualizações no arquivo principal.

4.5 — Vantagens e Desvantagens

Como vantagens e desvantagens deste sistema podemos citar os seguintes itens:

VANTAGENS:

- a) atribuição automática de valores a variáveis, dependendo dos valores de outras variáveis;
- b) atribuição automática de valores a variáveis, dependendo de intervalos de valores de outras variáveis;
- c) permite processar qualquer tipo de arquivo: seqüencial, fixo, variável, indexado etc.;
- d) permite processamento por quebra de até 10 variáveis especificadas para cada processo;
- e) permite a geração automática de rotinas de pesquisa de códigos declarados ou armazenados em arquivo;
- f) permite o processamento de arquivos paralelos, podendo ser lidos em forma de *balanced-line* com o arquivo principal;
- g) emissão do relatório de crítica de forma a poder ser usado na entrada de dados para posterior atualização;
- h) crítica automática de variáveis categorizadas.

DESVANTAGENS:

- a) especificação da linguagem em formulários formatados, obrigando que cada processo e comandos sejam especificados em colunas pré-determinadas;
- b) não permissão de aninhamento dos processos da linguagem.

5 — SISQUE

O SISQUE é um sistema para processamento de questionários, que consiste de uma pseudolinguagem dirigida para verificação de erros em questionários, e geração de variáveis e/ou tabelas a partir de variáveis pertencentes a um questionário armazenado num arquivo.

Foi desenvolvido pela Central de Processamento de Dados da Universidade Federal de Viçosa, para o trabalho de preparação dos dados de questionários de pesquisas sócio-econômicas.

A sintaxe da linguagem SISQUE é análoga à da linguagem SPSS, e o sistema funciona em quatro modalidades no tratamento dos dados: Fechamento, Consistência Interna, Agregação e Análise.

Os cartões-comandos que compõem a linguagem utilizam as Colunas 1 a 10 para o nome do comando. Da Coluna 11 até a 72 deverão ser colocados os parâmetros do comando, e da Coluna 73 a 80, a identificação da sequência dos cartões.

5.1 — Comandos permitidos

5.1.1 — Objetivo

Este comando pode ser especificado mais de uma vez no início do programa, e especifica a modalidade de processamento a ser executada.

5.1.2 — Arquivo

Neste comando são descritas as características do arquivo que contém os dados a serem trabalhados.

5.1.3 — Campos

Este comando descreve a distribuição lógica das variáveis dentro do arquivo, seu

formato, tamanho e registro a que pertencem.

5.1.4 — Compute

Este comando tem por finalidade a geração de novas variáveis por meio de uma expressão aritmética qualquer envolvendo outras variáveis e/ou constantes, funções matemáticas, etc.

5.1.5 — IF

Este comando permite a tomada de decisão dependendo da expressão lógica-relacional fornecida.

5.1.6 — Consiste

Este comando faz a verificação lógica da expressão contida entre parênteses. Se a expressão for falsa, o sistema imprimirá uma mensagem de erro acompanhada dos nomes e valores das variáveis envolvidas no comando e a identificação do questionário.

5.1.7 — Imprime

Causa a impressão das variáveis numéricas, tanto originais quanto criadas no processamento, descritas no campo de especificações de parâmetros do comando (Colunas 11-72).

5.1.8 — Grava

Com o uso deste comando é possível gerar um novo arquivo de dados com as variáveis selecionadas (tanto originais como criadas no processamento).

5.1.9 — Tabela

Serve para a geração de tabelas de variáveis com até três quebras. Estas tabelas contêm somatório, média, desvio, variância e número de ocorrências para cada célula da matriz de quebras.

5.1.10 — Selecione

Sua função é a de selecionar parte dos registros do arquivo de entrada que satisfaçam as condições explicitadas na expressão lógica-aritmética fornecida.

O primeiro comando da linguagem SISQUE a ser especificado é o comando Objetivo que definirá a modalidade da tarefa a ser realizada, que pode ser Fechar, Consistir, Agregar ou Analisar.

FECHAR — o programa executará os fechamentos horizontal e vertical para cada

questionário, ignorando os demais comandos especificados no programa.

O fechamento horizontal serve para verificar a exata transcrição dos dados em cada registro do questionário e para cada registro do questionário são somados todos os campos definidos para o registro, e é verificado se a soma confere com o campo de soma especificado no comando Campos, definido com o nome padronizado Fecha. Para os registros em que não tiver sido especificado este campo, não será efetuado o fechamento deste registro.

O fechamento vertical tem a função de verificar a exata presença dos registros que compõem o questionário, isto é, os registros podem estar ausentes por não conterem informações, ou devem estar presentes se as informações existem. Se este tipo de teste for para ser efetuado, então deve existir uma declaração de registro no comando Campos, com nome do campo da variável Fechavert. O sistema consistirá no somatório dos números de identificação dos registros com este campo, para cada questionário.

CONSISTIR — o programa executará todos os comandos especificados no programa, exceto os comandos Grava, Imprime e Tabela. O processamento é feito questionário a questionário, e os comandos Consiste que tiverem resultado lógico falso acusarão saída em disco para posterior classificação e impressão. A saída consta de número do questionário, nomes e valores das variáveis envolvidas no teste.

AGREGAR — esta opção está relacionada com a criação de um novo arquivo de dados, contendo os valores de variáveis originais ou geradas a partir dos comandos Compute e IF. As variáveis para este arquivo deverão ser descritas no comando Grava ou Imprime.

ANALISAR — esta opção prepara o sistema para a geração das tabelas que deverão ser impressas ao final do programa. Estas tabelas são definidas através dos comandos Tabela.

5.2 — Vantagens e Desvantagens

Para este sistema podem ser considerados como vantagens e desvantagens os seguintes itens:

VANTAGENS:

- a) crítica automática dos fechamentos horizontal e vertical;
- b) opção de especificação da modalidade de processamento desejada, sem que seja necessária a especificação de programas separados para cada modalidade.

DESVANTAGENS:

- a) campo de identificação do questionário consistindo, apenas, de dois campos — nº do questionário e tipo de registro;
- b) só permite processamento de arquivos sequenciais;
- c) não permite crítica automática de consistência das variáveis com suas especificações;
- d) especificação da linguagem de modo formatado, obrigando a especificação dos comandos em colunas pré-formatadas;
- e) não permite especificação nem crítica automática de variáveis categorizadas.

6 — AERO

O sistema AERO — Generalized Data Editing System — foi especificado por K. Szász, A. Szabó e I. Varasdy, do Departamento de Desenvolvimento de Aplicações do Centro de Computação do CSO (Oficina Central de Estatística) da Hungria.

Sua primeira versão foi implementada, em 1976, com o objetivo de utilização no Censo Demográfico de 1980, e foi usado, satisfatoriamente, na realização dos censos demográficos da Espanha, Hungria e Iugoslávia.

O AERO foi incluído como um dos pontos básicos de trabalho no UNDP Statistical Computing Project (SPC) e, como resultado desta cooperação; foi criada sua segunda versão incrementada.

Este sistema se caracteriza pela geração de programas para depuração de dados qualitativos, em linguagem PLI, que permitem as seguintes funções de processamento:

- a) validação de variáveis individuais com base numa lista de códigos armazenados num Dicionário de Dados;

- b) verificação de relacionamentos lógicos ou aritméticos entre as variáveis de um mesmo registro, usando regras de conflito especificadas (relações lógicas inaceitáveis entre os valores de duas ou mais variáveis);
- c) imputação automática dos registros errados através de um procedimento probabilístico, baseado nas próprias regras de conflito ou através de um procedimento determinístico baseado em regras de imputação determinística especificadas;
- d) verificação da composição de grupos de registros (por exemplo: o conjunto de pessoas que habitam o mesmo domicílio), e correção automática destes erros;
- e) produção de uma listagem com diagnóstico detalhado dos registros errados e das correções realizadas;
- f) produção de estatísticas sobre o resultado da validação de variáveis e verificação das regras de conflito.

Todas estas funções são executadas por programas gerados a partir de especificações fornecidas pelo usuário para cada situação específica.

6.1 — Especificações do usuário

Para cada aplicação o usuário deve especificar:

- a) a descrição do registro do arquivo de entrada;
- b) a lista de valores aceitáveis para cada variável pertencente ao registro;
- c) os parâmetros que descrevem as funções de depuração necessárias, que podem ser: validação das variáveis, verificação do cumprimento das regras de conflito, listagem dos erros, imputação automática etc.

A lista de valores aceitáveis para cada variável é especificada no módulo Dicionário de Dados do sistema, que pode ser usado por vários programas e contém comandos para inserção, deleção e modificação de variáveis no Dicionário.

Para cada variável é necessário especificar: nome, número máximo de caracteres que compõem a variável, tipo da variável (N-numérica e C-alfanumérica) e lista de códigos válidos.

A descrição do registro de entrada é feita fazendo-se referência às variáveis especifi-

cadas no Dicionário, indicando-se para cada variável do registro de entrada o nome, posição inicial e final dentro do registro, tipo de variável e tipo de codificação (se a variável deve ser considerada na depuração ou não).

As regras de conflito, ou regras Y, que especificam as condições inaceitáveis para as variáveis, e portanto, a serem detectadas como erradas têm o seguinte formato: Y(número): (Expressão Lógica).

As regras de imputação determinística, ou regras X, que determinam, através da especificação de uma situação inaceitável, a imputação a ser realizada para estes casos, têm o seguinte formato: X(número): (Expressão Lógica) DO (Atribuição).

Os parâmetros que especificam as funções a serem executadas pelo programa gerado podem ser:

CVAL — verifica se as variáveis têm valores válidos com os especificados no Dicionário de Dados;

RELA — verificação do cumprimento das regras de conflito;

ACOR — imputação automática;

ESTAT — elaboração das estatísticas de erros;

ELIST — listagem dos registros errados com informações dos erros ocorridos em cada um;

SEQN — listagem dos registros errados com informações dos erros ocorridos em cada um, indicando o número de ordem do registro.

O sistema AERO possui, ainda, a opção de uso de um Analisador de Regras para verificação da consistência lógica do conjunto de regras de conflito, regras Y, especificado pelo usuário num programa AERO.

Este analisador, além de eliminar redundâncias no conjunto de regras especificadas, detecta, também, inconsistências entre elas.

6.2 — Vantagens e Desvantagens

Como vantagens e desvantagens deste sistema podem ser citados os seguintes itens:

VANTAGENS:

- a) uso de um Dicionário de Dados contendo

a descrição das variáveis com seus códigos válidos, armazenados em máquina, permitindo que as variáveis possam ser usadas em arquivos diferentes, sem necessidade de redefini-las;

b) uso de imputação automática em dois modelos — o determinístico e o probabilístico;

c) conjunto de comandos simples a serem usados na especificação do usuário;

d) analisador de Regras para consistência das regras de conflito, especificadas pelo usuário.

DESVANTAGENS:

a) restrições do arquivo de entrada — só permite o processamento de arquivos seqüenciais, com variáveis do tipo qualitativo com formato alfanumérico ou decimal

zonado, e todos os registros de um mesmo tipo com tamanho máximo de 9.999 bytes;

b) na descrição do registro do arquivo de entrada, têm de ser especificadas todas as variáveis que compõem o registro, mesmo que estas não sejam usadas no processamento;

c) não permite o uso nem geração de arquivos paralelos ao de entrada;

d) não permite checagem automática da estrutura dos registros que compõem um questionário.

de variáveis e conjuntos de valores permitidos para cada uma das variáveis; os arquivos e campos que descrevem os arquivos que contêm os dados coletados; e as relações, atributos e valores que definirão a base de dados do RAPID;

b) especificação e análise das regras de crítica — nesta atividade são especificadas as regras de crítica a serem aplicadas aos dados, que são expressas sob forma de expressões lógicas que especificam combinações de valores de variáveis não permitidas.

Uma vez especificadas estas regras, o sistema as analisa fazendo diagnósticos de erros de consistência encontrados entre elas e gerando o conjunto de regras derivadas a partir das regras especificadas.

A fase de produção é, também, composta de duas atividades:

a) carregamento da base de dados — esta etapa consiste da conversão e armazenamento dos arquivos que contêm os dados coletados na base de dados do RAPID;

b) execução da crítica e correção dos dados — O CAN-EDIT, a partir do conjunto de regras especificadas pelo usuário e das regras derivadas por ele, gera um programa de crítica e correção das variáveis detectadas como erradas.

O processo de imputação automática usado pelo CAN-EDIT usa a filosofia de matrizes *hot-deck*, e todas as correções efetuadas são determinadas pelo sistema, a partir das regras de crítica especificadas, sem que haja especificação de comandos do usuário. Assim, as fases de pré-crítica e crítica são executadas simultaneamente.

7 — CAN-EDIT:

O sistema CAN-EDIT foi desenvolvido pelo centro de estatísticas do Canadá, para aplicação no Censo Demográfico de 1976, com o objetivo de detecção de erros e correção automática dos dados coletados.

Este sistema é voltado totalmente para a análise e correção de dados qualitativos.

Para a operação deste sistema é necessário que os dados sejam armazenados no RAPID e o Dicionário de Dados seja definido pelo MDM (*Meta Data Memory*) do RAPID.

O sistema CAN-EDIT é operado em duas fases chamadas de Preparação e Produção.

A fase de Preparação é composta de duas atividades:

a) definição dos dados — onde é feita a criação do Dicionário de Dados em termos

7.1 — Vantagens e Desvantagens

Como vantagens e desvantagens deste sistema podemos citar os seguintes itens:

VANTAGENS:

a) analisador das regras de crítica especificadas;

b) geração das regras de imputação para correção do arquivo, a partir das regras de crítica especificadas.

DESVANTAGENS:

a) necessidade do RAPID para uso do sistema;

b) só trabalha com dados qualitativos.

8 — OSIRIS IV

O OSIRIS IV (70), *Organized Set of Integrated Routines for Investigation of Social Science* é produto de um processo contínuo de desenvolvimento e refinamento, iniciado no *University of Michigan Computer Center*, no final da década de 50.

O sistema é formado por um conjunto integrado de rotinas para fazer Análise, Correção e Imputação de Dados, que permite o processamento de modo interativo, incluindo, também, uma variedade ampla de funções estatísticas.

8.1 — Tipos de arquivos processados:

Dois tipos, retangulares (um único tipo de registro) ou estruturados (vários tipos de registros relacionados hierarquicamente, ex.: domicílio, pessoa) e criados pelo OSIRIS IV via um comando especial (&BUILD) a partir de arquivos retangulares.

Estes arquivos são descritos para o OSIRIS IV utilizando-se um Dicionário de Dados, na própria linguagem, que contém a descrição da estrutura, caso exista, dos registros e das variáveis, incluindo informações para o tratamento da falta de informações ou especificação incorreta.

8.2 — Tipos de comandos:

- a) Comandos para monitoração (provêm documentação e funções de controle);
- b) Comandos de Programa:
 - de entrada/saída;
 - de seleção;
 - aritméticos;
 - de impressão;
 - para realização de funções específicas (ex.: RECODE, & WCC).

8.3 — Capacidade de processamento

- a) Checagem de consistência de valores — verifica relação de consistência entre os valores de um grupo de variáveis, de acordo com as especificações do usuário;
- b) Checagem de organização do arquivo de entrada — detecta e corrige erros no arqui-

vo de entrada, tais como: duplicação de registro, registro ausente ou inválido, permitindo obter um arquivo em bom estado;

c) Recodificação de variáveis — atribui novos valores a variáveis, através de comandos de atribuição, combinação de valores de duas ou mais variáveis, distribuição uniforme de valores aleatórios;

d) Checagem de consistência de variáveis — verifica se os valores das variáveis são consistentes com as especificações do dicionário e documenta tanto os valores errados quanto os ilegítimos.

8.4 — Correção automática

Permite que, além da identificação de um erro numa variável, esta possa ser corrigida de acordo com as especificações do usuário, através de atribuição de valores.

8.5 — Funções estatísticas

Várias funções estatísticas estão incorporadas ao sistema, tipo análise de correlação, regressão múltipla, análise de variância, multivariável ou não, factor análise e clusterização.

8.6 — Funções para geração de arquivo

Funções que permitem a Ordenação (SORT) e a Junção (MERGE) de arquivos.

8.7 — Vantagens e Desvantagens

VANTAGENS:

- a) pode funcionar de modo interativo;
- b) permite fazer análise de consistência das regras de crítica;
- c) permite fazer análise de consistência das variáveis especificadas no dicionário;
- d) possui uma ampla gama de funções estatísticas.

DESVANTAGENS:

- a) só permite arquivos sequenciais;
- b) não é recomendável para processamento de grandes volumes de dados.

9 — P-STAT

O P-STAT (78) é voltado, especificamente, para a crítica e correção de dados,

incluindo facilidades para manipulação de arquivos e de processamento estatístico, tais como: correlação, regressão, factor análise e tabulações cruzadas.

Sua principal aplicação é na área das pesquisas sociais (demografia, educação etc.).

O seu desenvolvimento teve início em 1962, em Princeton, e é totalmente desenvolvido em FORTRAN.

Tem duas modalidades de funcionamento: interativo e *batch*.

A linguagem é voltada ao usuário final, de formato livre e fácil de aprender. Porém, devido à grande variedade de comandos, não é fácil para o usuário se familiarizar, totalmente, com o sistema.

9.1 — Capacidade de processamento

a) — Geração de Arquivos — permite a geração de arquivos de formato livre, de modo interativo (só no caso de arquivos pequenos) e/ou *batch*.

b) — Checagem de Consistência de Variáveis — determina se o valor de uma variável está compatível com os limites de valores-estipulados para ela e, caso contrário, converte internamente o valor para *missing*, assim como faz validade dos campos numéricos, produzindo relatórios que contêm a identificação do registro, da variável e o valor errado.

c) — Checagem de Estrutura — verifica se todos os registros de um questionário estão presentes e reporta os erros detalhadamente. Aceita também questionários incompletos, caso o usuário assim o deseje, não sendo obrigatório prover registros *dummy* para os tipos dos registros ausentes.

d) — Processamento de Múltiplos Arquivos — mais de um arquivo pode estar ativo durante uma única execução, permitindo processá-los conjuntamente para fazer tabulação, ordenação, regressão, correlação e concatená-los horizontal ou verticalmente.

e) — Atualização de Arquivos — permite atualizar um arquivo a partir de um arquivo de transações, produzindo um novo *Master File*.

f) — Ordenação de Arquivos — permite ordenar arquivos até por 15 variáveis, ou

seqüencialmente, mantendo a ordem existente, caso tenha duas linhas com a mesma identificação, e emite relatório completo da execução do mesmo.

g) — Correção Automática — detecta erros e provê relatórios com a informação sobre os mesmos, que podem ser utilizados para indicar a correção. Correção que pode ser realizada automaticamente, via imputação de valores, após teste das condições que devem ser satisfeitas, criando um novo arquivo que pode ser utilizado para o resto do processamento.

h) — Limpeza de Arquivos — permite fazer a limpeza automática de arquivos, tratando as duplicações segundo as necessidades do usuário.

i) — Tabulação — possui uma variedade de processos de tabulação (cruzamento, freqüência, tabelas), permitindo combinações entre eles, além de ter um grande número de opções para obtenção dos correspondentes relatórios de saída e gráficos.

j) — Processamento estatístico — possui uma ampla variedade de funções estatísticas e tem interface com o SPSS.

l) — Transposição Automática de Arquivo de Entrada — tem esta facilidade que pode ser utilizada para arquivos pequenos.

9.2 — Vantagens e Desvantagens

VANTAGENS:

a) atribuição automática de valores a variáveis, dependendo dos valores de outras variáveis;

b) atribuição automática de valores a variáveis, dependendo de intervalos de valores de outras variáveis;

c) permite o processamento de arquivos paralelos, podendo ser lidos em forma de *balanced-line* com o arquivo principal;

d) emissão do relatório de crítica de forma a poder ser usado na entrada de dados para posterior atualização;

e) grande variedade de funções estatísticas.

DESVANTAGENS:

a) processamento de arquivos seqüenciais de até no máximo 5 000 variáveis, para a versão Jumbo;

b) só permite processamento de arquivos seqüenciais;

c) não permite crítica automática de con-

sistência das variáveis com suas especificações;

d) não permite especificação nem crítica automática de variáveis categorizadas.

OBS.: Notamos que a linguagem está especificamente voltada ao processamento de pequenos volumes de dados.

10 — ANÁLISE DAS FERRAMENTAS APRESENTADAS

Analisando-se cada uma das ferramentas apresentadas, constata-se sua eficiência ao atendimento dos objetivos propostos por cada uma delas.

Embora cada um destes sistemas apresentados se proponha a atender uma situação específica de processamento de dados e se promulgue para uso genérico em outros tipos de situações, isto nem sempre é garantido, devido às restrições apresentadas por cada uma destas ferramentas como: limitações para o arquivo de entrada, chaves que compõem o registro, tipos de variáveis, estrutura dos questionários etc.

Outra característica que se observa nestas ferramentas é que cada uma oferece uma ou mais funções altamente vantajosas, mas que se esbarram com outras funções do sistema a que pertencem, fazendo com que este não possa ser usado de modo genérico.

Como exemplo disto pode ser citado o fato do UNEDIT que permite uma checagem automática e correção para os casos de inconsistência dos valores das variáveis com suas especificações e checagem automática da estrutura de um questionário. No entanto, este sistema possui a função de correção automática muito limitado, consistindo, apenas, de comando de atribuição, não fazendo uso de correção automática por matrizes de imputação (*hot-deck*), como é o caso dos sistemas CONCOR, ATLAS e AERO.

Um outro aspecto, que se pode observar nestas ferramentas, é que embora construídas para serem usadas pelos usuários, e suas linguagens sejam simples, em todas elas é exigido um certo grau de conhecimento de lógica de processamento de pro-

gramas, dos usuários, o que nem sempre existe, nem deveria ser necessário. O usuário técnico estatístico, responsável pela pesquisa, normalmente faz seus planos de crítica dos formulários baseado nas variáveis que compõem estes questionários sem se preocupar se as variáveis pertencem ao mesmo registro ou não, o que, normalmente, é definido numa etapa posterior de arquitetura do sistema pela equipe de analistas que desenvolvem o projeto físico e, neste caso, para o uso destas ferramentas, não podem aproveitar diretamente as especificações das críticas especificadas pelos usuários.

Assim, podemos dizer que nenhuma destas ferramentas atende, integralmente, às características dos sistemas desenvolvidos na empresa.

Uma situação ideal seria a possibilidade de termos uma ferramenta em que pudéssemos juntar todas as grandes vantagens, oferecidas por cada uma destas ferramentas, incorporadas num só sistema, acrescida de novas características visando a atender aos objetivos propostos.

Assim, se pegarmos as características do ATLAS, sistema escolhido por sua maior abrangência de aplicação, por incorporar tanto processos de crítica como imputação e tabulação, e sua grande variedade de funções permitidas como:

- a) possibilidade de processamento de qualquer tipo de arquivo com qualquer tipo de variáveis;
- b) funções para grupamento de variáveis;
- c) uso de arquivos paralelos tanto de leitura como gravação;
- d) correção automática por atribuição ou matrizes de imputação;
- e) emissão de relatórios de crítica para correção manual e posterior atualização pelo próprio sistema;
- f) impressão de relatórios formatados pelo usuário;
- g) pesquisa automática de códigos;
- h) formatação das chaves de controle de quebra para os processos;
- i) funções de tabulação.

E adicionarmos as características principais dos outros sistemas como; a) verificação automática dos valores dos conteúdos das variáveis com suas especificações;

b) checagem automática da estrutura de um questionário; c) crítica direta de variáveis inter-registros. E se acrescentarmos novas funções que: a) permitissem que as próprias especificações dos planos de crítica do usuário pudessem ser inseridas para geração automática; b) lhe dessem uma nova forma de especificação da linguagem tornando-a mais livre sem necessidade de formulário; c) permitissem o aninhamento de suas funções; d) permitissem a especificação de novos comandos como macros, para especificação de expressões lógicas, aritméticas, e acesso a variáveis de regis-

tros diferentes; concluiríamos que teríamos uma ferramenta poderosíssima que, se não resolve totalmente, atenua em alto grau os problemas de *software* encontrados atualmente.

Seguindo esta linha de raciocínio, baseado nas ferramentas analisadas e nos problemas citados, é que foi desenvolvida a proposta do sistema CRIPTA apresentada na empresa que acreditamos que venha a contribuir, em muito, para uma melhor documentação e desenvolvimento dos diversos tipos de sistemas de apuração de dados desenvolvidos na empresa.

11 – QUADROS COMPARATIVOS ENTRE AS FERRAMENTAS APRESENTADAS

QUADRO 1
CARACTERÍSTICAS DE PROCESSAMENTO

	OBJETIVOS	ORIENTAÇÃO	TIPO DE ARQUIVO TRATADO	TIPO DE REGISTRO	TIPO DE VARIÁVEIS	FORMATO DE VARIÁVEIS
CONCOR (74)	- Consistência e Correção	Censo Demográfico	Sequencial	Fixo ($< = 999 \text{ bytes}$) 1 tipo		
UNEDIT (80)	- Crítica e Correção	Censo Demográfico	Sequencial	Fixo 9 tipos	Qualitativa Quantitativa	Zonado
ATLAS (80)	- Crítica, Imputação e Tabulação	Censos Demográfico e Econômico	Sequencial Indexado VSAM	Fixo ($< = 32767 \text{ bytes}$) Variável ($< = 32767 \text{ bytes}$) 99 tipos	Qualitativa Quantitativa	Zonado Binário Compactado Float
SISQUE (85)	- Crítica e Tabulação	Pesquisas Sócio-econômicas	Sequencial	Fixo ($< = 32767 \text{ bytes}$) 99 tipos	Quantitativa	Zonado
AERO (76)	- Crítica e Imputação	Censo Demográfico	Sequencial	Fixo ($< = 9999 \text{ bytes}$) 1 tipo	Qualitativa	Zonado
CAN-EDIT	- Crítica, Correção e Imputação	Censo Demográfico	RAPID	Formatado pelo <i>RAPID</i>	Qualitativa	
OSIRIS IV (66)	- Crítica, Imputação, Tabulação e Estatística	Pesquisas Sócio-econômicas	Sequencial	Variável ($< = 32767 \text{ bytes}$)		Zonado Binário Compactado Float
P-STAT	- Crítica, Imputação, Tabulação e Estatística	Censo Demográfico	Sequencial	Fixo		Zonado Binário
CRIPTA (proposto)	- Consistência, Crítica, Correção, Imputação e Tabulação	Censos Demográfico e Econômico	Sequencial Indexado VSAM	Fixo ($< = 32767 \text{ bytes}$) Variável ($< = 32767 \text{ bytes}$) 99 tipos	Qualitativa Quantitativa	Zonado Binário Compactado Float

QUADRO 2
CARACTERÍSTICAS DE FUNCIONAMENTO

	LINGUAGEM	VALIDAÇÃO DOS VALORES DAS VARIÁVEIS		CONSISTÊNCIA DOS QUESTIONÁRIOS	CORREÇÃO AUTOMÁTICA
		QL	QT		
CONCOR (74)	Semiformatada	Não	Sim	Permite	Comandos atribuição Matriz imputação
UNEDIT (80)	Formatada	Sim	Sim	Permite	Comandos atribuição
ATLAS (80)	Formatada	Sim	Não	Não permite	Comandos atribuição Matriz imputação
SISQUE (85)	Formatada	Não	Não	Permite	Comandos atribuição
AERO (76)	Livre	Sim	Não	Não permite	Comandos atribuição Matriz imputação Imputação probabilística
CAN-EDIT	Formatada	Sim		Permite	Imputação
OSIRIS IV (66)	Semiformatada	Sim	Sim	Permite	Comandos atribuição
P-STAT	Livre				
CRIPTA (proposto)	Livre	Sim	Sim	Permite	Comandos atribuição Matriz imputação

QUADRO 3
FUNÇÕES PERMITIDAS

	NÍVEL DE QUEBRA (processamento inter-registros)	AGREGAÇÃO DE VALORES	GRAVAÇÃO DE ARQUIVOS DERIVADOS	EMIÇÃO DE RELATÓRIO PARA CORREÇÃO E ATUALIZAÇÃO	RELATÓRIOS FORMATADOS PELO USUÁRIO
CONCOR (74)	2 níveis	Não permite	Permite	Não permite	Permite
UNEDIT (80)	3 níveis	Não permite	Não permite	Não permite	Não permite
ATLAS (80)	5 níveis	Permite	Permite	Permite	Permite
SISQUE (85)	Não faz	Permite	Permite	Não permite	Permite
AERO (76)	Não faz	Não permite	Não permite	Não permite	Não permite
CAN-EDIT		Não permite	Permite	Permite	Não permite
OSIRIS IV	Faz	Permite	Permite	Permite	Permite
P-STAT		Não permite	Permite	Permite	Permite
CRIPTA (proposto)	10 níveis	Permite	Permite	Permite	Permite

BIBLIOGRAFIA

- AERO: Systems philosophy. Versão 2. Budapeste, Hungarian Central Statistical Office, 1981.
- ATLAS: Sistema para a geração de programas de crítica, imputação e tabulação; Manual do usuário. Versão 2. Rio de Janeiro, IBGE. Diretoria de Informática, 1982, 89 p.
- CONCOR data editing package; user's guide. Washington, D. C., Bureau of the Census. 1986.
- GRAVES, Ronald B. CAN-EDIT. a specification driven editing system for the canadian census of population and housing. S. L., *Statistics Canada*, l.d.
- P-STAT: Manual do usuário
- SISQUE: sistema para consistência e preparação de dados; manual do usuário. Versão 1. Viçosa, Universidade Federal de Viçosa, 1985.
- UNEDIT: a generalized editing software package for census and survey data; user's guide. Versão 3. New York, Nações Unidas. 1980.

RESUMO

A partir da análise dos problemas existentes no desenvolvimento de sistemas de apuração de dados, surge a necessidade de uma ferramenta que permita a formalização, automação e integração das fases do desenvolvimento.

Assim, neste trabalho, fazemos um estudo comparativo entre algumas das ferramentas do tipo existente para *mainframe*.

As ferramentas analisadas são: CONCOR, UNEDIT, ATLAS, SISQUE, AERO, CAN-EDIT, OSIRIS-IV e P-STAT, utilizando-se como objetos de comparação as características de especificação, de processamento e recursos oferecidos, mapeados, no final do trabalho, em três tabelas comparativas.

ABSTRACT

Having in mind the existent problems in the development of Statistical Data Collected Systems, it is verified the necessity of a formal tool that permits automation and integration of the systems development phases.

Thus, this paper presents a comparative study among some tools of this sort, available to mainframe. The analysed tools are CONCOR, UNEDIT, ATLAS, SISQUE, AERO, CAN-EDIT, OSIRIS IV and P-STAT.

As comparative objects to this study we have considered specification, processing and resources characteristics offered by each one.

Those ones are represented at the end of the paper by three comparative tables.

CRÍTICA DE RAZÕES NO CENSO INDUSTRIAL DE 1985*

José Carlos C. Pinheiro**
Renato Martins Assunção**

1 — INTRODUÇÃO

Os Censos Econômicos de 1985, realizados pela Fundação Instituto Brasileiro de Geografia e Estatística — IBGE, têm o objetivo de investigar empresas e atividades industriais, comerciais, de serviços, de construção e de transportes, no tocante a sua atividade econômica.

As unidades básicas de levantamento são empresas e estabelecimentos, cuja definição é dada a seguir:

Empresa é a unidade jurídica caracterizada por uma firma ou razão social que engloba o conjunto de atividades econômicas exercidas em uma ou mais Unidade Local — UL e responde pelo capital investido nestas atividades.

Estabelecimento é uma partição da UL para fins de levantamento estatístico, podendo, em certos casos, ser a própria unidade local. Caracteriza-se por ser uma unidade de operação, localizada em área contínua, per-

tencente a uma única empresa onde se desenvolve basicamente um tipo de atividade econômica.

Na fase de coleta de dados foram utilizados diferentes modelos de questionário, de acordo com as características da atividade sob investigação e da empresa como um todo.

Os questionários podem ser classificados em dois grandes grupos:

1.1 — Questionários de empresa

— Modelo CE-001: aplicado na sede das empresas de médio e grande portes, contendo apenas dados sobre a empresa como um todo. É respondido pelas empresas que possuem faturamento igual ou superior a Cr\$ 245.000.000,00 durante 1985, ou tenham mais de uma UL e que estejam registradas no CGC;

— Modelo CE-002: versão simplificada, respondida pelas empresas de pequeno porte. Além de coletar dados da empresa, levanta informações das atividades nela desenvolvi-

* Recebido em 18 de agosto de 1988.

** Os autores são professores da Escola Nacional de Ciências Estatísticas da Fundação Instituto Brasileiro de Geografia e Estatística — IBGE.
R. bras. Estat., Rio de Janeiro, 49(191):101-118, jan./jun. 1988.

da. É o único questionário respondido por estas empresas.

1.2 — Questionários de atividades

— Modelos CE-301, CE-401 e CE-501: aplicados aos estabelecimentos industriais, comerciais e de serviços, respectivamente, pertencentes às empresas que respondem ao CE-001;

— Modelos CE-601 e CE-701: aplicados aos departamentos de construção e transportes, respectivamente, das empresas cuja atividade principal é de natureza industrial, comercial ou de serviços.

Antes de serem tabulados, os dados levantados pela rede de coleta devem ser criticados, de modo a identificar e, quando possível, eliminar incorreções que possam vir a comprometer a validade dos resultados finais. Em face das diferenças existentes entre os questionários, sistemas de crítica específicos foram elaborados para cada um deles. O modelo CE-301, que é o único no qual nos ocuparemos neste trabalho, passa por quatro fases de crítica, onde são identificados desde erros introduzidos na fase de transcrição até inconsistências existentes entre os dados dos estabelecimentos e os da sede da empresa.

Este trabalho insere-se na terceira fase de crítica do modelo CE-301 e objetiva identificar questionários com valores aberrantes em um ou mais de um conjunto de razões predefinidas. O estabelecimento de intervalos de aceitação para cada um das razões, levando-se em conta a disponibilidade existente de recursos e a importância relativa das razões, consiste no propósito central do estudo.

2 — DEFINIÇÃO DAS RAZÕES

As variáveis de interesse neste estudo dividem-se em dois grupos: razões de batimento e razões-índices.

Existem três razões de batimento, a primeira referente ao consumo de matéria-

-prima, a segunda ao consumo de combustível e a terceira ao consumo de peças e acessórios. Aquela referente à matéria-prima é definida como:

$$RMP = \frac{A}{B + C + D - E - F}$$

onde A = Consumo anual de MP em 85;
B = Compras anuais de MP em 85;
C = Recebimentos de MP durante 1985;
D = Estoque de MP em 31/12/84;
E = Estoque de MP em 31/12/84; e
F = Transferências¹ efetuadas durante o ano de 1985.

Os valores (em milhares de cruzeiros) das variáveis A,...,F constam do questionário de cada estabelecimento industrial.

O nome "razão de batimento" baseia-se no fato de que a equação abaixo deveria verificar-se:

$$A = B + C + D - E - F$$

isto é, os dois lados da equação deveriam "bater".

As definições das razões de batimento referentes a combustível — RC, e peças — RP, são análogas a esta onde agora as variáveis A,...,E dizem respeito ao consumo, compras, etc., de combustível e peças.

Por uma série de motivos, as razões de batimento nem sempre são exatamente iguais à unidade. Entre estes motivos encontram-se os erros de preenchimento, inconsistência dos dados fornecidos pelo estabelecimento, erros de digitação não detectados na fase de crítica I, inflação, etc.

Com relação a estas razões de batimento, o objetivo é identificar e rejeitar os questionários que têm pelo menos uma delas "muito grande" ou "muito pequena" em relação a 1. Para isto, precisamos definir o que significa ser "grande" ou "pequena" e com isto estabelecer uma faixa de variação tolerável em torno de 1 para estas razões.

Observe, por exemplo, que, como

$$\frac{A - (B + C + D - E - F)}{B + C + D - E - F} = |RMP - 1| \text{ e}$$

$$\frac{|A - (B + C + D - E - F)|}{A} = \left| 1 - \frac{1}{RMP} \right|,$$

¹Recebimentos e transferências referem-se à movimentação de material entre estabelecimentos de uma mesma empresa.

então RMP é função da diferença (relativa) entre A e $B + C + D - E - F$. Ao adotarmos a faixa de variação tolerável para RMP estaremos assegurando que os dados dos questionários não rejeitados tenham uma boa qualidade no sentido de apresentarem os erros relativos $|A - (B + C + D - E - F)|/A$ e $|A - (B + C + D - E - F)|/(B + C + D - E - F)$ pequenos. Entretanto, um erro relativo

$$\frac{|A - (B + C + D - E - F)|}{A}, \text{ digamos, de } 0,1,$$

quando temos valores pequenos de A , tem conseqüências muito diversas de quando temos valores grandes de A . Isto significa que a faixa de tolerância para RMP deve variar de acordo com o tamanho das variáveis A e $B + C + D - E - F$. Estas mesmas observações valem para as outras razões de batimento RC e RP. Resolver o problema exposto é o objetivo da metodologia apresentada nas Seções III e IV.

Quanto às razões-índices, elas se referem a quatro características de cada estabelecimento:

- tempo (em meses) que o estabelecimento gasta para renovar o estoque de matéria-prima (Giro de Matéria-Prima denotada GMP);
- tempo (em meses) que o estabelecimento gasta para renovar o estoque de produtos acabados e produtos de fabricação própria que estão estocados (Giro de Produtos Acabados denotada GPA);
- tempo (em meses) que o estabelecimento gasta para renovar o estoque de produtos em curso, produtos de fabricação própria que ainda não estão totalmente manufaturados (Giro de Produtos em Curso denotada GPC);
- Margem de Comércio definida como a proporção da receita com vendas de mercadorias em relação ao valor das compras destas mercadorias não são fabricadas pelo estabelecimento).

A variável Giro de Matéria-Prima é definida como:

$$GMP = \frac{(A + B)/2}{C/12},$$

onde A = Estoque de MP em 31/12/84;
 B = Estoque de MP em 31/12/85; e
 C = Valor do consumo (anual) de MP.

Assim, $\frac{A+B}{2}$ é um índice do valor do estoque de MP, $\frac{C}{12}$ é um índice do valor

do consumo mensal de MP e portanto GMP é um índice do tempo (em meses) gasto para renovar estoque de MP.

As variáveis A, B, C constam do questionário do Censo Econômico (CE-301).

De maneira análoga definem-se duas outras razões-índices. A variável Giro de Produtos Acabados é:

$$GPA = \frac{(A + B)/2}{C/12},$$

onde A = Estoque de produtos acabados em 31/12/84;

B = Estoque de produtos acabados em 31/12/85; e

C = Valor total da produção no ano.

As variáveis A, B, C constam do questionário do Censo Econômico.

A variável Giro de Produtos em Curso é definida como:

$$GPC = \frac{(A + B)/2}{C/12},$$

onde A = Estoque de produtos em curso em 31/12/84;

B = Estoque de produtos em cursos em 31/12/85; e

C = Valor total de produção no ano de 1985.

As variáveis A, B, C constam do questionário do Censo Econômico.

A variável Margem de Comércio é definida como:

$$MC = \frac{A}{B + C + D - E - F},$$

onde A = Vendas de mercadorias para revenda durante 1985;

B = Compras de mercadorias para revenda durante 1985;

C = Recebimento de mercadorias para revenda durante 1985;

D = Estoque de mercadorias para revenda em 31/12/84;

E = Estoque de mercadorias para revenda em 31/12/85; e

F = Transferências efetuadas de mercadorias para revenda durante 1985.

Diferente das razões de batimento, as razões-índices não possuem um valor típico

em torno do qual deveriam oscilar. Entretanto, espera-se que cada uma destas razões-índices tenha uma distribuição de probabilidade bem definida e, portanto, de posse desta distribuição, poderíamos definir um limite para os valores toleráveis das razões-índices. Assim, o objetivo seria determinar este limite de maneira que sejam rejeitados todos os questionários com razões-índices acima deste limite de tolerância.

2.1 — Correção do efeito da inflação

As variáveis, que entram na definição das razões, são afetadas pela inflação ao longo do ano e a correção desta influência deveria ser feita expressando todas as variáveis envolvidas em unidades monetárias de um mesmo mês.

Assim, escolhendo-se o mês de janeiro como o mês de referência, todas as variáveis referentes a estoque em 31/12/85 seriam deflacionadas utilizando-se um índice de inflação anual. Para as outras variáveis, referentes a valores agregados (tais como: consumo anual de matéria-prima, compras anuais de matéria-prima, valor total da produção no ano, etc.) seria adotado o seguinte procedimento: o valor agregado total V corresponde a uma quantidade total

Q de bens, onde $Q = \sum_{i=1}^{12} q_i$ e q_i = quan-

tidade de bens correspondente ao mês i . Então o valor agregado total V satisfaz

$V = \sum_{i=1}^{12} V_i$, onde V_i = valor agregado cor-

respondente aos q_i produtos do mês i . Se $q_i = q_1 V_i$, então $V_i = I_i V_1$, onde I_i = índice da inflação acumulada no período de janeiro ao mês i e portanto:

$$V = \sum_{i=1}^{12} V_i I_i \quad V_1 = \frac{V}{\sum_{i=1}^{12} I_i}$$

Denote $\sum_{i=1}^{12} I_i$ por I

O valor agregado anual corrigido pelo efeito de inflação seria então:

$$V^* = 12V_1 = \frac{12}{I} V$$

Deste modo, as razões de batimento corrigindo-se os efeitos de inflação seriam da forma:

$$R = \frac{\frac{12A}{I}}{\frac{12B}{I} + \frac{12C}{I} + D - \frac{12F}{I} - \frac{E}{I_{12}}}$$

e as razões-índices seriam das formas:

$$R = \frac{\frac{1}{2} (A + \frac{B}{I_{12}})}{\frac{C}{I}} \quad (\text{para os giros de estoque})$$

$$R = \frac{\frac{12A}{I}}{\frac{12B}{I} + \frac{12C}{I} + D - \frac{12F}{I} - \frac{E}{I_{12}}}$$

(para a margem de comércio), onde $A, B, \dots, F$ são definidas da mesma maneira que antes em cada razão.

Esta correção da influência da inflação não foi implementada no programa de crítica, entre outros motivos, porque o efeito desta correção realizada nos dados do Censo de 1980 se mostrou pequeno. Isto não é de se estranhar: para o ano de 1980, $I_{12} =$

$= 1,9532$ e $\sum_{j=1}^{12} I_j = 17,4004$ (usando o INPC) e portanto as razões-índices referentes a giro de estoque corrigidas são da forma:

$$GMP = \frac{\frac{1}{2} (A + \frac{B}{1,9532})}{\frac{C}{17,4004}}$$

Como a razão não-corrigida é da forma

$$\frac{\frac{1}{2} (A + B)}{\frac{C}{17,4004}}, \text{ então o quociente destas razões}$$

(corrigida e não-corrigida) é:

$$q = \frac{\frac{\frac{1}{2} (A + \frac{B}{1,9532})}{\frac{C}{17,4004}}}{\frac{\frac{1}{2} (A + B)}{\frac{C}{12}}} =$$

$$= \frac{(A + \frac{B}{1,9532}) (17,4004)}{(A + B) 12}$$

Como $B \approx (1,9532)A$ então $q \approx 0,98$. O mesmo cálculo para a inflação de 1985 mostrou que $q \approx 0,90$. A superestimação que a razão não-corrigida implica, foi considerada pequena e acrescentando-se outros motivos operacionais foi decidido não corrigir o efeito da inflação.

2.2 — Transformação das razões de batimento

As razões de batimento definidas anteriormente foram transformadas obtendo-se um novo conjunto de variáveis. Como será explicado abaixo, esta transformação teve basicamente o objetivo de eliminar a assimetria que teriam os limites de aceitação caso fossem utilizadas as variáveis iniciais.

Para estabelecer uma regra de crítica para as razões de batimento

$$R = \frac{A}{B+C+D-E-F} = \frac{A}{V} \text{ devemos obter}$$

uma faixa de aceitação tolerável para estas razões em torno do número 1. Temos que se $A < V$ então $R \in (0, 1)$ e se $A > V$ então $R \in (1, \infty)$ (excluímos os casos $A = 0$ e $V = 0$). Deste modo, a menos que $P(A < V) \gg P(V < A)$, a faixa de aceitação com limites L_1 e L_2 (com $L_1 < 1 < L_2$) deve ter $1 - L_1 < L_2 - 1$.

Para eliminar esta assimetria tomamos os valores absolutos dos logaritmos das razões de batimento, definimos:

$$R^* = |\log R| = \left| \log \frac{A}{V} \right| = |\log A - \log V|,$$

e esta nova variável é simétrica em relação à A e V , isto é, R^* é invariante para as duas

possíveis definições de R , $R = \frac{A}{V}$ ou $R = \frac{V}{A}$. Além disso, ainda reflete uma variação

relativa entre os valores A e V .

A nova variável R^* pertence ao intervalo $[0, \infty]$ e o valor "correto" desta variável é 0. Nosso objetivo passa a ser, buscar um limite crítico c para esta variável, de modo

que se um questionário tem $R^* > c$ então ele será rejeitado.

2.3 — Correção da informação referente a estoque

É muito natural esperar que as variáveis A e B sejam fortemente relacionadas, onde:

A = Estoque de matéria-prima em 21/12/84; e

B = Estoque de matéria-prima em 31/12/85.

Pode-se mesmo supor, para um estabelecimento fixo, que $B \approx I.A$, onde I = índice de inflação durante 1985, e, deste modo, esperava-se que $\log E$ deveria ser relacionado a $\log D$ através do modelo linear.

$$\log B = \log I + \log A + \varepsilon.$$

Deve ser notado que a inclinação da reta de regressão é 1 e que o intercepto é o logaritmo do índice de inflação no ano.

O Gráfico 1 do Apêndice representa os valores dos pares $(\log B, \log A)$ obtidos dos dados de 8 500 estabelecimentos industriais pesquisados no Censo Industrial de 1980.

O índice de inflação para o ano de 1980 foi 1,9532 e portanto $\log 1,9532 = 0,67$. Isto e mais a inclinação da "nuvem" de pontos mostram que nossa hipótese é bem plausível, a menos dos pontos que formam um "L" a partir da origem.

Verificou-se que este comportamento anormal detectado no gráfico explicava-se pelo início ou encerramento das atividades do estabelecimento durante o ano. São estes estabelecimentos que foram o "L" no Gráfico 1. Para tomar os logaritmos de B e A , somou-se 1 em cada variável, de modo a evitar o caso $\log(0)$.

Para estes estabelecimentos, o índice de giro de estoque de matéria-prima

$$(GMP = \frac{1}{2} \frac{A + B/I_{12}}{\sum I_i}) \text{ não daria um indi-}$$

cador correto do tempo de renovação de estoques.

Deste modo, uma correção da informação referente ao estoque de matéria-prima se fez necessária. Então, para estes estabele-

²Ver "Correção da informação referente a estoque", à frente.

cimentos que tenham $A = 0$ ou $B = 0$ esta correção consistiu em redefinir GMP como

$$\text{GMP} = \frac{\frac{1}{2} \max(A, \frac{B}{I_{12}})}{\frac{C}{I_i + \dots + I_j}}$$

onde A , B , C e I_{12} têm o mesmo significado que antes, i = número do mês de início das atividades e j = número do mês de encerramento das atividades. Estas últimas informações adicionais também constam do questionário do Censo Econômico e como o número de estabelecimentos que necessitam ser corrigidos é relativamente pequeno, a correção não terá custo muito elevado.

Uma correção análoga deve ser feita nas razões-índices GPA e GPC. Deve ser notado que as razões de batimento e margem de comércio não necessitam desta correção.

2.4 — Critério de rejeição

Para a determinação dos limites críticos das regiões de variação tolerável para as razões devemos considerar dois aspectos que atuam em sentidos contrários. Se quisermos obter, ao final do processo de crítica, um alto nível de qualidade para os dados, seremos obrigados a tomar regiões de variação muito estreitas, o que levará a uma grande proporção de questionários rejeitados. Entretanto, o tempo e o alto custo envolvidos na revisão destes questionários rejeitados atuam no sentido de restringir o nível de qualidade final desejado. Assim, é preciso decidir sobre o balanceamento entre o nível de qualidade final e os recursos disponíveis. Esta decisão é de natureza gerencial e a direção do Censo Econômico adotou, em face das limitações existentes, o nível de 5% do total de questionários como um limite máximo para a proporção de questionários a serem rejeitados no processo de crítica.

Como um questionário será rejeitado, se pelo menos uma das sete razões consideradas for rejeitada, então a limitação acima pode ser escrita como:

$$0,05 = P(\text{rejeitar um questionário}) =$$

$$= P\left(\bigcup_{i=1}^n A_i\right), \text{ onde } A_i = [\text{rejeitar a } i\text{-ésima}$$

razão] isto é,

$$0,05 = 1 - P\left(\bigcap_{i=1}^n A_i^c\right) = 1 - P(\text{todas as razões "corretas"}) (*)$$

Num estudo realizado com os dados dos Censos Econômicos de 1980, verificou-se que hipótese de independência entre as razões era bastante aceitável. Adotando-se esta hipótese, (*) torna-se:

$$0,05 = 1 - \left(\prod_{i=1}^n (1 - P(A_i))\right) (**)$$

Assim, os limites críticos para cada uma das sete razões (que determinam $P(A_i)$) devem satisfazer a restrição (**).

A determinação das probabilidades $P(A_i)$ foi baseada na importância relativa entre as razões. Isto significa atribuir pesos aos níveis de rejeição $P(A_i)$: quanto maior o peso, maior $P(A_i)$, o que implica num nível de qualidade maior para a razão i . A direção do Censo Econômico decidiu que as razões referentes à matéria-prima, combustível, giro de produtos acabados e giro de produtos em cursos teriam um percentual de rejeição cinco vezes maior que as razões referentes a giro de matéria-prima lucro de venda e peças. Deste modo: (**) torna-se,

$$0,05 = 1 - (1 - 5\theta)^4 (1 - \theta)^3,$$

onde θ é o nível percentual de rejeição para GMP, MC e RP. Resolveu-se esta equação utilizando o método de Newton, encontrando $\theta = 0,0022 \Rightarrow 5\theta = 0,011$.

3 — CURVAS DE REJEIÇÃO

Quando são empregadas razões entre variáveis para fins de crítica em um questionário, um aspecto fundamental da questão permanece usualmente encoberto: o da magnitude dos valores envolvidos nos cálculos. Tome-se, por exemplo, uma razão de batimento (ou seja, a resultante da divisão entre variáveis que deveriam apresentar o mesmo valor no questionário). Obviamente o valor correto deveria igualar a unidade, sendo necessário, para a crítica, estabelecer uma faixa de variação tolerável em torno deste número. Ocorre que, se para pequenos valores um erro relativo de até 10% pode ser considerado aceitável (em termos

de sua influência sobre o agregado das informações, por exemplo), para valores elevados este mesmo percentual pode revelar-se desastroso.

Uma maneira de contornar esta dificuldade é através da combinação da informação associada ao erro relativo (representada, no exemplo anterior, pela razão) com alguma medida correlacionada com a magnitude dos valores utilizados no cálculo. Nesta seção descreve-se uma abordagem metodológica de crítica que preenche estes requisitos (combinação dos erros relativos e absolutos), desenvolvida para a fase de crítica II do Questionário 301 do CE-85. Saliente-se que a metodologia em questão não tem sua utilização limitada a razões de batimento, tendo sido empregada também para razões de outros tipos, como giros de estoque.

3.1 — Motivação do estudo

Dois aspectos necessitam ser considerados quando da elaboração de um plano de crítica: o nível final de qualidade desejado para os dados e a disponibilidade de recursos. A atuação dos dois dá-se em sentido contrário, sendo de natureza mais gerencial do que técnica a decisão sobre o balanceamento entre ambos.

A motivação por trás do desenvolvimento da metodologia aqui descrita foi a de atingir o maior nível de qualidade possível para os dados dentro da limitação de recursos existentes. A idéia básica é de que as informações possuem importância diferenciada, em termos de influência sobre a qualidade final dos cálculos. Assim os recursos devem ser prioritariamente destinados à identificação de erros associados a valores de maior peso na composição dos agregados.

No caso particular da crítica de razões, essa linha de raciocínio conduz naturalmente ao estabelecimento de critérios diferenciados de rejeição, que levem em conta a grandeza dos valores utilizados nos cálculos. Essencialmente, o que se pretende é maior flexibilidade com razões associadas a valores pouco significativos e maior rigidez com aquelas que envolvem valores expressivos.

3.2 — Metodologia

A princípio, considerou-se a alternativa de estratificação dos questionários segundo alguma medida correlacionada com as variáveis envolvidas nos cálculos das razões de interesse, seguida da especificação de valores críticos diferenciados por estrato. Estes últimos seriam estabelecidos de tal forma que a probabilidade de rejeição crescesse com a média dos valores assumidos pela medida de estratificação no estrato.

Considere-se, por exemplo, a razão de batimento associada a consumo e gasto com matéria-prima. Assuma-se que os questionários tenham sido divididos em três categorias: pequenos (60%), médios (30%) e grandes (10%) consumidores de matéria-prima e que seja tolerada a rejeição de até 1% dos questionários devido a esta razão. Uma questão fundamental para a aplicação desta abordagem é a definição da importância relativa entre os estratos, para fins de crítica. Uma maneira bastante natural de resolver este problema é através da atribuição de pesos relativos aos percentuais de rejeição por estrato. Suponha que, no exemplo em questão, tenham sido atribuídos pesos um, três e cinco, respectivamente, aos estratos dos pequenos, médios e grandes consumidores de matéria-prima. Isto significa que seriam rejeitados, em termos percentuais, três vezes mais questionários no segundo e cinco vezes mais no último estrato que no primeiro. Neste exemplo particular, os percentuais de rejeição seriam: 0,5% no primeiro, 1,5% no segundo e 2,5% no último estrato. A regra de crítica ficaria estabelecida por completo com a determinação dos percentis das distribuições das razões nos estratos, associados aos percentuais de rejeição previamente mencionados.

A abordagem de crítica por estratos encerra, contudo, alguns problemas operacionais que podem dificultar sua utilização prática. O principal deles talvez seja o da construção dos estratos de questionários, consistindo-se em tarefa especialmente delicada, a especificação da quantidade e limites dos mesmos. A metodologia desenvolvida neste estudo procura aproveitar a idéia central da abordagem por estratos, contor-

nando, dentro do possível, as desvantagens a ela associadas.

3.3 — Determinação teórica da curva de rejeição

Em certo sentido a metodologia de crítica por curvas de rejeição pode ser entendida como uma extensão da abordagem por estratos, com a importante diferença de que os questionários não necessitam ser estratificados, sendo considerados individualmente para fins do estabelecimento da regra de crítica.

O método pode ser resumidamente descrito da seguinte forma: a partir de uma medida de tamanho correlacionada com as variáveis que entram no cálculo das razões de interesse, determina-se uma função, denominada curva de rejeição, que fornece para cada questionário um valor crítico (note-se que a cada razão associa-se uma curva de rejeição). Traçando-se um paralelo com a abordagem anterior, os questionários poderiam ser considerados estratos unitários, adicionando-se, agora, a hipótese de existência de uma relação funcional entre os valores críticos dos estratos.

O procedimento de crítica para uma determinada razão R consiste então em:

(a) através da curva de rejeição e do valor da medida de tamanho observada no questionário, obter o valor crítico; e

(b) comparar o valor da razão no questionário com o valor crítico anteriormente obtido.

Duas funções, cuja notação é introduzida abaixo, são de fundamental importância para o desenvolvimento da metodologia crítica utilizada:

$C_R(T)$ — a curva de rejeição associada à razão R , baseada na medida de tamanho T ; e

$P_R(T)$ — a probabilidade condicional de rejeição da razão R , dado o valor T para a medida de tamanho.

Existe uma certa dualidade entre as duas no sentido de que, conhecidas as distribuições de probabilidade de R e T , uma pode ser obtida a partir da outra. Na abordagem a seguir descrita, faz-se, a princípio, a determinação da função $P_R(T)$ a partir da especificação

de um modelo e de um conjunto de restrições. A seguir, deriva-se a curva de rejeição associada, $C_R(T)$.

Pela discussão apresentada na seção anterior, a função $P_R(T)$ deve ser crescente em T , de modo a direcionar os recursos prioritariamente para os questionários de maior importância na composição dos cálculos. Além desta, duas outras restrições devem ser obedecidas pela função $P_R(T)$.

$$0 < P_R(T) < 1 \quad (1) \quad \text{e}$$

$$E(P_R(T)) = \alpha \quad (2),$$

onde $E(.)$ representa a esperança matemática e α o percentual de rejeição associado a R .

A partir destas condições, arbitrou-se o seguinte modelo para a função de probabilidade condicional:

$$P_R(T) = Ae^{BT} + C \quad (3),$$

onde A , B e C são constantes a serem determinadas.

Note-se que, pelas restrições mencionadas, as seguintes desigualdades devem ser satisfeitas:

$$A, B < 0 \quad (4);$$

$$1 > C > 0 \quad (5);$$

$$|A| < C \quad (6)$$

A obtenção das constantes A , B e C é feita a partir de um conjunto de equações derivadas das restrições citadas anteriormente:

$$P_R(t_0) = Ae^{Bt_0} + C = \delta_0\alpha \quad (7);$$

$$P_R(t_1) = Ae^{Bt_1} + C = \delta_1\alpha \quad (8);$$

$$E(P_R(T)) = AEe^{BT} + C = A\psi_T(B) + C = \alpha \quad (9);$$

onde, como antes, α representa o percentual de rejeição associado a R ; t_0 e t_1 são dois valores preestabelecidos da medida T ; δ_0 e δ_1 os pesos relativos de rejeição arbitrados para questionários de tamanhos t_0 e t_1 e ψ_T representa a função geradora de momentos de T .

Uma solução trivial para o sistema de equações acima, quando $\delta_0 = \delta_1 = 1$, é dada por $B = 0$ e $A + C = \alpha$. Como ficará claro no decorrer do texto, isso significa que a curva de rejeição assume um valor constan-

te³, igual ao percentil de ordem $1 - \alpha$ da distribuição de R. No presente estudo, ficar-se-á restrito aos casos em que $\delta_0 \neq \delta_1$ e $B < 0$, significando que $P_R(T)$ é estritamente crescente em T.

As Equações (7), (8) e (9) podem ser reescritas na forma:

$$A = \frac{(\delta_1 - \delta_0) \alpha}{e^{Bt_1} - e^{Bt_0}} \quad (7'); \quad (7')$$

$$C = \left[\delta_1 - \frac{(\delta_1 - \delta_0) e^{Bt_1}}{e^{Bt_1} - e^{Bt_0}} \right] \alpha \quad (8')$$

$$\frac{\psi_T(B) - e^{Bt_1}}{e^{Bt_1} - e^{Bt_0}} = \frac{1 - \delta_1}{\delta_1 - \delta_0} \quad (9')$$

Note-se que a constante B independe de α .

Uma simplificação importante é obtida tomando-se $\delta_1 = 1$, ou seja, assumindo-se que um questionário com tamanho t_1 terá a razão R rejeitada com probabilidade α . Neste caso, as equações se simplificam para:

$$A = \frac{(1 - \delta_0) \alpha}{e^{Bt_1} - e^{Bt_0}} \quad (7''); \quad (7'')$$

$$C = \frac{(\delta_0 e^{Bt_1} - e^{Bt_0}) \alpha}{e^{Bt_1} - e^{Bt_0}} \quad (8''); \quad (8'')$$

$$\psi_T(B) = e^{Bt_1} \quad (9'') \quad (9'')$$

Observe-se, agora, que B depende apenas de t_1 e da distribuição de T.

Obtidos valores para A, B e C, o problema passa a ser a determinação da curva de rejeição $C_R(T)$.

Como visto nas seções precedentes, todas as regras de crítica utilizadas no trabalho visam a identificar razões que assumem valores por demais elevados, ou seja, são da forma:

“rejeitar o questionário quando $R > C_R$ ”, onde C_R é o valor crítico associado à razão R. Na abordagem aqui apresentada, o valor crítico é dado pela função $C_R(T)$. Tem-se, então:

$$P_R(T) = P(R > C_R(T) | T) = 1 - F_R(C_R(T) | T) \quad (10), \quad (10),$$

onde $F_R(C_R(T) | T)$ é a função de distribuição de R condicionada em T.

Assumiu-se, no desenvolvimento do projeto, a hipótese de independência entre as razões de crítica⁴ e a medida de tamanho T. A análise posterior dos dados dos Censos Econômicos de 1980 indicou ser esta hipótese bastante aceitável. A Equação (10) pode ser então simplificada:

$$P_R(T) = 1 - F_R(C_R(T)) \quad (10') \quad (10')$$

Segue daí a relação fundamental para obtenção da curva de rejeição:

$$C_R(T) = F_R^{-1}(1 - P_R(T)) \quad (11), \quad (11),$$

ou seja, o valor crítico $C_R(T)$ é dado pelo percentil de ordem $1 - P_R(T)$ da distribuição de R.

3.4 — Estimação

O desconhecimento das distribuições de probabilidade das variáveis T e R torna necessário o uso de técnicas de estimação para derivação das curvas de rejeição. A distribuição de T é necessária para determinar as constantes A, B e C, que definem a função $P_R(T)$, enquanto que a de R é utilizada na obtenção da curva $C_R(T)$. A metodologia de estimação empregada é a seguir apresentada:

— estimação de $P_R(T)$.

A distribuição da variável T é fundamental para resolução da Equação⁵ (9''), que fornece o valor da constante B, da função $P_R(T)$.

$$\psi_T(B) = e^{Bt_1} \quad (9'') \quad (9'')$$

Dois procedimentos alternativos para estimação de B são apresentandos. Um primeiro, de características não-paramétricas, utiliza a expansão da função geradora de momentos em série de Taylor, e outro,

³Está implícito, nesta afirmação, que a regra de crítica visa a identificar apenas razões que excedam um determinado valor, não sendo fixado um limite inferior de variação.

⁴Esclarecimentos que esta hipótese não contradiz a idéia original da estratificação apresentada no início da Seção Metodologia, já que a medida de tamanho pode ser correlacionada com as variáveis que entram no cálculo das razões, mas não com as próprias razões.

⁵Será assumida a simplificação mencionada, considerando-se $\delta_1 = 1$.

baseia-se na especificação da distribuição de T.

3.4.1 — Abordagem não-paramétrica

A hipótese básica aqui adotada é de que a variável T possui todos os momentos finitos, podendo-se escrever:

$$\psi_T(B) = 1 + \sum_{k=1}^{+\infty} \frac{m(k)B^k}{k!} \quad (12),$$

onde $m(k)$ representa o k-ésimo momento da variável T.

O processo de estimação consiste em escolher um p suficientemente grande, tal que:

$$\psi_T(B) = 1 + \sum_{k=1}^p \frac{m(k)B^k}{k!} \quad (13)$$

e substituir os parâmetros $m(k)$ pelos momentos amostrais correspondentes, obtendo-se:

$$\hat{\psi}_T(B) = 1 + \sum_{k=1}^p \frac{\hat{m}(k)B^k}{k!} \quad (14)$$

O estimador de B seria, então, dado pela solução da equação:

$$\hat{\psi}_T(B) = e^{Bt_1} \quad (15)$$

que pode ser obtida numericamente através de métodos como o de Newton-Raphson, por exemplo.

A determinação do p adequado é feita iterativamente em conjunto com a estimação de B. Define-se, inicialmente, um valor p_0 , obtendo-se a estimativa associada $\hat{B}_0$. Toma-se, então, $p_1 = p_0 + 1$, com a respectiva estimativa $\hat{B}_1$. Prossegue-se até o passo n , onde se verifica a convergência das estimativas a um valor $\hat{B}$.

3.4.2 — Abordagem paramétrica

Neste caso, é assumida uma distribuição P_Θ para a variável T, com vetor de parâmetros Θ . Obtidas as estimativas para Θ , encontra-se a estimativa paramétrica de B a partir da equação:

$$\psi_{T, \hat{\Theta}}(B) = e^{Bt_1} \quad (16),$$

Por exemplo, assumindo-se $T \sim N(\mu, \sigma^2)$ e representando-se por $\hat{\mu}$ e $\hat{\sigma}^2$ os estimadores

da esperança e da variância, obtém-se o estimador $\hat{B}$ como solução de:

$$e^{\hat{\mu}B} + \frac{\hat{\sigma}^2 B}{2} = e^{Bt_1}$$

que é dada por:

$$\hat{B} = 2(t_1 - \hat{\mu}) / \hat{\sigma}^2$$

Nem sempre será possível derivar-se a solução fechada para a equação, tornando-se necessário o emprego de métodos numéricos de resolução. Considere-se, por exemplo, $T \sim \Gamma(\alpha, \beta)$ e sejam $\hat{\alpha}$ e $\hat{\beta}$ os estimadores associados. A Equação (17) neste caso é dada por:

$$\left(\frac{\hat{\beta}}{\hat{\beta} - B} \right)^{\hat{\alpha}} = e^{Bt_1}$$

que não possui solução fechada.

Encontrada uma estimativa $\hat{B}$ para B, segundo uma das abordagens mencionadas anteriormente, obtém-se estimativas para as constantes A e C por substituição de B por $\hat{B}$ nas Equações (7'') e (8''). A função $P_R(T)$ será, então, estimada por:

$$\hat{P}_R(T) = \hat{A}e^{\hat{B}T} + \hat{c} \quad (17)$$

3.4.3 Estimação de $C_R(T)$

A metodologia utilizada para estimação da curva $C_R(T)$ é de natureza essencialmente não-paramétrica, baseando-se em $\hat{P}_R(T)$ e na função de distribuição empírica de R, que será representada por $\hat{F}_R$.

A curva de rejeição foi definida em (11), pela relação:

$$C_R(T) = F_R^{-1}(1 - P_R(T))$$

Como já salientado anteriormente, os valores críticos $C_R(T)$ são dados pelos percentis de ordem $1 - P_R(T)$ da distribuição de R. A relação (11) pode ser alternativamente representada por:

$$F_R(C_R(T)) = 1 - P_R(T) \quad (11')$$

ou, ainda,

$$F_R(C_R(T)) = \Theta \quad (11'')$$

$$1 - P_R(T) = \Theta$$

Em todo o processo de estimação dispôs-se de uma amostra de 8 500 questionários referentes ao Censo Industrial de 1980. Na

amostra, a medida de tamanho T utilizada (no caso, o logaritmo do total de pessoal ocupado no estabelecimento) apresentava valores entre um mínimo ($t_{\min}$) e um máximo ($t_{\max}$). Assumindo-se a mesma variação para T no Censo de 1985, observa-se que só há necessidade de definir-se a curva $C_R(T)$ no intervalo $[t_{\min}, t_{\max}]$.

A partir da função empírica $\hat{F}_R$ e de $\hat{P}_R(T)$, podemos obter um conjunto de pontos $(T_p, \hat{C}_R(T_p))$ tais que para $p \in N$,

$$\begin{aligned} [n \cdot (1 - \hat{P}_R(T_{\max}))] \leq p \leq [n \cdot (1 - \hat{P}_R(T_{\min}))] \\ \hat{F}_R(\hat{C}_R(T_p)) = p/n \\ 1 - \hat{P}_R(T_p) = p/n \end{aligned} \quad (18)$$

Note-se que, pela discussão anterior, só necessitamos determinar a curva $C_R(T)$ em $[t_{\min}, t_{\max}]$, o que equivale a obter as soluções das Equações (11'') para Θ no intervalo $[1 - P_R(t_{\max}), 1 - P_R(t_{\min})]$. A limitação imposta sobre a variação de p faz com que p/n tome valores no intervalo $[1 - \hat{P}_R(t_{\max}), 1 - \hat{P}_R(t_{\min})]$. Fazendo $R_{(p)}$ representar a p -ésima estatística de ordem da variável R na amostra, obtêm-se as soluções:

$$\begin{aligned} T_p &= 1n \left((1 - p/n - \hat{C})/\hat{A} \right) / \hat{B} \\ \hat{C}_R(T_p) &= R_{(p)} \end{aligned} \quad (19)$$

Assumiu-se, então, a hipótese de que os pontos $(T_p, \hat{C}_R(T_p))$ seguiam o mesmo comportamento de $(T, C_R(T))$, a menos de perturbações aleatórias. Mais especificamente, adotou-se o modelo:

$$\hat{C}_R(T) = \beta_0 e^{\beta_1 T} \cdot \xi \quad (20),$$

ou, equivalentemente,

$$1n(\hat{C}_R(T)) = 1n(\beta_0) + \beta_1 T + \xi^* \quad (21);$$

a curva de rejeição teórica, neste modelo, é dada por:

$$C_R(T) = \beta_0 e^{\beta_1 T} \quad (22)$$

Os parâmetros β_0 e β_1 foram estimados por mínimos quadrados ordinários, no modelo linearizado, assumindo-se as hipóteses usuais "sobre os erros ξ^* ".

A curva de rejeição estimada, utilizada no processo de crítica, é dada por:

$$\hat{C}_R(T) = \hat{\beta}_0 e^{\hat{\beta}_1 T} \quad (23)$$

4 — RESTRIÇÃO DO TRABALHO

Durante a realização deste trabalho, um importante problema surgiu pela impossibilidade de acesso aos dados dos Censos Econômicos de 1985, que se encontravam ainda em fase de digitação.

A alternativa adotada foi a de utilizarem-se os dados do questionário CE-301 referentes ao Censo de 1980.

Por serem as razões utilizadas na crítica variáveis adimensionais (razões de batimento) ou não-monetárias (razões-índices), a conversão dos valores para 1985 não apresentou maiores problemas. No entanto, foi necessário assumir-se a hipótese de que as distribuições de probabilidade das razões permaneceram inalteradas de um Censo para outro. Exceção foi feita à razão de margem de comércio, que mereceu tratamento diferenciado das demais.

Efetou-se uma análise comparativa das distribuições das razões por gênero industrial e região do Brasil, não se constatando diferenças que justificassem o uso de regras diferenciadas de rejeição, por gênero ou região.

A impossibilidade de acesso aos dados de 1985 implicou também na escolha de uma medida de tamanho (T) de características não-monetárias, com vistas à estimação das curvas de rejeição definidas na seção anterior. A escolha recaiu sobre a variável "pessoal ocupado em 31/12/85", que será referenciada no restante do texto por PO. Novamente foi assumida a hipótese de igualdade entre as distribuições de probabilidade da variável para os Censos de 1980 e 1985.

Um derradeiro aspecto da pesquisa a ser ressaltado, diz respeito à mudança do âmbito de aplicação do questionário CE-301 do Censo de 1980 para o de 1985. No primeiro, o questionário era respondido por todos os estabelecimentos de indústria, enquanto que no segundo o universo de informantes restringiu-se às empresas com faturamento

igual ou superior a Cr\$ 245.000.000,00, no ano de 1985, ou que possuísem mais de uma unidade local. A compatibilização das informações foi efetuada por meio da seleção, no arquivo dos dados de 1980, das empresas que teriam respondido o questionário CE-301, caso fossem adotadas as mesmas restrições utilizadas no Censo de 1985.

5 — RESULTADOS

Nesta seção, são apresentados os principais resultados obtidos a partir da aplicação da metodologia apresentada nas seções anteriores aos dados do Censo de 1980. Todos os cálculos foram efetuados com base em uma amostra sistemática de um terço do total de questionários disponíveis. O uso da amostra deveu-se não só à necessidade de diminuir-se o volume de informações, de modo a acelerar as rotinas de cálculo, como também ao interesse de validar-se os procedimentos de crítica em um conjunto independente de questionários, no caso os dois terços restantes.

5.1 — Medida de tamanho (T)

Conforme mencionado na seção anterior, fez-se uso da variável PO (pessoal ocupado em 31/12/85) como medida de tamanho. Na verdade, a fim de reduzir-se a acentuada assimetria que caracteriza a distribuição dessa variável, tomou-se a transformação logarítmica de PO, doravante representada por LPO.

A análise do histograma da variável LPO indicou que a distribuição gama ajustava-se bem aos dados (ver Gráfico 2). Os parâmetros α e β da distribuição foram estimados pelo método de momentos, tendo-se obtido:

$$\begin{aligned}\hat{\alpha} &= 3,215 \\ \hat{\beta} &= 1,234\end{aligned}$$

Desta forma, foi assumido que LPO seguia uma distribuição $\Gamma(3,625, 1,234)$.

5.1.1 — Probabilidade condicional de rejeição ($P_R(T)$)

A determinação de $P_R(T)$ é feita através da estimação dos três parâmetros que a definem (ver Seção III). Inicialmente, procedeu-se à estimação do parâmetro B, que depende apenas da especificação da distribuição da medida de tamanho e do ponto t_1 para o qual deseja-se rejeitar a razão com probabilidade α . No caso, definiu-se este ponto como sendo a mediana de T, obtendo-se então:

$$t_1 = 2,83321$$

A constante B foi estimada a partir da equação:

$$\left(\frac{1,234}{1,234-B} \right)^{3,615} = e^{B \cdot (2,83321)}$$

tendo-se, portanto, optado pela abordagem paramétrica de estimação, apresentada na Seção III. A expressão acima é a Equação (17) daquela seção, aplicada à distribuição $\Gamma(3,613, 1,234)$ e $t_1 = 2,83321$. Resolvendo-se por Newton-Raphson, obteve-se⁶:

$$\hat{B} = -0,0846$$

Para a estimação dos demais parâmetros, torna-se necessário definir o valor t_0 , seu peso relativo de rejeição δ_0 e a probabilidade de rejeição não-condicional α . No caso, tomou-se:

$$\begin{aligned}t_0 &= \min(T) = 0 \\ \delta_0 &= 0,2\end{aligned}$$

após discussão com a gerência do Censo.

A probabilidade não-condicional de rejeição varia de razão para razão, fazendo com que os parâmetros A e C de $P_R(T)$ também variem com ela. Dois fatores respondem por esta variação, os pesos diferenciados de rejeição e os modelos assumidos para as distribuições das razões. Esta questão será abordada em detalhe no decorrer desta seção. Por ora, observe-se apenas que os parâmetros A e C podem ser expressos na forma abaixo:

$$\begin{aligned}A &= A' \cdot \alpha_R \\ C &= C' \cdot \alpha_R\end{aligned}$$

⁶ A solução não-paramétrica indicou $\hat{B} = -0,0819$, com $P = 9$, convergindo para uma tolerância de 10^{-7}

onde A' e C' são os mesmos para todas as razões⁷. Substituindo-se os valores, definidos ou estimados anteriormente nas expressões de A' e C' , obtém-se:

$$\begin{aligned}\hat{A}' &= -3,754 \\ \hat{C}' &= 3,954\end{aligned}$$

Assim sendo, a curva de rejeição estimada para uma razão R qualquer será dada por:

$$\hat{P}_R(T) = (-3,754 e^{-0,0846T} + 3,954) \cdot \alpha_R,$$

onde α_R representa o percentual de rejeição não-condicional para a razão R .

5.2 — Percentuais de rejeição por razão(α_R)

A metodologia de determinação dos percentuais de rejeição por razão foi apresentada ao final da Seção II. Faz-se menção, ali, ao uso de pesos diferenciados para os percentuais de rejeição, levando ao estabelecimento de dois grupos de razões: as muito e as menos importantes, sendo que as razões do primeiro grupo são rejeitadas com um percentual cinco vezes maior que o percentual das do segundo.

Afora este, um outro ponto contribui para a diferenciação dos percentuais de rejeição entre as razões: o próprio modelo assumido para a distribuição da razão. Foram utilizados dois tipos de modelo, um para as razões de batimento:

$$R_B = (X) + (1 - X) \cdot R'_B$$

e outro para as razões-índices:

$$R_I = (1 - X) \cdot R'_I$$

Em ambos os modelos, X representa uma variável com distribuição de Bernoulli, com probabilidade p de sucesso igual ao percentual de questionários com valor igual a 1 (no caso das razões de batimento) ou 0 (no caso das razões-índices). R'_B e R'_I representam respectivamente as componentes aleatórias das razões de batimento e índice. As regras de crítica são formuladas para as compo-

nentes aleatórias das razões, devendo os percentuais de rejeição dados inicialmente serem corrigidos para levar em conta a parte determinística do modelo. Assim, para uma razão qualquer R , cujo percentual de rejeição α_R tenha sido obtido segundo a metodologia descrita na Seção II, obtém-se o percentual de rejeição real α_R^* da seguinte forma:

$$\alpha_R^* = \frac{\alpha_R}{1 - p},$$

onde, como antes p , representa o percentual de questionários com razões iguais a 1 (batimento) ou 0 (índice).

Os percentuais α_R^* são aqueles utilizados nos procedimentos de crítica descritos nesta seção.

Os valores de α_R para as diferentes razões de crítica são dados abaixo:

Razão	α_R	P	α_R^*
Matéria-prima	0,0110	0,6411	0,03065
Combustível	0,0110	0,8663	0,0823
GPC	0,0110	0,8442	0,0706
CPA	0,0110	0,5514	0,0245
GMP	0,002	0,3994	0,00366
Peças	0,002	0,6411	0,0061
Margem de comércio	0,0022	0,8562	0,0153

5.3 — Curvas de rejeição ($C_R(T)$)

O procedimento de estimação das curvas de rejeição $C_R(T)$ a partir das funções $\hat{P}_R(T)$ e $\hat{F}_R$ (distribuição empírica da razão R) é descrito em detalhe ao final da Seção III. Mencionou-se, então, a necessidade de estimar-se $C_R(T)$ apenas em um intervalo $[t_{\min}, t_{\max}]$, onde a medida de tamanho T toma valores com probabilidade 1. No caso da variável LPO, este intervalo de estimação foi dado por $[0,8,53621]$. A hipótese por trás do uso deste intervalo é a de invariância da distribuição de LPO de um censo para outro.

⁷ $A' = (1 - \delta_0)/(e^{Bt} - e^{Bt_0})$

$C' = (\delta_0 \cdot e^{Bt} - e^{Bt_0})/(e^{Bt} - e^{Bt_0})$

São apresentadas, a seguir, as curvas de rejeição estimadas para as diferentes razões de crítica (ver também Gráfico 3):

Razão	$\hat{C}_R(T)$
Matéria-prima	$0,3193 e^{-0,0729T}$
Combustível	$0,2629 e^{-0,028T}$
GPA	$7,3214 e^{-0,097T}$
GPC	$2,5655 e^{-0,155T}$
GMP	$30,468 e^{-0,0348T}$
Peças ⁸	$0,3139 e^{-0,0729T}$

A razão de margem de comércio recebeu um tratamento diferenciado das demais, pela impossibilidade de assumir-se a hipótese de preservação de sua distribuição de probabilidade de um censo para outro, já que esta é fortemente influenciada pelos níveis de inflação vigentes, bastante diferentes em um período e outro. A metodologia de crítica utilizada para esta razão é descrita em detalhe ao final desta seção.

Após a análise das curvas de rejeição por parte dos técnicos do DEIND, procedeu-se à alteração do percentual de rejeição da razão associada a Giro de Matéria Prima — GMP, já que os pontos críticos de rejeição foram considerados muito flexíveis (valores muito absurdos seriam aceitos). O percentual então utilizado foi o mesmo das razões consideradas importantes, ou seja, 0,011. O percentual global de rejeição foi elevado para 0,058 por esta modificação. A curva de rejeição estimada para GMP nestas condições foi $23,19 e^{0,102T}$, que levou a valores mais coerentes com as expectativas dos analistas do DEIND, sendo, portanto, adotada.

5.4 — Validação dos procedimentos

De modo a checar-se a validade da metodologia sugerida, submeteu-se o restante dos questionários não utilizados na etapa de estimação aos procedimentos de crítica obtidos anteriormente. Determinou-se, então, o percentual de rejeição por razão e no total, tendo-se obtido os resultados abaixo:

Razão	Percentual de rejeição
Matéria-prima	0,015
Combustível	0,009
GPA	0,012
GPC	0,014
GMP	0,009
Total	0,056

Observe-se que o percentual de rejeição total não é soma dos percentuais individuais de cada razão, já que um percentual considerável dos questionários rejeitados o é em mais de uma razão. Os resultados referentes à razão de consumo de peças não puderam ser calculados, já que não havia informações sobre a variável no Censo de 1980.

A análise dos percentuais obtidos indicou que a metodologia de crítica sugerida atingiu os objetivos aos quais se propunha de forma satisfatória. Como era esperado, observaram-se oscilações do percentual de rejeição real em torno do valor teórico, mas em magnitude tal que não compromete os objetivos fixados para os procedimentos de crítica. Note-se que o percentual total de rejeição ficou um pouco acima do esperado (0,058), já que o valor acima (0,056) deve ser aumentado pelas rejeições devido à margem de comércio e ao consumo de peças.

5.5 — Margem de comércio

A construção de limites críticos para a variável margem de comércio — MC, foi feita de maneira diferente das demais.

Esta variável, medida em cada estabelecimento industrial, refere-se a uma atividade secundária: venda de produtos não-industrializados no estabelecimento em questão, isto é, a variável refletia o lucro obtido com revenda de produtos. Há um pequeno número de estabelecimentos que exercem esta atividade marginal, a qual era muito pouco conhecida na época deste trabalho. Um dos motivos de incluí-la no pro-

⁸Em face da inexistência de dados sobre compra e consumo de peças no Censo de 1980, utilizou-se a função estimada para a razão de matéria-prima também para esta razão.

cedimento de crítica era garantir dados de boa qualidade para estudos exploratórios deste aspecto do funcionamento dos estabelecimentos.

Por motivos variados não era de interesse uma crítica desta variável com base em curvas de rejeição. Buscava-se portanto apenas um intervalo de aceitação simples para os valores de MC em 1985.

O escasso conhecimento sobre a variável não incluía a influência da inflação ao longo do tempo. A característica de lucro numa atividade marginal pouco conhecida dava bases pouco sólidas ao pressuposto da invariância da variável entre períodos de tempo com níveis de inflação muito diferenciados. Assim, utilizar intervalos de aceitação calculados com dados de 1980 para a crítica dos dados de 1985 parecia um pouco arriscado.

A alternativa adotada foi obter uma pequena amostra desta variável em questionários CE-301 do Censo de 1985. Devido a dificuldades de acesso aos dados de 1985, a amostra deveria ser obtida com uma restrição: só estavam disponíveis questionários de uma população de estabelecimentos pertencentes à Coleta Especial — CE⁹. Grosseiramente, esta Coleta Especial — CE referia-se a um grupo de empresas, em geral de grande porte, que recebiam tratamento diferenciado em todo o processo de realização do censo.

Para descrever o procedimento vamos definir:

POPA = população A = Conjunto de valores da variável MC no censo do ano A em todos os estabelecimentos que exercem a atividade de revenda.

SUBPOPA = Conjunto de valores da variável MC no Censo do ano A em todos os estabelecimentos da CE que exercem a atividade de revenda.

Usando-se os dados de 1980 estimou-se a relação entre os valores de POP80 e SUBPOP80:

- seja $POP80 = \{Y_1, \dots, Y_N\}$

e $SUBPOP80 = \{z_1, \dots, z_n\} \subset POP80$

- tomou-se os n percentis $Y_{(k)}$ de ordem

$$\left(K \cdot \frac{N}{n}\right) \quad 100\% \text{ de } POP80,$$

onde $K = 1, \dots, n$. Fez-se uma regressão das estatísticas de ordem $z_{(1)}, \dots, z_{(n)}$ de SUBPOP80 em $Y_{(1)}, \dots, Y_{(n)}$ obtendo-se $\hat{z}_{(k)} = A + B Y_{(k)}$ $K = 1, \dots, n$ e outra regressão de $Y_{(k)}$ em $Z_{(k)}$, obtendo-se $\hat{Y}_{(k)} = A' + B' z_{(k)}$ $K = 1, \dots, n$

- Estas relações usam dois grupos de dados de um mesmo ano, 1980, e portanto entre os grupos não há diferenças devido à inflação.

- Foi feita a hipótese de que as relações lineares acima não mudaram em 1985.

Para lidar com a mudança da distribuição de 80 para 85, utilizou-se SUBPOP80 e a amostra de SUBPOP85:

- Sejam $x_1, \dots, x_m$ os valores ordenados obtidos da amostra de SUBPOP85.

- Sejam $z'_1 < \dots < z'_m$ os m percentis empíricos de SUBPOP80 ($n > m$).

- Fez-se uma regressão de $z'_1, \dots, z'_m$ em $x_1, \dots, x_m$, obtendo-se

$$z'_i = a + bx_i \quad i = 1, \dots, m$$

Foi obtida uma relação entre POP80 e POP85:

- Sejam C_1 e C_2 os percentis de ordem $\frac{\alpha}{2}$

100% e $(1 - \frac{\alpha}{2})$ 100% respectivamente de

POP80.

Estimou-se por C'_1 e C'_2 os percentis de ordem

$\frac{\alpha}{2}$ 100% e $(1 - \frac{\alpha}{2})$ 100% respectivamente de POP85 como

$$C'_1 = A' + B' (a + b (A + BC_1))$$

$$C'_2 = A' + B' (a + b (A + BC_2))$$

O intervalo $[C'_1, C'_2]$ foi então utilizado como intervalo de aceitação para a variável margem de comércio no processo de crítica do Censo de 1980¹⁰. Espera-se que este intervalo rejeite aproximadamente α 100% dos valores da variável.

⁹Estes questionários também não estavam digitados. Apesar de toda a subpopulação estar disponível, foi preciso retirar uma pequena amostra, pois os dados tinham de ser obtidos manualmente, diretamente do questionário a um custo altíssimo.

¹⁰Até 26/09/88 foram criticados 41% dos questionários do Censo Industrial de 1985 e os percentuais de rejeição de questionários em cada razão estão de acordo com o esperado.

APÊNDICE

Os gráficos aqui incluídos ajudam a esclarecer determinados tópicos apresentados durante a exposição da metodologia e dos resultados do trabalho.

GRÁFICO 1

RELAÇÃO ENTRE OS LOGARITMOS DOS VALORES
NOS ESTOQUES EM 31/12/79 E 31/12/80

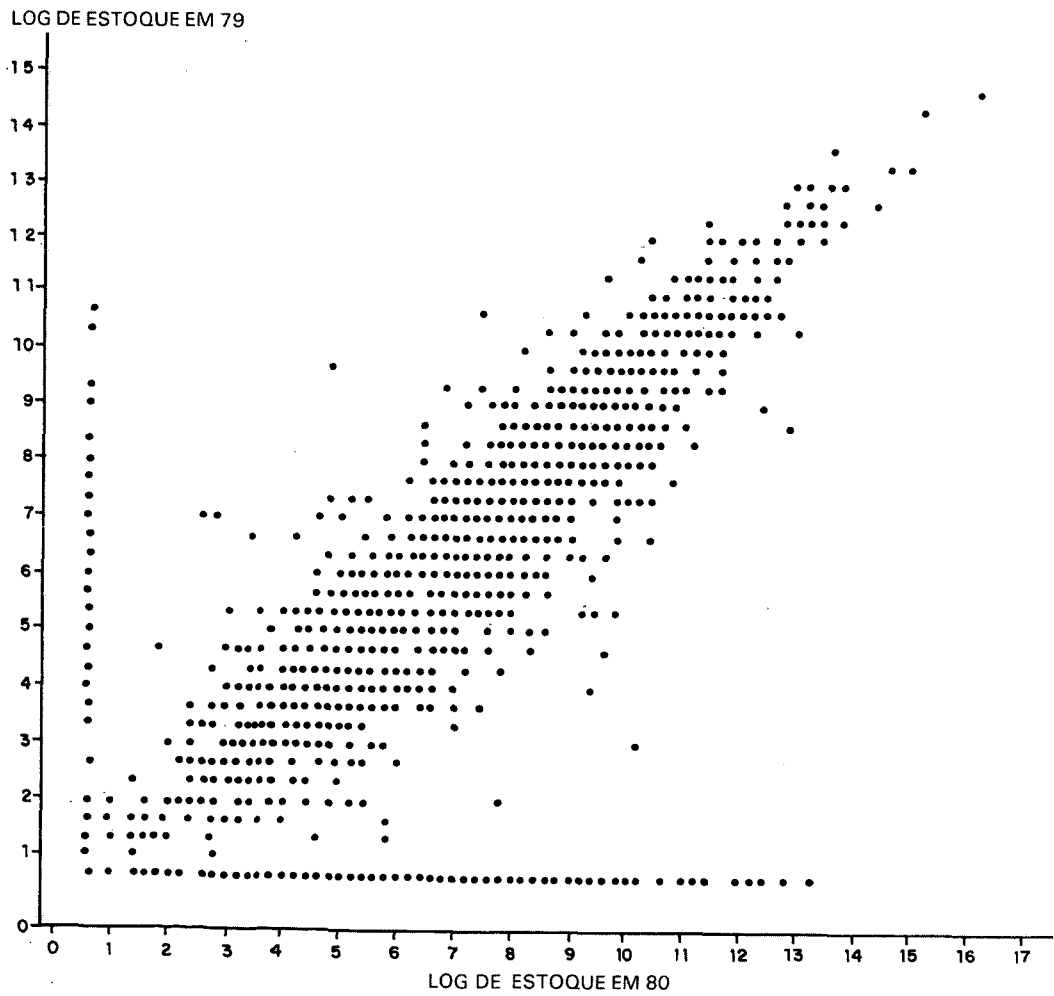


GRÁFICO 2
HISTOGRAMA DA MEDIDA DE TAMANHO Ç LPO

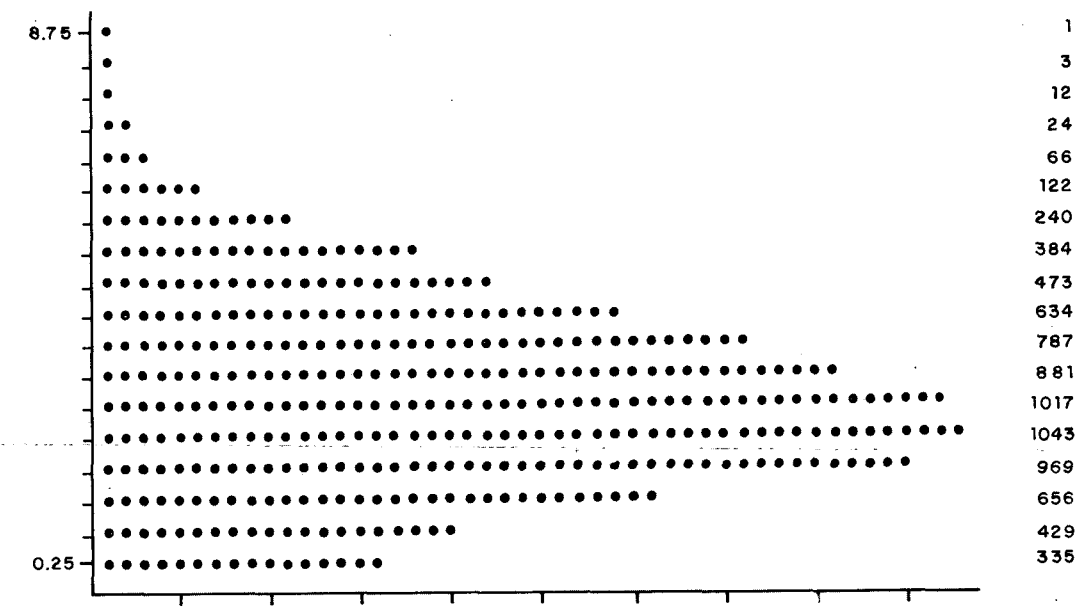
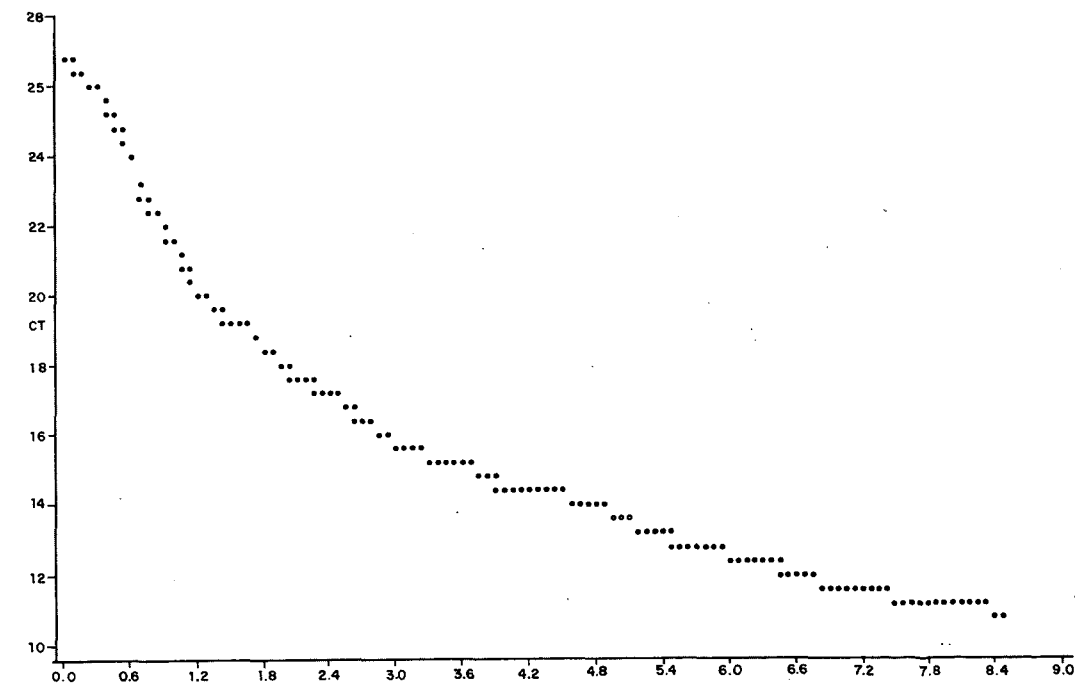


GRÁFICO 3
EXEMPLO DE CURVA DE REJEIÇÃO
(NO CASO REFERENTE AO GIRO DE MATÉRIA-PRIMA — GMP)



BIBLIOGRAFIA

- HANDBOOK of mathematical functions. Abramowitz, M. and Segun, I.A. (editors). Dover Publication, Inc., New York, 1968.
- JOHNSON, N.L.; KOTZ, S. *Continuous univariate distributions*, v.1 e 2 Houghton Mifflin Company, Boston, 1970.
- SAS User's guide: Basics*, Version 5 edition SAS Institute Inc., Cary, NC, 1985.

RESUMO

O IBGE faz um trabalho de crítica dos dados fornecidos pelo respondente nos questionários dos Censos Econômicos. Uma parte deste trabalho procura determinar intervalos de aceitação para um conjunto de razões envolvendo diferentes campos do questionário para indústria CE-301. A inovação mais importante na metodologia de crítica foi o emprego de curvas de rejeição ao invés de valores fixos, o que permitiu um tratamento diferenciado dos questionários de acordo com o porte dos estabelecimentos.

ABSTRACT

Brazilian National Bureau of Census (IBGE) carries out data checking provided by respondents on the Economic Census questionnaires. Part of this work is setting up acceptance intervals to ratios set involving different items of the industry questionnaire CE-301. The most important methodological innovation was the use of rejection curves instead of fixed values, allowing a differential treatment of questionnaires according to the establishments size.

RESENHAS BIBLIOGRÁFICAS

— **OUTLIERS IN STATISTICAL DATA**, Vic Barnett and Toby Lewis. 2nd. edition. Wiley. New York. 1984. XIV + 463 pages.

Sem dúvida, é muito importante para qualquer estatístico o estudo formal do tema *outliers*. Aquelas observações "afastadas" da maioria dos dados colhidos, quando são registrados os resultados duma experiência, têm preocupado principalmente aos estatísticos experimentais em todas as épocas. Sem muito rigor essas observações têm recebido o nome de *outliers*. Devido a essa preocupação é que a comunidade estatística deve receber, com ansiedade, a aparição duma verdadeira enciclopédia sobre esse tema.

A primeira edição do livro, que ora reviso, foi publicada em 1978. Muito tem sido feito desde lá até 1984, e ainda se faz hoje sobre este assunto tão empolgante. Assim, a segunda edição não é apenas uma reedição da primeira, na verdade, é um trabalho novo.

Possivelmente, a melhor alternativa para se ter uma idéia do conteúdo de ambas as edições é ler o que os autores dizem nos prefácios.

O propósito deles, na primeira edição, era colocar num texto só, da forma mais completa e organizada possível, toda a grande

variedade e técnicas de tratamento de *outliers* que surgiram, principalmente, a partir de 1900. Por um lado, revisaram todo (ou quase todo) o conhecido e apresentaram os detalhes dos diferentes procedimentos usados nas diversas situações. Por outro lado, eles classificaram as diferentes atitudes, objetivos e modelos de *outliers*. Procuravam assentar as bases para um desenvolvimento mais organizado dos trabalhos futuros. Dessa maneira, esperavam servir tanto aos estatísticos experimentais (analistas de dados), como aos pesquisadores interessados na indagação de caminhos novos.

Os primeiros trabalhos sobre *outliers* davam pouca atenção ao mecanismo ou modelo probabilístico que explicaria a aparição dessa classe de observações. Barnett e Lewis trataram de remediar essa falha através da definição duma certa quantidade de "modelos para *outliers*". Eles basearam essas sugestões sobre considerações teóricas e práticas.

A edição de 1978 resultou num trabalho no qual todo estatístico — aluno, analista de dados e/ou pesquisador "teórico" — poderia encontrar, estabelecidos sobre uma base lógica formal, os trabalhos mais destacados do século sobre *outliers*.

O nível de conhecimentos necessários ao leitor era o fornecido pela maioria dos cursos introdutórios à Probabilidade e Estatística dos programas da pós-graduação.

As proposições matemáticas foram demonstradas segundo a conveniência dos autores para cumprir com o seu objetivo. Daí que as provas são mais intuitivas que matematicamente elegantes, porém são rigorosas.

Barnett e Lewis completaram esse primeiro trabalho colocando uma grande quantidade de tabelas, algumas delas já publicadas anteriormente, para ajudar aos analistas no uso dos diferentes procedimentos explicados no texto.

Finalmente, pensando nos interessados em desenvolver pesquisas teóricas e/ou aplicadas colocaram, ao longo do livro, uma extensa bibliografia.

As características mais salientes dessa primeira edição foram mantidas na segunda, principalmente nos objetivos e forma da exposição. Porém a nova edição foi consideravelmente aumentada no que diz respeito aos procedimentos estudados. Devemos levar em conta o que Barnett e Lewis dizem no prefácio desta edição: desde 1978 até 1984 foram publicados mais de 300 trabalhos sobre *outliers* em diversas revistas, livros e/ou relatórios técnicos, isto é, houve um crescimento de quase 75% nesses seis anos, a respeito de tudo o que foi feito desde 1900 até 1978. Isto dá uma idéia do vigor e atualidade do tema.

A nova edição traz a análise de novos procedimentos para tratamento de *outliers*: certos procedimentos próprios da Inferência Robusta e outros mais informais sugeridos, digamos, pela Estatística Descritiva Gráfica.

Na segunda edição Barnett e Lewis analisam a relação entre *outliers* e "pontos influentes", particularmente no modelo de regressão.

Finalmente, nessa edição, os autores analisam situações e modelos mais complexos: séries temporais (pouco), planejamento de experimentos, dados numa direção etc... Também dizem alguma coisa sobre o ponto de vista bayesiano.

A seguir, resumi o conteúdo do livro, num esquema com os cabeçalhos de cada capítulo.

Alguns comentários breves são acrescentados em certos capítulos.

CHAPTER I: INTRODUCTION

Aqui é apresentada a questão geral em torno dos *outliers*, para vários modelos e situações, através de exemplos simples.

CHAPTER II: WHY DO OUTLYING OBSERVATIONS ARISE AND WHAT SHOULD ONE DO ABOUT THEM?

É apresentada uma breve história sobre a metodologia para tratamento dos *outliers*. A última seção deste capítulo é imperdível, pois, aí, os autores definem os diversos modelos para *outliers* nos casos de amostras univariadas.

CHAPTER III: THE ACCOMODATION APPROACH: ROBUST ESTIMATION AND TESTING

CHAPTER IV: ACCOMODATION PROCEDURES FOR UNIVARIATE SAMPLES

Na minha opinião o tratamento da Robustez, nestes dois últimos capítulos, é muito formal, isenta de um apelo à intuição. Porém, é útil para quem não necessita de motivação para estudar Robustez.

CHAPTER V: TESTING FOR DISCORDANCY: PRINCIPLES AND CRITERIA

CHAPTER VI: SPECIFIC DISCORDANCY TESTS FOR OUTLIERS IN UNIVARIATE SAMPLES

Os dois últimos capítulos formam uma das partes mais importantes e novas do livro.

CHAPTER VII: OUTLIERS IN DIRECTIONAL DATA

CHAPTER VIII: OUTLYING SUB-SAMPLES: SLIPPAGE TESTS

CHAPTER IX: OUTLIERS IN MULTIVARIATE DATA

CHAPTER X: THE OUTLIER PROBLEM FOR STRUCTURED DATA: REGRESSION, THE LINEAR MODEL, AND DESIGNED EXPERIMENTS

Neste último capítulo, Barnett e Lewis estudam a relação entre "pontos influentes" e *outliers*.

CHAPTER XI: OUTLIERS IN TIME SERIES: A LITTLE-EXPLORED AREA

Na verdade, os autores falam muito pouco sobre o assunto.

CHAPTER XII: BAYESIAN APPROACHES TO OUTLIERS

CHAPTER XIII: PERSPECTIVE

Mesmo com o muito que se vem fazendo sobre *outliers* ainda fica sem resposta um problema fundamental: a formalização do conceito de "observação surpreendente". Mas as dificuldades não dispensam aos estatísticos da inadiável responsabilidade de contribuir para uma elaboração dinâmica duma metodologia para manejar *outliers*.

Em resumo, este livro deverá constituir parte de toda biblioteca que deseje servir aos estatísticos: alunos, analistas e pesquisadores. Uma ressalva: não parece recomendável ser usado como livro-texto, mas, sim, como uma "enciclopédia" sobre *outliers*.

Oscar Bustos — IMPA

— **CLASSIFICATION AND REGRESSION THEES**, por Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. California: Wadsworth International Group, x + 358 pp. (The Wadsworth Statistics/Probability Series)

Segundo seus autores, a metodologia de árvores, discutida neste livro, é filha da era do computador. Diferentemente de muitas outras que evoluíram do lápis para a calculadora e então para o computador, o seu uso em estatística seria inimaginável antes do computador. Motivados pela necessidade de tratar com dados e problemas reais, cientistas sociais introduziram árvores de regressão no programa AID (Automatic Interaction Detection), desenvolvido no Institute for Social Research, da Universidade de Michigan, por Morgan e Sonquist, no início da década de 60.

Os autores do presente livro iniciaram seus trabalhos no assunto em 1973, com métodos de árvores para classificação, e desenvolveram um programa de computador denominado CART (Classification and Regression Trees) que evoluiu até o presente livro.

O livro apresenta técnicas não paramétricas para resolver os problemas estatísticos de regressão e de classificação que se baseiam na construção de árvores, obtidas por sucessivas partições binárias da amostra disponível. Nestas árvores, o nó inicial corresponde à amostra inteira que é subdividida em duas subamostras, e assim sucessivamente até se atingir a partição final da amostra em várias subamostras. Cada subdivisão é definida por critério apropriado, que depende das variáveis observadas, e também do problema sob consideração. Os problemas analisados são os de Regressão e de Classificação, podendo-se facilmente, por modificação do critério de partição dos nós, fazer adaptação para resolução de outros problemas estatísticos.

A metodologia utilizada é definida por três elementos: o critério de partição dos nós; a regra de parada; e a classificação dos nós terminais da árvore.

O livro dispensa maior atenção ao problema de classificação, onde, como sabemos, medições são feitas em algum caso ou objeto para com base nestas prever sua classe. O ponto de partida para a solução do problema é uma amostra de "aprendizado", cujos elementos têm suas classes já conhecidas, sendo também observadas, para cada elemento da amostra, diferentes variáveis que irão definir a regra de classificação. A solução clássica do problema adota a hipótese de normalidade conjunta das variáveis observadas, originando funções de classificação lineares ou quadráticas, dependendo de se conhecer ou não a matriz de covariância das observações. O método proposto, ao contrário, não faz hipóteses relativas à distribuição das observações ou à forma da função de classificação. O produto final do método é uma árvore com seus nós terminais tendo classes determinadas. A classe de um novo caso é a do nó terminal a que ele pertence.

Antes de qualquer coisa, para a utilização desta metodologia é imprescindível se dispor de um algoritmo eficiente, devido à complexidade computacional dos métodos envolvidos, exemplificando a nova filosofia de utilização intensiva do computador em Estatística.

A apresentação dispensa conhecimentos prévios sobre Regressão e Classificação. Vários exemplos de aplicação são fornecidos e a leitura do livro exige, exceto pelo último capítulo, poucos pré-requisitos.

O livro é bastante didático e sua leitura bastante estimulante, com exemplos muito bem selecionados. É, sem dúvida, uma obra de grande utilidade para os que têm interesse em aplicações da estatística ou em sua teoria. São nele discutidos vários problemas teóricos interessantes, em aberto.

Comparando-se às técnicas clássicas, as apresentadas no livro têm as seguintes vantagens:

1. Não adotam hipóteses relativas à distribuição de probabilidades dos dados ou à forma das funções de Classificação ou de Regressão;
2. Podem ser usadas mesmo quando a dimensão do vetor de dados varia com os indivíduos;
3. São facilmente adaptáveis para o caso de dados omissos (*missing values*);
4. Produzem saídas fáceis de interpretar e de explicar ao leigo em estatística. As árvores obtidas mostram claramente quais as variáveis mais importantes e como atuam conjuntamente na Classificação ou Regressão;
5. Definem subamostras, na divisão de cada nó, obtidas por condicionamentos sucessivos de eventos definidos sobre as variáveis observadas, permitindo assim a análise das interações entre as variáveis;
6. Asseguram um ajuste do modelo adaptado à região particular do espaço das variáveis observadas, tirando partido de possível mudança na estrutura de dependência das variáveis em diferentes regiões; e
7. São facilmente adaptáveis a outros problemas estatísticos, bastando alterar-se o critério de subdivisão das amostras e a análise dos nós finais.

Contudo, devem-se notar as seguintes deficiências das Árvores de Regressão e Classificação:

1. Têm grande complexidade computacional, exigindo programa de computador eficiente para sua aplicação;
2. Podem depender fortemente da amostra disponível, sendo instável o ajuste. A sua regra de parada não é fácil de implementar, podendo gerar superajuste. O controle de graus de liberdade, que é feito nas técnicas clássicas, não é possível nesta metodologia; e
3. Fazem otimização passo a passo por ocasião das subdivisões, podendo produzir solução bem distinta da globalmente ótima. A grande complexidade computacional do problema, por outro lado, torna impossível a busca desta última.

O livro está distribuído em 12 capítulos, sucintamente descritos a seguir:

CAPÍTULO 1 — BACKGROUND

Introduz o problema estatístico da classificação

CAPÍTULO 2 — INTRODUCTION TO TREE CLASSIFICATION

Apresenta exemplos para motivar as árvores de classificação, inclusive analisando a situação onde a dimensão do vetor de observações varia entre os indivíduos. Introduz ainda neste capítulo o critério para definir a partição dos nós de modo a diminuir o seu "grau de impureza", tornando mais homogêneos os nós descendentes. Para isto utiliza o critério de Gini. Apresenta dois exemplos interessantes: um sobre reconhecimento de dígitos e outro de detecção de forma de ondas.

CAPÍTULO 3 — RIGHT SIZE TREES AND HONEST ESTIMATES

Analisa a regra de parada no processo de partição dos nós e, após a parada, a obtenção de estimadores das probabilidades de classificação incorreta.

CAPÍTULO 4 — SPLITTING RULES

Estuda critérios de otimização para definir as sucessivas subdivisões da amostra dos nós da árvore.

CAPÍTULO 5 — STRENGTHNING AND INTERPRETING

Fornece critérios mais gerais, em função das variáveis observadas, para definir as partições. Analisa também a situação de dados omissos.

CAPÍTULO 6 — MEDICAL DIAGNOSIS AND PROGNOSIS

Apresenta análise completa de várias pesquisas na área médica: prognóstico após ataque cardíaco; diagnóstico de ataques cardíacos; imunossupressão e diagnóstico de câncer; estudo de postura. Analisa ainda neste capítulo o problema de detecção de valores atípicos (*outliers*).

CAPÍTULO 7 — MASS SPECTRA CLASSIFICATION

Analisa exemplo na área de ecologia, construindo árvores para detectar a presença de certos elementos químicos, em compostos, através de sua massa espectral.

CAPÍTULO 8 — REGRESSION TREES

Estuda Árvores de Regressão. Relaciona a técnica ao AID (Automatic Interaction Detection) proposto por Morgan e Sonquist em 1963.

CAPÍTULO 9 — BAYES RULES AND PARTITIONS

Introduz *background* teórico para julgar a qualidade do ajuste feito através de árvores.

Usa para isto o enfoque de teoria de decisão, com o critério de Bayes para definir regras ótimas.

CAPÍTULO 10 — OPTIMAL PRUNING

Trata do problema do tamanho ótimo da árvore. Quando se deve parar de subdividir a amostra? Partições mais finas diminuem o erro estimado mas implicam superajuste à amostra disponível.

CAPÍTULO 11 — CONSTRUCTION OF TREES FROM LEARNING SAMPLES

Estuda procedimentos para diminuir o vício na estimação dos erros de classificação.

CAPÍTULO 12 — CONSISTENCY

Mostra a consistência dos procedimentos de classificação e regressão em árvores sob condições bem gerais. É um capítulo de teoria que requer, para sua leitura, conhecimentos de Análise e de Teoria das Probabilidades.

O livro, apesar de escrito para estatísticos, tem sido freqüentemente citado na literatura de sistemas especialistas, na área de computação. As técnicas nele descritas fornecem exemplo de algoritmo para incorporar o conhecimento contido na amostra, na classificação de novos indivíduos.

Djalma G.C. Pessoa — ENCE/IBGE

INSTRUÇÕES BÁSICAS PARA PREPARO DOS ORIGINAIS

Os originais entregues para publicação devem obedecer às seguintes normas:

1 — O texto deve ser datilografado em papel branco, tamanho ofício, em um só lado, em espaço duplo, com margem de 3 cm em todos os lados do papel, sem rasuras ou emendas que dificultem sua leitura e compreensão. As páginas devem ser numeradas seqüencialmente, contendo até 30 linhas de 72 batidas cada. Os autores interessados poderão solicitar ao Editor responsável laudas-padrão. Todas as cópias submetidas devem ser legíveis.

2 — A primeira página do original (folha-de-rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a filiação e endereço permanente para correspondência. Agradecimentos a colaboradores, outras instituições e auxílios recebidos devem figurar também nesta página.

3 — 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.

4 — O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Subseções podem ser usadas, mas não subdivisões de subseções. Onde forem usadas subseções, todas, inclusive a primeira, devem ser numeradas e receber título apropriado.

5 — Sentenças ou palavras entre parênteses, sublinhadas ou em tipos diferentes (itálico) não devem ser usadas. Notas de rodapé devem ser evitadas. Abreviações e siglas, tais como, i.i.d., g.l., A.A.S., ANOVA e símbolos especiais tais como $\forall$ e $\frac{b}{a}$ não devem ser empregados.

6 — A citação de referências no texto deve ser feita de acordo com os exemplos apresentados a seguir:

a) Pereira e Portugal (1987)

b) Costa (1987, p. 39)

c) Hansen et alii (1953, cap. 5)

As referências listadas ao final devem corresponder exatamente às que foram citadas no texto. Referências a livros devem indicar a edição, citando no texto: página, seção ou capítulo. Referências a livros ou artigos devem incluir: título do livro (ou artigo), editor(es), número da primeira e última páginas do artigo, entidade que publicou e localidade onde foi publicado. Na lista final

das referências, todos os autores devem ser mencionados, e a abreviatura "et alii" nunca deve ser usada. Referências a documentos não publicados pressupõem que o autor poderá fornecer cópia do material citado. A listagem final das referências deve ser apresentada em ordem alfabética do último sobrenome do autor e, para um mesmo autor, em ordem cronológica de publicação. Veja os exemplos a seguir:

Costa, L. N. da (1987). Aplicação da amostragem na coleta dos Censos Demográficos no Brasil. *Revista Brasileira de Estatística*, 48, n.º 189/190, pp. 35-64, Rio de Janeiro.

Hansen, M. H., Hurwitz, M. N. e Madow, W. G. (1953). *Sample survey theory and methods*, V.I. New York, John Wiley & Sons.

7 — As tabelas e gráficos devem ser apresentados em folhas separadas, precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numerados seqüencialmente (Tabela 1, Figura 3 etc....) e referidos 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 títulos e numeração, tal como as demais seções do trabalho.

8 — As fórmulas matemáticas devem ser apresentadas com clareza, para evitar problemas de interpretação. Siga as regras indicadas e os exemplos mostrados.

a) arrange os parênteses em ordem { [()] };

b) siga as convenções usuais para e, exp, log etc.;

c) prefira a forma $x^{1/2}$; nunca use $\sqrt{x}$;

d) notações tais como $\hat{x}$, ou $\underline{x}$ ou $\widehat{(x+y)}$ nunca devem ser usadas;

e) sub e superíndices (inclusive de segunda ordem) devem ser, se possível, alinhados horizontalmente, ou, do contrário, claramente marcados em tinta de cor; evite sub e superíndices de ordem maior que 2;

f) use sigma grego maiúsculo Σ apenas para indicar somatórios;

g) procure fazer distinção clara entre caracteres que se confundem facilmente, tais como:

w (dáblio) e ω (ômega minúsculo)

v (vê) e ν (ni minúsculo)

o (o minúsculo), O (o maiúsculo) e 0 (zero)

1 (um) e l (ele minúsculo); e

h) procure usar a notação indicada nos exemplos a seguir:

Use

var(x)

cov

pr para probabilidade

tr para traço de matriz

E(X) para esperança de X

$\log(x)$

$a/(bc)$ ou $a(bc)^{-1}$

0,2

1.000,00

$x_1, \dots, x_n$

$\Gamma\left(\frac{1}{2}n + \frac{1}{2}\right)$

ab

Não Use

var x ou VAR(x)

Cov ou COV

P ou Pr

Traço ou traço

EX ou $\mathcal{E}(X)$

$\log_e(x)$ ou $\ln(x)$

a/bc

0.2 ou .2 ou ,2

1000 ou 1000,00

$x_1, x_2, \dots, x_n$

$\Gamma[(n+1)/2]$ ou $\Gamma\left(\frac{n+1}{2}\right)$

a.b ou $a \times b$

9 — Equações devem ser numeradas somente se citadas no texto; nesse caso, a numeração deve ser dada entre colchetes [] e ser colocada alinhada junto à margem direita. Expressões ou equações longas e/ou importantes devem ser destacadas, isto é, apresentadas em linhas separadas. Fórmulas curtas devem ser deixadas no texto para poupar espaço, quando possível. Nenhuma fórmula deixada no texto deve ter mais de uma linha de altura. Portanto, $\sum_{i=1}^n x_i$ não deve ser deixada no texto, ou então deve

ser escrita $\sum x_i$ (se os limites de somação forem óbvios). A forma $\sum_1^n$ não deve ser usada sob qualquer hipótese. Da mesma forma, $\binom{a}{b}$ não deve ser deixada no texto. Equações ou expressões muito longas devem ser evitadas, sempre que possível, introduzindo-se notação apropriada.

10 — Gráficos e diagramas para publicação devem ser traçados em papel branco, com nitidez e boa qualidade, para permitir que a redução seja feita mantendo-se a qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros, quer no desenho quer nas legendas ou títulos.