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, nº 192, p. 1-168, jul./dez. 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.

Edição de Texto e Editoração Eletrônica

Escola Nacional de Ciências Estatísticas — ENCE e Gerência de Editoração — GEDIT.

Esta publicação foi composta e diagramada pelo sistema de editoração eletrônica, com emissão em "laser HP". Tendo em vista as dificuldades iniciais de operação do sistema, as normas editoriais não puderam ser cumpridas integralmente.

Capa
Pedro Paulo Machado

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) nov. 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)

Impresso no Brasil/Printed in Brazil

SUMÁRIO

NOTA DO EDITOR

ARTIGOS

OUTLIERS E ROBUSTEZ 7
Oscar H. Bustos

ESTUDO DO PONTO DE CORTE EM UMA REGRA DE CLASSIFICAÇÃO PARA DADOS BINÁRIOS E UMA APLICAÇÃO EM PNEUMOLOGIA 31
Beatriz Mendes Luz

DIFERENCIAIS DE RENDIMENTOS ENTRE CENSOS DEMOGRÁFICOS E PNADs - ASPECTOS CONCEITUAIS E OPERACIONAIS 49
André Cezar Medici

REDUÇÃO DA AMOSTRA DA PESQUISA MENSAL DE EMPREGO: ESTRATÉGIA PARA REDUZIR O CUSTO DA PESQUISA 65
Pedro Luis do Nascimento Silva
Fernando Antonio da Silva Moura

DISTRIBUIÇÃO ESPACIAL DA POPULAÇÃO BRASILEIRA E ALGUMAS CARATERÍSTICAS SÓCIO-ECONÔMICAS ENTRE 1960-1980 97
Antonio de Ponte Jardim

RESENHAS BIBLIOGRÁFICAS

GENERALIZED LINEAR MODELS 155
Gauss M. Cordeiro

BAYESIAN ANALYSIS OF LINEAR MODEL 157
Helio S. Migon

PUBLICAÇÕES RECEBIDAS 163

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBE's 165

ISSN 0034-7175

R. bras. Estat. Rio de Janeiro, v.49, n.192,p.1-168, jul./dez. 1988

NOTA DO EDITOR

Informamos aos leitores da Revista Brasileira de Estatística que o International Statistical Institute (ISI) publicará, de agora em diante, os sumários dos artigos desta Revista, no periódico *Statistical Theory and Methods Abstracts*.

OUTLIERS E ROBUSTEZ*

Oscar H. Bustos (IMPA/CNPq)**

7 – UMA TÉCNICA DE "ACOMODAMENTO" OU "CONVIVÊNCIA" COM OS OUTLIERS: USO DE ESTIMADORES ROBUSTOS

Em um sentido amplo e sem tecnicismos, a "Estatística Robusta" tem a ver com o fato de que muitas suposições que são feitas habitualmente em Estatística são somente aproximações da realidade. Neste sentido, podemos dizer que qualquer método para lidar com *outliers*, será objeto de interesse para a "Estatística Robusta".

Relendo a história desse ramo do conhecimento estatístico, veremos que, aqui também, a subjetividade está presente de forma notável. Somente nas últimas décadas se fizeram esforços para construir uma "Teoria da Robustez".

Hoje em dia existe uma grande variedade de formas de encarar o problema da robustez. Entre elas, devemos destacar, por sua presença vigorosa, aquela que se preocupa em colocar todos os elementos essenciais dos métodos robustos usados para abordar problemas reais,

*Nota do editor. Escrito a convite, este artigo objetiva oferecer uma retrospectiva sobre "Outliers e Robustez". Dividido em 3 partes, a 1ª foi publicada no número 191, a 2ª neste número e, a 3ª será publicada *no próximo número desta revista*.

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

dentro de um rigoroso contexto matemático. Têm surgido assim novos conceitos, definições e critérios de otimalidade cujas propriedades formais devem ser objeto de estudo de estatísticos matemáticos.

Também os "analistas de dados" encontram dentro da Robustez um esquema formal para questões tais como: dado um conjunto de dados, existem alguns subconjuntos que "dizem" coisas diferentes das contadas por todo o conjunto?; que diz a "maioria" dos dados?; qual é a influência, no resultado final de uma análise, dos diferentes subconjuntos extraídos do conjunto total de dados?; quais são os dados de maior importância para a escolha de um modelo e/ou para os resultados finais?; que dados devem ser analisados com maior cuidado?; que proporção de dados "maus" pode tolerar tanto o desenho experimental, quanto o modelo e/ou os procedimentos de estimação ou teste de parâmetros?; quais são os procedimentos mais seguros e eficientes? A tentativa de responder formalmente (com respostas "matemáticas") a essas e outras muitas perguntas constitui a principal motivação da "Teoria da Robustez".

Sem dúvida, onde mais se tem avançado em todos os aspectos, é no que poderíamos chamar de Robustez em Estatística Paramétrica. Isto é, supondo que as hipóteses com as quais se trabalha habitualmente em Estatística Paramétrica são aproximadamente válidas. Daí ser importante fazer uma boa escolha do modelo com o qual se trabalhará inicialmente. Deve ser estatisticamente significativo, mais que matematicamente atraente.

Os principais objetivos da Estatística Robusta são:

(i) Descrever o modelo que melhor se ajusta à maioria dos dados.

(ii) Identificar observações que se desviam da maioria para tentar dar–lhes um tratamento futuro.

(iii) Identificar e diagnosticar a presença de "pontos influentes". Isto é particularmente importante quando se trabalha com dados não balanceados.

(iv) Estudar o que se passa quando as estruturas de correlação assumidas não são estritamente válidas. Na verdade, estudos deste tipo começaram a ser encarados teoricamente depois de 1980, pioneiramente por Graf (1983).

Uma descrição mais detalhada destes objetivos, com exemplos esclarecedores para o leitor, pode ser encontrada em Hampel e outros (1986), havendo também alguma coisa em Bustos (1986).

Na Seção 5 nos referimos de passagem à influência negativa das observações distantes sobre o desvio – padrão amostral. Vejamos agora brevemente o que se passa com o consagrado método de estimação por máxima verossimilhança, quando a amostra apresenta *outliers* e pontos contaminantes. Na verdade, este método, sugerido por Fisher, é de duvidoso valor, a menos que coincidam os três modelos presentes no estudo: o verdadeiro, o suposto e o que realmente seguem as observações. Porém, este método pode ser modificado de modo a fazê–lo confiável ainda quando não se der esse caso de coincidência, coisa que quase sempre sucede na

prática. Um estudo muito detalhado, atualizado e rico em referências, sobre "estimadores de máxima verossimilhança modificados" pode ser encontrado no livro já citado de Tiku e outros (1986).

O método de máxima verossimilhança, aplicado a amostras univariadas que se supõe serem distribuídas segundo a distribuição normal, indica como estimadores ótimos para a média e a variância populacionais as correspondentes medidas amostrais. Ótimos no sentido de que esses estimadores têm eficiência assintótica máxima entre os não viciados, sob normalidade. Contudo, Jeffreys e outros estatísticos sugeriram que a maioria dos conjuntos de dados de boa qualidade são bem ajustados por distribuições tais como a t_3 ou a t_9 (t–Student com 3 e 9 graus de liberdade) e também por curvas da família de distribuições chamadas "Curvas de Pearson". Pode–se ver, por exemplo, em Fisher (1922), que:

-Eficiência assintótica de média amostral sob $t_v = 1 - 6/(v(v+1))$.

-Eficiência assintótica de variância amostral sob $t_v = 1 - 12/(v(v-1))$.

Assim, para t_9, t_5 e t_3 a eficiência assintótica da média amostral é, respectivamente: 93, 80 e 50% eà da variância amostral é 83, 40 e ... 0%. Quer dizer que a média amostral, e a variância amostral em maior medida, são muito instáveis com respeito à eficiência assintótica. O perigo de usar esses estimadores, quando não se tem a certeza de que as observações são exatamente i.i.d. normais, fica claro ao observarmos quão semelhantes são as curvas da $N(0,1)$, t_3 e t_5, o que faz com que sejam bem difíceis de distinguir as amostras geradas por uma ou outra (ver a Figura 10).

Outros exemplos importantes são devidos a Tukey e seus colaboradores, no Laboratório de Computação de Stanford. Estão baseados na suposição de que, quando se quer observar uma certa variável aleatória, certas observações são realizações dessa variável, mas outras (em proporção pequena) são realizações de outra, ou outras, variáveis aleatórias que tenham uma distribuição diferente. Estas poderiam considerar–se "contaminantes". O modelo popularizado por estes analistas no caso univariado é o que supõe que os n erros de observação são realizações de n variáveis aleatórias independentes e identicamente distribuídas com distribuição comum

$$F(x) = (1-\varepsilon)\Phi(x) + \varepsilon G(x)$$

onde $\Phi(x)$ é a função de distribuição acumulada de $N(0,1)$ e $G(x)$ é outra função de distribuição acumulada diferente de Φ, que se supõe conhecida a menos de um parâmetro. Habitualmente (e para simplificar) se considera

$$G(x) = \Phi(\frac{x}{\sigma})$$

com $\sigma > 1$. Habitualmente se supõe também que ε é conhecido. Apesar das simplificações que tem este modelo, parece que se adapta bem a muitas situações experimentais. Tukey

demonstra que a eficiência assintótica da média amostral decresce rapidamente, passando de 1 (caso $\varepsilon = 0$), para 0.70 (caso $\varepsilon = 0.1$). Essa medida, quando aplicada à variância amostral, põe em evidência sua instabilidade de forma mais notável. Mais detalhes podem ser encontrados no excelente livro de Huber (1981).

Na verdade, não está isenta de discussão a questão da necessidade de métodos robustos. Com efeito, dando uma rápida olhada na abundante literatura sobre o assunto, sentimos uma sensação de confusão ante a mistura de pareceres conflitantes. Porém, no básico, não existe confusão: para conjuntos de dados de boa qualidade, ao menos para aqueles "pré–limpados" ou sem *outliers*, os métodos robustos não são necessários. A contribuição principal desses métodos aparece quando são aplicados em situações, por sinal muito freqüentes, nas quais, apesar dos dados já haverem sido "limpados", subsistem, ainda, *outliers* "mascarados" entre aqueles com que se vai trabalhar. Contudo, mesmo quando usados em situações com dados de alta qualidade, os métodos robustos podem oferecer melhorias notáveis sobre as técnicas clássicas. É muito interessante estudar estes aspectos e detalhes nos vários trabalhos sobre o assunto, que vêm sendo publicados de forma crescente. Por exemplo: Stigler (1977), Hill e Dixon (1982), Mallows (1975), Rocke e outros (1982), etc. Finalmente, naquelas situações reais onde a presença de pontos *outliers* é evidente, a necessidade do uso de técnicas de robustez é inquestionável. A Teoria da Robustez não se limita a estudar técnicas de estimação e/ou testes de hipóteses que possam ser mais estáveis frente ao não cumprimento de certas hipóteses que sustentam um modelo. Com efeito, incorporou–se ao jargão da Estatística uma série de termos que tratam de conceituar certas idéias, já latentes desde os primórdios da Estatística. Também definiram–se novos critérios para avaliar a *performance* das diversas técnicas, tendo como premissa básica o reconhecimento da realidade: nunca se pode saber com absoluta certeza se o modelo suposto coincide exatamente com o verdadeiro. Parece ter sido inevitável a utilização de uma matemática mais sofisticada que a usada na Estatística Clássica. Talvez isto seja um obstáculo para que muitos profissionais não especialmente treinados para lidar com as abstrações matemáticas usem essas técnicas em suas análises. De qualquer modo, é de se desejar que os "práticos" e os "teóricos" se esforcem em entender–se. Nesse sentido, os estatísticos matemáticos dedicados à robustez vêm tratando de tornar mais claras as idéias que se escondem por detrás do rigorismo lógico. De nossa parte, veremos algumas definições e resultados que são aplicáveis fundamentalmente ao estudo de amostras univariadas. Sua extensão a situações mais complexas é imediata em vários casos; outros ainda resistem a tal extensão.

De passagem, notamos o muito que há para se evoluir dentro da Robustez... isso sem levar em conta o que se deve fazer no campo da computação, para contar–se com *software* que permita utilizar com segurança e eficiência as técnicas robustas. Para dizer a verdade, já há "software–para–robustez" de excelente qualidade, mas dia a dia surgem novas técnicas e

esse *software* deve ser atualizado. Ao final desta seção listaremos alguns dos programas e/ou programotecas que estão dedicadas especialmente à robustez.

Vejamos agora alguns dos conceitos e resultados da Teoria da Robustez aplicados à análise de amostras univariadas.

7.1 – Definição de estimadores através de funcionais

Sem dúvida, uma das idéias mais frutíferas troduzidas em Robustez por Hampel (1968), baseadas em idéias anteriores utilizadas em Probabilidade por Prokhorov, Von Mises e outros, é a de definir os estimadores usando "funcionais", isto é, funções definidas sobre uma certa família de probabilidades. Para melhor entender sua relevância vejamos sua motivação intuitiva.

Sejam $y_1, \cdots, y_n$ números reais. Chama-se distribuição empírica de $y_1, \cdots, y_n$ à probabilidade definida sobre R por

$$m[y_1, \cdots, y_n](B) = \frac{1}{n} \sum_{i=1}^{n} I_B(y_i)$$

onde B é um subconjunto boreliano de R e I_B é a função indicadora de B.

Seja agora $Z(R)$ o conjunto de todas as probabilidades sobre R. Para cada n, seja $Z_n(\mathrm{R})$ o conjunto de todas as distribuições empíricas definidas por amostras de tamanho n; logo, $Z_n(\mathrm{R}) \subset Z(\mathrm{R})$.

Agora, todos os estimadores razoáveis de alguma medida de posição e/ou dispersão em uma amostra univariada têm a propriedade de não alterar seu valor se se permutarem os índices das observações, isto é, esses estimadores satisfazem:

$$T_n(y_1, \cdots, y_n) = T_n(y_{p(1)}, \cdots, y_{p(n)})$$

qualquer que seja a permutação p definida sobre $1, \cdots, n$. Vemos assim que o estimador depende não tanto da amostra, mas sim da distribuição empírica da mesma. Por isso é natural (depois de Hampel...) pensar em definir os estimadores não por meio de funções definidas sobre R^n, mas sim por funções definidas sobre um conjunto de probabilidades, ou seja : um subconjunto de $Z(\mathrm{R})$. Chegamos assim à seguinte

Definição: Diz-se que um estimador T_n está definido por um funcional T, definido sobre um subconjunto D de $Z(\mathrm{R})$, se:

$$T_n(y_1, \cdots, y_n) = T(m[y_1, \cdots, y_n]),$$

toda vez que essa fórmula tenha sentido.

Por exemplo, a média amostral está definida pelo funcional T, definido sobre $D = = \{G \in Z(\mathbf{R}): G$ é uma distribuição sobre R com momento de primeira ordem finito $\}$, dado por

$$T(G) = \int x dG(x) = E_G(X)$$

Também é usual trabalhar com uma ligeira modificação do conceito anterior, que se aplica quando se estuda o comportamento assintótico de certos estimadores.

Definição: Seja D um subconjunto de $Z(\mathbf{R})$, T um funcional definido sobre D com valores em $\mathbf{R}$; para cada n inteiro positivo, seja T_n uma função definida sobre $\mathbf{R}^n$ com valores em $\mathbf{R}$. Diz–se que a sucessão (T_n) está definida assintoticamente por T se:

$$T_n(X_1, \cdots, X_n) \to T(G)$$

cada vez que $X_1, \cdots, X_n$ sejam variáveis aleatórias independentes e identicamente distribuídos com distribuição comum G, e isso para toda G em D. Neste caso, diz–se que $T(G)$ é o valor assintótico de (T_n) em G.

Para a definição seguinte necessitamos de uma notação. Seja G em $Z(\mathbf{R})$. Com $L(G;Y)$ denotamos a distribuição de Y sob G; isto é, $L(G;Y)$ é a probabilidade sobre R definida por:

$$L(G;Y)(B) = G(Y \in B)$$

para todo boreliano B de $\mathbf{R}$.

Definição: Suponhamos que para cada G em D, domínio do funcional T, existe um número $V(T,G)$ tal que:

$$L(G; \sqrt{n}(T_n - T(G)) \to N(O, V(T,G))(D)$$

onde (D) significa convergência em distribuição. Então, chamamos ao número $V(T,G)$ "variância assintótica de (T_n) em G".

Suponhamos agora que estamos no caso paramétrico, ou seja, assumimos que a distribuição subjacente é uma de uma certa família $\mathcal{P} = \{F_\mu: \mu \in \Theta\}$. Suponhamos também que $\mathcal{P}$ está contida no domínio do funcional T. Então diz–se que T é "Fisher consistente para estimar μ" se:

$$T(nF_\mu) = \mu$$

para toda F_μ em $\mathcal{P}$.

O conceito que definiremos a seguir é um dos mais importantes introduzidos pela Teoria da Robustez.

Definição: Função de Influência: Seja T um funcional definido sobre D, seja G uma distribuição em D. Suponhamos que para cada y em $\mathbf{R}$ e cada t em um intervalo $[0,d]$ contido em $[0,1]$ está definida

$$f(t) = f(t;T,G,y) = T((1-t)G + tm[y])$$

onde $m[y]$ é a distribuição com massa 1 no ponto y. Chama-se função de influência de T em G avaliada em y a:

$$ICH(y) = ICH(y; T, G) = lim_{t\to 0} \frac{f(t) - f(0)}{t}$$
$$= \lim_{t\to 0} \frac{T((1-t)G + tm(1-T))G + tm[y]}{t}$$

A importância deste conceito em Robustez está em que, como é fácil de ver, ele nos dá uma formalização da velocidade de alteração do valor de T quando a distribuição G é contaminada por uma distribuição com massa 1 em y. Os funcionais que vão produzir estimadores robustos serão aqueles cuja função de influência seja limitada. Uma conta simples mostra que, por exemplo, a média amostral tem função de influência:

$$ICH(y) = y$$

para todo y, logo não é limitada e portanto não é considerada robusta. Por outro lado, a mediana amostral tem função de influência:

$$ICH(y) = \frac{sinal(y)}{2\varphi(0)}$$

onde φ é a densidade da $N(0,1)$, quando G é a distribuição $N(0,1)$.

Uma propriedade importante da função de influência é sua relação com a variância assintótica. De fato, se pode ver que na maioria dos casos de interesse prático se verifica:

$$V(T,G) = (ICH(y;T,G))^2 dy$$

O leitor interessado pode estudar diversos resultados de interesse, tanto teórico quanto prático, sobre a função de influência em diversos trabalhos, por exemplo: Hampel (1974), Hampel e outros (1986), Boos e Serfling (1980), Fernholz (1983), James e Bustos (1980), Bustos (1986).

A partir da função de influência se definem vários outros conceitos que são usados para analisar o comportamento dos estimadores definidos por funcionais, do ponto de vista da robustez.

Entre eles destacamos dois que são de interesse por sua relação com *outliers*:

Definição: Chama-se "sensibilidade" a erros grosseiros de T em G ao valor:

$$GES(T,G) = sup\{|ICH(y;T,G)| : y \in \mathbf{R}\}$$

Definição: Chama-se "ponto de rejeição" de T em G ao valor:

$$\rho(T,G) = inf\{r > 0 : ICH(y;T,G) = 0, |y| > r\}.$$

O primeiro conceito diz qual é a influência máxima que pode ocasionar um só *outlier*, enquanto que o segundo diz quando um ponto pode ser considerado como muito afastado da massa dos dados.

Um estimador ótimo do ponto de vista robusto com respeito a estes valores será aquele que tenha GES e ρ tão baixos quanto possível.

Outro conceito de grande importância definido na Teoria da Robustez é o de "ponto de ruptura" (*breakdown point*) de um estimador. A grosso modo, mede a máxima proporção de observações contaminantes que pode admitir uma certa técnica de estimação, antes de deixar de dar qualquer informação relevante sobre a quantidade que se deseja estimar. Vamos precisar um pouco esta noção no caso mais simples: estimação de um parâmetro assumindo um modelo paramétrico. Seja F_H a distribuição suposta (hipotética ou objetivo) das observações, (T_n) uma sucessão de estimadores de um parâmetro, digamos μ. Suponhamos que a distribuição real das observações seja F_O, e diferente de F_H. Diz-se que $\varepsilon*$ é o ponto de ruptura de (T_n) em F_H, quando (T_n) nos dá "informação válida" sobre μ se e só se a distância de F_O a F_H é menor ou igual a $\varepsilon*$. Considera-se então que quanto maior é $\varepsilon*$ mais robusta é (T_n) com respeito ao ponto de ruptura. Agora, dependendo de como se formalizem as noções de "informação válida" e "distância entre distribuições", ter-se-ão diversas definições de "ponto de ruptura". Qual delas é a mais adequada depende do contexto em que se está trabalhando. A totalidade destas definições pode ser colocada em uma de duas classes: definições assintóticas (para n tendendo para infinito) e definições aplicáveis a um tamanho fixo de amostra. Entre as várias propostas de uma e outra classe, apresentadas até agora, podem-se assinalar as definidas nos seguintes trabalhos: Hampel (1971), Huber (1981), Donoho e Huber (1983), Hampel e outros (1986). Todas elas apresentam vantagens e desvantagens segundo o ponto de vista de cada um e conforme seja a situação particular com que se esteja trabalhando. Na Seção 11 veremos uma delas aplicada à análise de vários estimadores para o caso em que se deseja estimar a média de uma população supostamente normal.

Existem vários outros conceitos utilizados em Robustez, por exemplo: robustez qualitativa, robustez min-max, sensibilidade a deslocamentós locais, curva de sensibilidade etc. Não os veremos aqui, mas se deve advertir que alguns deles, por exemplo "robustez qualitativa", requerem conhecimentos de matemática que vão bem além do Cálculo Diferencial.

É necessário destacar um aspecto de interesse prático importante. A grande maioria das técnicas robustas requerem o uso de um computador para serem calculadas. Já existe *software* disponível para algumas delas, principalmente quando se trabalha nos modelos de posição e regressão. Esse *software* está disponível em alguns *packages* muito difundidos, por exemplo: **BMDP** (1981) e **TROLL**. A partir de 1976 têm sido publicados, em vários informes técnicos, *software* especialmente desenhados para serem usados em Robustez, por exemplo: **LINWDR**, **PROGRESS**, e o mais extenso e detalhado **ROBETH**. Todos eles estão sendo

continuamente estendidos e melhorados, graças a um esforço conjunto de vários estatísticos e matemáticos.

Os conceitos definidos recentemente são usados como critérios de avaliação da *performance* das diversas técnicas de estimação. Uma técnica ótima do ponto de vista da robustez será uma que tenha as seguintes propriedades: variância (ou variância assintótica) mínima, (ou seja, eficiência) não só sob a distribuição hipotética mas também sob outras distribuições próximas dela, em um certo sentido; seja Fisher consistente; função de influência limitada; GES e ρ o menor possível; ponto de ruptura próximo de 0.5; seja qualitativamente robusta; etc., e... não seja muito difícil de calcular. É de se suspeitar que uma tal técnica ótima não exista na maioria dos casos práticos. É fácil de ver, por outro lado, que alguns desses critérios são conflitantes entre si, ou seja, se se quer otimalidade com respeito a um critério, se deve sacrificar a otimalidade com respeito a outro. Uma possível forma de resolver este problema seria buscar técnicas que satisfaçam uma certa solução de compromisso entre esses critérios, solução que dependerá de cada problema em particular. Outra opção seria adotar uma escala de prioridades entre critérios, com o caso, Martin e Yohai (1985)) robustez qualitativa, eficiência, robustez min–max e, especificamente para o caso de amostras finitas, alto ponto de ruptura ao mesmo nível de prioridade que robustez qualitativa. Talvez o roteiro de ação mais recomendável do ponto de vista prático seja adotar o sugerido por Hogg (1979) (também em Martin e Yohai (1985)): 1) Efetuar a análise habitual, usando uma técnica clássica; 2) usar um procedimento robusto; 3) se os resultados anteriores coincidirem satisfatoriamente, realizar o relatório de síntese estatística usual; 4) quando houver diferenças substanciais, estudar todo o problema novamente e de forma global, ou seja, considerando aspectos como adequação do modelo, legitimidade das técnicas usadas, qualidade dos dados etc.

Como vemos, a subjetividade, ou melhor, a experiência da equipe envolvida na análise é indispensável. Uma teoria sólida, junto com habilidade prática e bons recursos computacionais são cruciais em toda experiência.

8 – ESTIMADORES ROBUSTOS EM AMOSTRAS UNIVARIADAS

Consideraremos procedimentos de "acomodamento" de *outliers* para o caso de amostras univariadas quando se trabalha com um modelo paramétrico.

Suponhamos que nos interessa conhecer o mais acertadamente possível a média e/ou outros parâmetros da distribuição de uma certa variável Y. Denotaremos esta distribuição por F_Y.

Modelo:

Seja f uma função de densidade de probabilidade cuja forma explícita supomos conhecida e consideramos a distribuição hipotética: F_H com densidade dada por

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$$

para certo μ em R e $\sigma > 0$.

Problema : estimar μ. Às vezes σ se supõe conhecido, para simplificar o estudo de um certo procedimento de estimação de μ.

As observações são realizações, digamos $x_1, \cdots, x_n$, de n variáveis aleatórias independentes e identicamente distribuídas $X_1, \cdots, X_n$ com distribuição comum F_0 que supomos definida por uma densidade f_0.

O ponto de partida da Estatística Paramétrica Clássica é supor que F_H e F_0 coincidem. O método de estimação mais aconselhável deste ponto de vista é o chamado de máxima verossimilhança. Recordemo–lo:

Chama–se "função de verossimilhança baseada em $X_1, \cdots, X_n$
avaliada em (μ, σ)" a :

$$L((\mu,\sigma),(X_1,\cdots,X_n)) = f_{\mu,\sigma}(X_1)\cdots f_{\mu,\sigma}(X_n).$$

Diz–se que as estatísticas $\hat{\mu}$ e $\hat{\sigma}$, baseadas em $X_1, \cdots, X_n$ constituem um estimador de máxima verossimilhança para μ e σ se

$$L((\hat{\mu},\hat{\sigma}),(X_1,\cdots,X_n)) \geq L((\mu*,\sigma*),(X_1,\cdots,X_n))$$

quaisquer que sejam μ^* em R e $\sigma^* > 0$.

São bem conhecidas as propriedades de otimalidade deste estimador sob a suposição de que F_H e F_0 coincidem. Não as repetiremos aqui. O leitor que necessite recordá–las poderá recorrer a qualquer dos textos de cursos clássicos de Estatística Paramétrica.

Também recordemos que se f é a densidade da $\mathbf{N}(0,1)$, então:

$$\begin{aligned} \hat{\mu} = &\quad \text{média amostral de} \quad \{X_1, \cdots, X_n\} = \bar{x} \\ \hat{\sigma} = &\quad \text{desvio padrão de} \quad \{X_1, \cdots, X_n\} \\ &= (\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2)^{1/2} = s. \end{aligned}$$

Tudo vai bem se além da coincidência $F_H = F_0$, se tem $F_\nu = F_H$, coisa de que, honestamente, nunca poderemos estar plenamente seguros. Em outras palavras, se as observações são realmente realizações genuínas daquela variável Y cuja distribuição queremos estudar.

Se não se dão aquelas igualdades os estimadores de máxima verossimilhança podem ser pouco úteis para estimar μ e/ou σ, segundo o que já havíamos comentado nas seções anteriores.

Um primeiro passo para a obtenção de técnicas mais confiáveis será supor que $F_H = F_\nu$ mas que F_0 poderia não ser F_H. Em tal caso é recomendável dar alguma forma de tratamento especial aos *outliers* concebidos como "observações extremas" da amostra, tratando de minimizar sua influência nos estimadores de μ e/ou σ que serão utilizados de fato.

Na Seção 5 vimos que, em geral, existem duas categorias de tratamentos. Uma delas, drástica, formada pelos tratamentos que consistem em identificar os *outliers* e eliminá-los da amostra aplicando a seguir o método de máxima verossimilhança (ou este modificado, ver Tiku e outros (1986)). A outra categoria é formada pelos procedimentos que consistem em "acomodar" os *outliers*. Esta última, por sua vez, podemos considerar dividida em outras duas classes. Uma formada pelos "procedimentos robustos-protetores", a outra classe é formada pelas técnicas construídas pensando em fazer frente aos *outliers* considerados como observações contaminantes. Não existe uma distinção muito nítida entre quais são os de uma classe e quais os de outra, tudo depende do que se faz antes de aplicar um determinado procedimento, mais precisamente, se se aplica ou não previamente uma técnica para identificar observações "contaminantes". Estas técnicas se chamam "testes de discrepância". Veremos alguns deles mais adiante.

Vejamos agora alguns procedimentos robustos que são aplicáveis ao modelo proposto acima.

8.1 – Procedimentos Robustos

Nesta seção nos preocupamos principalmente em estimar o parâmetro de posição μ e não a escala σ. De qualquer modo, na prática, a escala (ou dispersão) deve ser estimada, assim é que ,sem entrar em mais detalhes, consideraremos um estimador de escala de σ , digamos s.

8.1.1 – M-estimadores para μ

Definição: Seja ψ uma função definida sobre R com valores reais, se chama M-estimador de μ definido por ψ , baseado em $X_1, \cdots, X_n$, a uma estatística $M_n = M_n(X_1, \cdots, X_n)$ tal que:

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - M_n}{s}\right) = 0 \tag{8.1}$$

Uma definição alternativa deste estimador pode-se obter como se fosse um estimador de mínimos quadrados ponderados, isto é:

$$\sum_{i=1}^{n} W_i(X_i - M_n) = 0$$

ou equivalentemente:

$$M_n = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

onde

$$W_i = \frac{\psi(\frac{X_i - M_n}{s})}{\frac{X_i - M_n}{s}}$$

De passagem, notemos que esta última forma dá um método de cálculo iterativo da equação implícita que define M_n.

É conveniente que vejamos agora qual é a idéia motivadora desta definição. Já dissemos que a Estatística Paramétrica Clássica sugere o uso dos estimadores de máxima verossimilhança. Consideremos então a "função de verossimilhança baseada em $X_1, \cdots, X_n$ avaliada em (μ, σ)":

$$L((\mu, \sigma), (X_1, \cdots, X_n)) = f_{\mu,\sigma}(X_1) \cdots f_{\mu,\sigma}(X_n)$$

Daí:

$$log(L((\mu, \sigma), (X_1, \cdots, X_n))) = -\sum \rho(X_i, (\mu, \sigma))$$

onde

$$\rho(X_i, (\mu, \sigma)) = -log(f(\frac{X_i - \mu}{\sigma}))$$

Assim, se tem:

$$\sum_{i=1}^{n} \rho(X_i, (\hat{\mu}, \hat{\sigma})) \leq \sum_{i=1}^{n} \rho(X_i, (\mu*, \sigma*)).$$

Agora, sob condições de regularidade bastante gerais tem-se que esta última condição implica:

$$\sum_{i=1}^{n} \psi(\frac{X_i - \hat{\mu}}{\hat{\sigma}}) = 0 \tag{8.2}$$

onde $\psi(t) = \psi'(t) = -\frac{f'(t)}{f(t)}$.

Comparando então esta última equação com a (8.1), vemos que os M–estimadores estão definidos por uma equação similar à que define os estimadores de máxima verossimilhança, resultando daí o nome dado àqueles. Agora, a ψ da (8.1) não tem por que ser a mesma que a de (8.2). Em geral, escolhe-se a ψ que define um M–estimador segundo diversos critérios: facilidade de utilização do ponto de vista matemático, numérico, otimalidade com respeito a certos critérios usados para medir a *performance* de estimadores, experiência etc.

Vejamos alguns exemplos de funções ψ que se usam habitualmente :

1) ψ que define os MV (máxima verossimilhança) no caso normal

$\left(f(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})\right)$

$$\psi(x) = x.$$

Logo, neste caso :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

2) ψ que define os MV no caso de ser f a dupla exponencial ($f(x) = \frac{1}{2}\exp\{-|x|\}$):

$$\psi(x) = sinal(x)$$

Em tal caso:

$$\hat{\mu} = \text{ mediana } (X_1, \cdots, X_n)$$

3) ψ *que define os MV no caso de ser* f a Cauchy
($f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$):

$$\psi(x) = \frac{2x}{1+x^2}$$

Neste caso não existe uma forma explícita para $\hat{\mu}$, deve ser calculado resolvendo a equação (8.2), geralmente por um processo iterativo (ver, por exemplo, Andrews e outros (1972)).

4) ψ tipo Huber :

$$\psi_{H,k}(x) = \begin{cases} -k & \text{se} \quad x < -k \\ x & \text{se} \quad -k \leq x \leq k \\ k & \text{se} \quad x > k \end{cases}$$

5) ψ tipo Hampel:

$$\psi_{H,a,b,c}(x) = \begin{cases} a \cdot (x) & \text{se} \quad |x| \leq a \\ a \text{ sinal } (x) & \text{se} \quad a \leq |x| \leq b \\ a\dfrac{c.\text{sinal}(x) - x}{c-b} & \text{se} \quad b \leq |x| \leq c \\ 0 & \text{se} \quad |x| \geq c \end{cases}$$

6) ψ tipo senóide ou Andrews :

$$\psi_{A,d}(x) = \begin{cases} \text{sen}\left(\dfrac{x}{d}\right) & \text{se} \quad |x| \leq d\pi \\ 0 & \text{se} \quad |x| > d\pi \end{cases}$$

7) ψ tipo biquadrada ou Tukey:

$$\psi_{B,k}(x) = \begin{cases} x\left(1 - \left(\dfrac{x}{k}\right)^2\right)^2 & \text{se} \quad |x| \leq k \\ 0 & \text{se} \quad |x| > k \end{cases}$$

Não entraremos em detalhes sobre as vantagens de usar uma ou outra. Somente destacamos que em robustez se usam as limitadas, como as ψ definidas em 4) a 7) e preferencialmente as chamadas *redescending*, como as 5) a 7). Estas últimas apresentam certos problemas para calcular os M–estimadores definidos por elas devido ao fato de que a solução da equação implícita não é única. É conveniente que recordemos a forma aproximada destas funções cujos gráficos estão na Figura 11.

Na Seção anterior vimos que o conceito de curva de influência era um dos mais importantes dentro da Robustez. Resumidamente, diz–se que um estimador é robusto quando sua função de influência é limitada. Não é difícil de provar que a função de influência dos estimadores tem a mesma forma que a função ψ que os define; de fato , coincide com ela a menos de um fator.

Tendo agora em vista o gráfico das ψ mostrado na Figura 11 vemos quais são os M–estimadores definidos por ψ limitada e dali quais são os preferencialmente usados como robustos.

8.1.2 – M–estimadores ou M–estimadores *one-step* para μ

Definição : Seja ψ uma função definida sobre R com valores reais, chama–se M–estimador de μ definido por ψ , baseado em $X_1, \cdots, X_n$, a uma estatística $m_n = m_n\ (X_1, \cdots, X_n)$ tal que:

$$m_n = m_n^0 + s_n^0 \frac{\sum_{i=1}^n \psi(W_i^o)}{\sum_{i=1}^n \psi'(W_i^0)}$$

onde $W_i^0 = \frac{X_i - m_n^0}{S_n^0}$

Os m_n^0 e s_n^0 são estimadores iniciais para μ e σ, respectivamente.

Foram definidos por Bickel (1975). Observemos que não são outra coisa que os estimadores obtidos depois de aplicar o primeiro passo do processo iterativo que define os M–estimadores, processo que descrevemos sem detalhe mais acima. Apresentam a vantagem sobre os M–estimadores de que são mais rápidos em tempo de computação para seu cálculo, e que não apresentam problemas de unicidade no caso de $\psi's$ do tipo *redescending*. Por outro lado, segundo Hampel e outros (1986), seu comportamento assintótico é similar ao dos M–estimadores, ao menos quando ψ é uma função ímpar e a F_0 é simétrica. Sem dúvida, estes M–estimadores deverão ser usados quando os estimadores iniciais são "deficientemente" robustos. Hampel e outros (1986) recomendam tomar como m_n^0 e s_n^0 a mediana amostral e a mediana do desvio absoluto da mediana (ver Seção 5), respectivamente.

8.1.3 – Outros estimadores

Vários tipos de estimadores têm sido propostos. Por exemplo, em Andrews e outros (1972) se estudam 65 estimadores diferentes para μ. Já dissemos que é impossível obter "o estimador ótimo", devido a que certos critérios conduzem a condições conflitantes. Por isso, sempre será válido conhecer ao menos a existência e as propriedades de estimadores que tratam de atender requisitos diversos de otimalidade. O importante será, em cada caso, captar as idéias subjacentes às diversas propostas. Neste trabalho nos limitaremos a ver somente as definições de alguns dos estimadores mais usados.

8.1.4 – L–estimadores para μ

Definição: Sejam $a_1, \cdots, a_n$ números reais tais que $\sum_{i=1}^{n} a_i = 1$. Chama–se L–estimador de μ definido por $a_1, \cdots, a_n$ baseado na amostra $X_1, \cdots, X_n$:

$$L_n = \sum_{i=1}^{n} a_i X_{(i)} \tag{8.3}$$

onde $X_{(1)}, \cdots, X_{(n)}$ são as estatísticas de ordem de $X_1, \cdots, X_n$.

Dando diversos valores aos $a_i's$ obtêm–se vários estimadores já conhecidos desde muito antes de surgirem os M–estimadores. Por exemplo :

1) Mediana amostral. L–estimador definido por (8.3) assim:
para $n = 2k - 1$ (n ímpar):

$$\begin{aligned} a_k &= 1 \\ a_i &= 0 \quad \text{se} \quad i \neq k \end{aligned}$$

para $n = 2k$ (n par):

$$\begin{aligned} a_k = a_{k+1} &= 1/2 \\ a_i &= 0 \quad \text{para os } i\text{'s diferente de } \quad k \quad \text{ou} \quad (k+1). \end{aligned}$$

2) Média α–truncada. Seja $0 \leq \alpha < 1/2$, chama–se "média–truncada" ao L–estimador definido por (8.3) com :

$$\begin{aligned} a_i &= 1/(n - 2[n\alpha]) \quad \text{se} \quad i = [n\alpha] + 1, \cdots, n - [n\alpha] \\ a_i &= 0 \quad \text{caso contrário} \end{aligned}$$

3) Média α–winsorizada. Seja α como antes, chama–se "média α-winsorizada" ao L-estimador definido por (8.3) com :

$$a_i = 1/n \quad \text{se} \quad i = [n\alpha]+2, \cdots, n-[n\alpha]-1$$
$$a_i = [n\alpha]/n \quad \text{se} \quad i = [n\alpha]+1, \quad \text{ou} \quad i = n-[n\alpha]$$
$$a_i = 0 \qquad \text{caso contrário}$$

Vários outros L–estimadores têm sido sugeridos nos últimos anos. Um deles, que inclusive aparece em alguns textos modernos de Estatística, é o proposto por Gastwirth e Rubin (1969):

4) "tri–média" ou estimador de Gastwirth–Rubin.

$$TRG = \frac{1}{4}(X_{(q^-)}) + 2.\ \text{Mediana}\ (X_1, \cdots, X_n) + (X_{(q^+)})$$

onde $q^- = [n/4]+1$ e $q^+ = n-[n/4]$.

Para facilitar o estudo da teoria assintótica dos L–estimadores, freqüentemente se modificam um pouco as definições dadas anteriormente, obtendo estimadores assintoticamente equivalentes.

8.1.5 – R–estimadores para μ

Esta família é constituída por uma rica variedade de estimadores. Na opinião do autor deste trabalho alguns destes estimadores competem bastante bem com certos M–estimadores, pelo menos no modelo de posição. A idéia que sustenta o uso dos R–estimadores está baseada em técnicas de testes de hipóteses não paramétricas. Constituem assim um bom exemplo da influência mútua que se verifica entre as diferentes partes da Estatística.

Vale a pena deter–se um pouco mais nessa idéia. Depois veremos uma definição formal.

Seja G uma distribuição simétrica em R. Suponhamos que $U_1, \cdots, U_n$ é uma amostra de G e que $V_1, \cdots, V_n$ é uma amostra da distribuição G_μ definida por $G_\mu(t) = G(t-\mu)$. Sejam os valores observados : $u_1, \cdots, u_n$ e $v_1, \cdots, v_n$, respectivamente. Suponhamos que se deseja testar a hipótese de que ambas amostras provêm da mesma distribuição. A Teoria de Testes Não–Paramétricos recomenda o seguinte procedimento (ver detalhes, por exemplo, em: Bickel e Doksum (1977), Hájek e Sidak (1967) ou Lehmann (1975)):

Seja $J : (0,1) \to R$ uma certa função não–decrescente tal que $J(1-t) = -J(t)$. Para cada n seja a_n a função definida sobre $1, \cdots, 2n$ por

$$a_n(k) = J\left(\frac{k}{2n+1}\right) \tag{8.4}$$

e seja:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i)$$

onde R_i é o posto de u_i na amostra combinada $u_1, \cdots, u_n, v_1, \cdots, v_n$.

Se a hipótese nula é correta, então S_n deve estar próximo de zero. Assim, o teste não-paramétrico baseado em S_n rejeita a hipótese nula para valores grandes de S_n.

Vejamos a conexão do que foi apresentado com a estimação de μ. Para simplificar, suponhamos σ conhecida e igual a 1. Se a amostra $x_1, \cdots, x_n$ provêm de F_H (isto é, $F_H = F_O$) e esta é simétrica em torno de μ, então $u_i = (x_i - \mu) i = 1, \cdots, n$ é uma amostra de $f_{0,1}$ que é simétrica em torno de zero, logo, $v_i = -(x_i - \mu) i = 1, \cdots, n$ também será uma amostra de $f_{0,1}$; segue-se que $S_n()$ deverá estar próxima de zero, onde

$$S_n(\mu) = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i). \tag{8.5}$$

Assim, obtém-se um estimador razoável para μ, baseado em $x_1, \cdots, x_n$, é um T_n tal que

$$S_n(T_n) = 0.$$

Resumindo:

Definição: Seja $J : (0,1) \to \mathbf{R}$ uma função não-decrescente tal que $J(1-t) = -J(t)$, a_n como em (8.4) e S_n como em (8.5). Chama-se R-estimador de μ definido por J baseado em $x_1, \cdots, x_n$ a um $\mu(R, J)$ tal que

$$S_n\left(\mu(R,J), (X_1, \cdots, X_n)\right) = 0.$$

ou, equivalentemente:

$$S_n^*\left(\mu(R,J), (X_1, \cdots, X_n)\right) = 0.$$

Onde $S_n^* : \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ é a função definida por:

$$S_n^*(m, (x_1, \cdots, x_n)) = \sum_{i=1}^{n} a_n(R_i^*)$$

sendo R_i^* a amplitude de x_i no conjunto $x_1, \cdots, x_n, 2m - x_1, \cdots, 2m - x_n$.

Utilizando a notação habitual da Teoria de Testes Não-Paramétricos, chama-se *scores* aos $a_i's$ e *geratriz dos scores* à função J.

Diversos R-estimadores podem ser obtidos por meio de diferentes $J's$. Estas funções J, assim como as $\psi's$ dos M-estimadores podem ser definidas usando a distribuição hipotética

das observações (a F_H), mas aqui também é mais proveitoso usar outras funções J, definidas pensando–se em atingir robustez. De todo modo temos, em geral:

R–estimador definido por J através de uma densidade g é o R–estimador definido mais acima com

$$J(t) = -\frac{g'(G^{-1}(t))}{g(G^{-1}(t))},$$

sendo G função de distribuição definida por g.

De acordo com a g que se use nesta última fórmula, obtêm–se diversos R–estimadores. Por exemplo :

R–estimador com *scores* gaussianos (ou normais) é o definido por J através da densidade $\varphi(x)$ da $N(0,1)$, isto é:

$$J(t) = \Phi^{-1}(t)$$

sendo Φ a função de distribuição da $N(0,1)$.

8.1.6 – R–estimador com *scores* definidos pela dupla exponencial

É também chamado R–estimador com *scores* baseados na mediana amostral. É o definido por J através da densidade da dupla exponencial ($f(t) = e^{-|t|}/2$). Pode–se ver fácilmente que neste caso se tem:

$$J(t) = \text{ sinal } \left(t - \frac{1}{2}\right)$$

R–estimador com *scores* definidos pela logística. Chama–se também R–estimador de Hodges–Lehmann ou R–estimador com *scores* de Wilcoxon, aquele definido por J através da densidade da logística ($f(t) = e^{-t}/(1+e^{-t})^2$ para t real). Prova–se que J se reduz a

$$J(t) = t - \frac{1}{2}$$

R–estimador definido através de um funcional. Pode–se ver (por exemplo em Bustos (1986)) que os R–estimadores podem ser definidos assintoticamente pelo seguinte funcional T definido sobre um subconjunto conveniente do conjunto das distribuições sobre R:

$$\int J\left(\frac{G(y) + 1 - G(2T(G) - y)}{2}\right) dG(y) = 0$$

8.1.7 – D–estimadores ou de "distância mínima" para μ

A história desta classe de estimadores, na verdade pouco usada, em que pese suas boas propriedades, remonta ao trabalho de Wolfowitz (1957). A idéia básica sobre a qual se baseiam é a seguinte: quando G_n é a distribuição empírica da amostra, se escolhe como estimador para μ a estatística Tn tal que

$$d(G_n, F_{T_n,*}) \leq d(G_n, F_{\mu^*,\sigma^*})$$

quaisquer que sejam $\mu*$ em R e $\sigma* > 0$, onde $F_{\mu*,\sigma*}$ é a distribuição dada pela densidade $f_{\mu*,\sigma*}$. Aqui d é uma certa medida da discrepância ou "distância" entre distribuições e S é um estimador da escala σ.

8.1.8 – P–estimadores ou de Pitman para μ

São uma generalização dos estimadores de Pitman (ver, por exemplo, Andrews e outros (1972)). Foram definidos por Johns (1979). Seja h uma certa função definida sobre R com valores reais não negativos, chama–se P–estimador definido por h a estatística $T_n == T_n(X_1, \cdots, X_n)$ definida por

$$T_n = \frac{\int u(\prod_{i=1}^n h(X_i - u))du}{\int \prod_{i=1}^n h(X_i - u)du}$$

Johns provou que estes estimadores podem ser definidos assintoticamente pelo funcional T que satisfaz :

$$\int (h'(u - T(G))/h(u - T(G)))dG(u) = 0$$

onde G toma valores sobre um certo subconjunto de distribuições.

8.1.9 – S–estimadores para μ

Foram definidos por Rousseeuw e Yohai (1983). Estão baseados na minimização de uma estatística que pode ser usada como estimador da escala. Seja h uma função não negativa e limitada tal que :

$$K = \int h(t)dF_H(t) < \infty$$

Suponhamos que para cada μ^* em R o conjunto

$$SS(\mu^*) = \left\{ s > 0 : \frac{1}{n} \sum_{i=1}^{n} h(\frac{X_i - \mu^*}{s}) = K \right\}$$

é não vazio, seja então $\sigma(\mu^*)$ tal que:

$$\sigma(\mu^*) \leq \quad s \quad \forall \quad s \in SS(\mu^*)$$

Se h é escolhida convenientemente, o conjunto $SS(\mu^*)$ terá só um ponto mínimo e assim $\sigma(\mu^*)$ estará bem definida. Neste caso, chama-se S-estimador para μ definido por h baseado em $X_1, \cdots, X_n$ a estatística T_n tal que

$$\sigma(T_n) \leq \sigma(\mu^*) \quad \forall \quad \mu^* \in \mathbb{R}$$

A seguir, pode definir-se um novo estimador de escala, simplesmente:

$$S = \sigma(T_n).$$

Na verdade, estes estimadores foram introduzidos visando a otimalidade com respeito ao "ponto de rutura", no modelo de regressão. É neste modelo que se manifestam de forma mais notável suas excelentes propriedades.

8.1.10 – W–estimadores ou estimadores de mínimos quadrados ponderados para μ

Foram introduzidos por Tukey (1970-71) como a média ponderada das observações. Isto é, chama-se W–estimador para μ com ponderações (pesos) $w_1, \cdots, w_n$ baseado em $X_1, \cdots, X_n$ a estatística T_n dada por:

$$T_n = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Os pesos podem ser tomados como dependentes das observações através de uma certa função W:

$$W_i = W(X_i - T_n).$$

Se for necessário introduzir um estimador, digamos S, para a escala, recomenda-se tomar a "mediana dos desvios absolutos com relação à mediana" (ver Seção 5).

Finalmente, T_n deve satisfazer à seguinte equação:

$$T_n = \frac{\sum_{i=1}^{n} X_i w(\frac{X_i - T_n}{S})}{\sum_{i=1}^{n} w(\frac{X_i - T_n}{S})}$$

que habitualmente se resolve através de um processo iterativo iniciado a partir de um estimador robusto, que costuma ser a mediana das observações.

O correspondente W–estimador do tipo *one–step* é muito usado devido à sua facilidade de cálculo, sendo sua *performance* razoavelmente boa.

8.1.11 – LMS–estimador (*least median square*) para μ

Definição: Chama–se LMS–estimador para a estatística T_n tal que

$$\text{Med} \quad (r_1(T_n)^2, \cdots, r_n(T_n)^2) \leq \text{Med} \quad (r_1(\mu^*)^2, \cdots, r_n(\mu^*)^2)$$

para todo μ^*, onde: Med significa "mediana" e $r_i(\mu^*) = (X_i - \mu^*)/S$, para todo $i = 1, \cdots, n$.

Do mesmo modo que os S–estimadores (dos quais é um caso particular) suas propriedades mais notáveis se evidenciam no problema de estimação no modelo de regressão. Aparece definido e estudado em Rousseeuw (1984). Um programa de computador para seu cálculo é o principal objetivo do já citado **PROGRESS**. Para o caso aqui considerado (modelo univariado), é analisado em Andrews e outros (1972), com o nome de *short*, um "parente" próximo deste estimador.

A idéia em que se baseia é a de minimizar a mediana dos resíduos ao quadrado, em vez de minimizar a soma de tais resíduos ao quadrado que, como sabemos, é o que faz a média amostral no modelo univariado e o estimador de mínimos quadrados no modelo de regressão. A propriedade mais notável do LMS–estimador é seu excelente comportamento com respeito ao "ponto de ruptura".

8.1.12 – MML–estimadores (*modified maximum likelihood*) para μ

Não veremos aqui sua definição. O leitor interessado poderá vê–la em Tiku e outros (1986) onde se faz uma cuidadosa e exaustiva análise do mesmo. Pode–se ver ali que ele tem propriedades de robustez bastante boas e seu cálculo não é muito difícil, ao menos quando se supõe que F_H é a distribuição normal e a amostra não é de tamanho muito grande.

Certamente a lista apresentada aqui não esgota a plêiade de estimadores que se pode encontrar na literatura estatística.

Apenas destacamos os que são mais usados ou os que parecem ter propriedades boas com respeito a vários critérios.

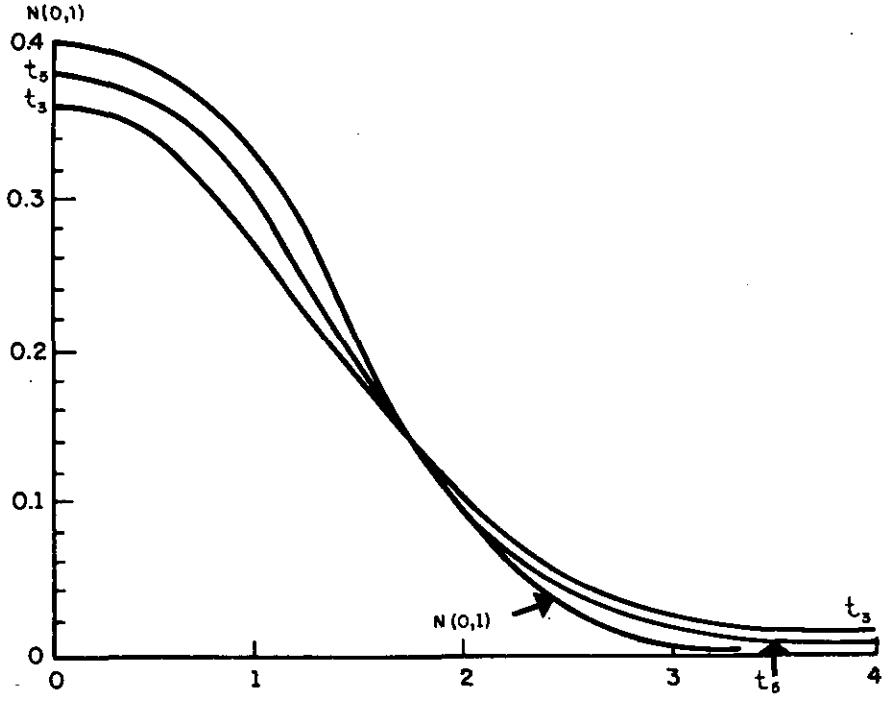


Figura 10

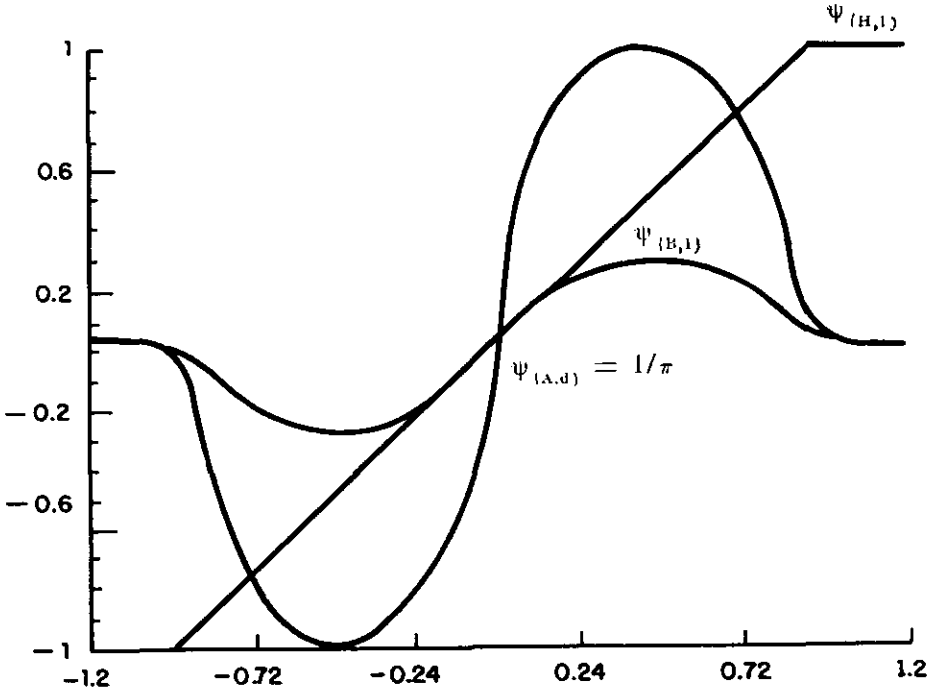


Figura 11

BIBLIOGRAFIA

ANDREWS, D. F. et alii (1972) *Robust estimates of location: survey and advances.* Princeton, New Jersey, Princeton University Press.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, (70): 428 – 434.

BICKEL, P. J., & DOKSUM, K. A. (1977), *Mathematical statistics*, San Francisco, California, Holden-Day.

BMDP. Statistical software (1981). Los Angeles, California, University of California Press.

BOOS, D. D. & SERFLING, R. J. (1980). A note on differentials and the CLT and LIL for statistical functions, with Application to M-estimates. *Ann. Statist.*, (8): 618 – 624.

BUSTOS, O. H. (1986). Algumas idéias de robustez aplicadas à estimação paramétrica em séries temporais. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 7, 1986, Campinas. Anais... Campinas: 1986.

DONOHO, D. L. & HUBER, P. J. (1983). *The notion of breakdown point.* In: A festschrift for Erich L. Lehmann. Belmont, California, Wadsworth.

FERNHOLZ, L. T. 1983). *Von mises calculus for statistical functionals.* New York, Springer-Verlag. (Lecture notes in statistics, 19).

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London, Ser. A.* (22): 309 – 368.

GASTWIRTH, J. L. & RUBIN, H. (1969). On robust linear estimators. *Ann. of Math. Stat.*. (40): 24 – 39.

GRAF, H. P. (1983). *Long-range correlations and estimation of the self-similarity parameter.* Zurich, ETH. Ph. D. Dissertation.

HAJEK, J. E. & SIDAK, Z. (1967). *Theory of rank tests.* New York, Academic Press.

HAMPEL, F. R. (1968). *Contributions to the theory of robust estimation.* California, University of Berkeley. Ph. D. Dissertation.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Annals of Math. Stat.*, (42): 1887 – 1896.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *JASA*, (69): 383 – 393.

HAMPEL, F. R. et alii (1986). *Robust statistics, the approach based on influence functions.* New York, Wiley.

HILL, M. & DIXON, W. J. (1982). Robustness in real life: A Study of clinical laboratory data. *Biometrics*, (38): 377 – 396.

HOGG, R. V. (1979). Statistical robustness: one view of its use and applications today. *The American statistician*, (33): 108 – 115.

HUBER, P. J. (1981). *Robust statistics*, New York, Wiley.

JAMES, K. L. & BUSTOS, O. H. (1980). *Procedimentos robustos.* Rio de Janeiro, Simpósio Nacional de Probabilidade e Estatística.

JOHNS, M. V. (1979). *Robust Pitman-like estimators, in robustness in statistics.* New York, Academic Press. p. 49–60.

LEHMANN, E. L. (1975). *Non-parametrics: statistics methods based on ranks.* San Francisco, Holden-Day.

MALLOWS, C. L. (1975). *On some topics in robustness.* Murray Hill, New Jersey, Bell Labos. (Tech. memo).

MARTIN, R. D. & YOHAI, V. J. (1985). *Robustness in time series and estimating ARMA models.* In: HANDBOOK OF STATISTICS. New York, Elsevier Sciencs Publisher. v. 5.

ROCKE, D. M. et alii (1982). Are robust estimators really necessary?. *Technometrics*, (24): 95 – 101.

ROUSSEEUW, P. J. & YOHAI, V. J. (1983). *Robust regression by Means by means of S-estimators*, in: ROBUST AND NON -LINEAR TIME SERIES ANALYSIS. New York, Springer-Verlag. (Lecture notes in statistics, 26).

SHAPIRO, S. S. & WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, (52): 591 – 611.

STIGLER, S. M. (1977). Do robust estimators work with real data" *Ann. Statist.* (5): 1055 – 1098.
TIKU, M. L. et alii (1986). *Robust inference*. New York, Marcel Dekker.
TUKEY, J. W. (1970, 1971). *Exploratory data analysis*. Reading, Massachussetts, Addison-Wesley.
WOLFOWITZ, J. (1957). The minimax distance method. *Ann. of Math. Stat.*, (28): 75 – 88.

ESTUDO DO PONTO DE CORTE EM UMA REGRA DE CLASSIFICAÇÃO PARA DADOS BINÁRIOS E UMA APLICAÇÃO EM PNEUMOLOGIA

Beatriz Mendes Luz*

1 – INTRODUÇÃO

Um paciente dá entrada em um laboratório de testes de função respiratória para se submeter a exames. O motivo de sua vinda pode ser desde uma crise de asma até simples rotina pré-operatória. Os médicos, após os exames, e baseados em uma regra de classificação construída a partir de uma amostra de treinamento, decidem se o paciente é sadio ou portador de algum tipo de doença respiratória. Problemas deste tipo, onde o resultado é binário, ocorrem freqüentemente em Medicina.

O Modelo Logístico é um modelo apropriado a este tipo de dados e é comum o uso de uma regra de classificação obtida a partir da Regressão Logística.

Este trabalho, baseado em uma tese de mestrado defendida no IMPA, Instituto de Matemática Pura e Aplicada, em 1988, pesquisa um ponto de corte ótimo a ser usado nesta regra de classificação. Para isso, através de simulações e de um conjunto de dados relacionados com o problema sugerido no início desta Introdução, fazemos uma comparação entre três pontos de corte: dois deles bastante conhecidos e o terceiro uma nova proposta, obtida a partir do estudo de um estimador para o vício da taxa aparente de erro da regra de classificação.

* Mestre em Estatística. A autora agradece a sua orientadora Kang Ling James, ao professor Oscar Bustos pela ajuda computacional e a dois pareceristas da RBEs.

Um estudo mais detalhado da variável aleatória C_0, ponto de corte definido por esse terceiro procedimento, incluindo a construção de intervalos de confiança e aplicação de testes, será desenvolvido futuramente. Por exemplo, um numero maior de simulações, usando-se técnicas de redução de variância, poderia ser utilizado.

1.1 – Regressão Logística

1.1.1 – O Problema

Consideremos um experimento onde temos n variáveis respostas independentes:

$$\begin{aligned} y = 1 \quad & \text{com probabilidade} \quad \pi_i \\ y = 0 \quad & \text{com probabilidade} \quad 1 - \pi_i \end{aligned}$$

e p variáveis explanatórias consideradas não aleatórias, qualitativas ou quantitativas:

$$x_1, x_2, \cdots, x_p.$$

Queremos relacionar a probabilidade de sucesso $\pi_i = E(y_i)$ e as p variáveis

$$x_1, x_2, \cdots, x_p.$$

Este tipo de problema, onde temos apenas dois resultados de interesse, ocorre especialmente em Medicina. A resposta binária, geralmente chamada de sucesso ou fracasso e codificada como 1(um) ou 0(zero), pode representar, por exemplo, sobrevivência ou morte após determinado tratamento, presença ou ausência de alguma enfermidade. As variáveis explanatórias representariam características dos indivíduos tais como idade, sexo, resultados de exames de laboratório etc.

O modelo linear geral supõe:

$$y = \Pi\alpha + \varepsilon$$

onde $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ é o vetor aleatório , Π uma matriz $n \times (p+1)$ de constantes conhecidas, $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_p)'$ o vetor de parâmetros desconhecidos e $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$ o vetor de erros. Sabe-se, entretanto, que este não é um modelo apropriado para a situação descrita acima. Uma análise das violações às suposições do modelo pode ser vista em Cox (1970), cap. 2, e uma forma de contornar os problemas será discutida na próxima seção.

1.2 – O Modelo Logístico Linear

Uma das maneiras de se representar a dependência da probabilidade de sucesso π e as p variáveis explanatórias impondo a condição:

$$0 < \pi < 1,$$

é supor:

$$\pi_i = \exp(\mathbf{t}_i'\alpha)[1 + \exp(\mathbf{t}_i'\alpha)]^{-1}, \quad i = 1, 2, \cdots, n, \tag{1.1}$$

onde $\mathbf{t}_i = (t_{i0}, t_{i1}, \cdots, t_{ip})' = (1, x_1, x_2, \cdots, x_p)'$ é um vetor de constantes conhecidas e α o vetor de parâmetros a determinar.

Equação (1.1) acima equivale a definir, $\forall \quad i = 1, 2, \cdots, n$, a transformação logística:

$$\lambda_i = \log\left[\pi_i(1 - \pi_i)^{-1}\right] = \mathbf{t}_i'\alpha = \sum_{s=0}^{p} t_{is}\alpha_s$$

ou:

$$\lambda = \Pi\alpha.$$

Para estimar o vetor de parâmetros α usaremos o método da Máxima Verossimilhança, através do programa Logistic Regression do pacote computacional BMDP.

2 – UMA REGRA DE CLASSIFICAÇÃO

Obtidos os estimadores de Máxima Verossimilhança $\hat{\alpha}$:

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_p)'$$

através da Regressão Logística, podemos obter para cada indivíduo, por substituição em (1.1), uma estimativa de sua probabilidade de sucesso:

$$\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_n)'.$$

Assim, se temos uma nova observação $\mathbf{t}_0$, podemos obter $\hat{\pi}_0$ a partir de (1.1) e usar essa estimativa para predizer se o indivíduo cujas medições são dadas por $\mathbf{t}_0$ será "sucesso" ou "fracasso". Temos, então, a seguinte regra de classificação:

* classifique como sucesso se $\quad \hat{\pi}_0 > C_0$
* classifique como fracasso se $\quad \hat{\pi}_0 \leq C_0$ (2.1)

onde C_0 é conhecido como ponto de corte, e $0 < C_0 < 1$.

2.1 – Uma medida de erro de predição

Gostaríamos agora de, a partir de algum critério, investigar a qualidade dessa regra de classificação construída.

Uma medida natural é a taxa aparente de erro, que consiste na proporção de casos erroneamente classificados na amostra que usamos para estimar os parâmetros e que, daqui para frente, chamaremos de amostra de treinamento. A taxa aparente de erro, denotada por $\overline{err}$, é então a proporção de casos erroneamente classificados na amostra de treinamento.

Podemos a partir de $\hat{\pi}$ obter o vetor de predições

$$\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2, \cdots, \hat{\eta}_n)'$$

da seguinte maneira:

$$\begin{aligned} \hat{\eta}_i = 1 \quad &\text{se} \quad \hat{\pi}_i > C_0 \\ \hat{\eta}_i = 0 \quad &\text{se} \quad \hat{\pi}_i \leq C_0 \qquad i = 1, \cdots, n. \end{aligned} \tag{2.2}$$

A taxa aparente de erro é então:

$$\overline{err} \equiv \sharp\{y_i \neq \hat{\eta}_i\}/n \tag{2.3}$$

onde $\sharp A$ denota a cardinalidade do conjunto A.

Percebemos agora que é crucial a escolha do ponto de corte. O uso de $C_0 = 0.5$ é comum e, na Seção 2.2, veremos por quê. Outro procedimento bastante conhecido, que veremos em 2.3, é o de se usar a taxa aparente de erro para definir o ponto de corte. Na Seção 2.4 proporemos um novo procedimento para se determinar o valor de C_0 .

2.2 – O ponto de corte 0.5

Um dos motivos da popularidade do uso de $C_0 = 0.5$ deve-se ao fato de que, quando as $\mathbf{x}$ têm distribuição normal multivariada, usar regra (2.1) com ponto de corte 0.5 equivale, conforme pode ser visto em Mendes Luz(1988), a usar a regra de classificação:

$$\begin{aligned} &\text{classifique como sucesso se} \quad \log\left\{[p_1(\mathbf{t}_0)][p_2(\mathbf{t}_0)]^{-1}\right\} > 0 \\ &\text{classifique como fracasso se} \quad \log\left\{[p_1(\mathbf{t}_0)][p_2(\mathbf{t}_0)]^{-1}\right\} \leq 0 \end{aligned} \tag{2.4}$$

onde $p_1(\mathbf{t}_0) \equiv pr\{y_0 = 1 \mid \mathbf{t}_0\}$ e $p_2(\mathbf{t}_0) \equiv pr\{y_0 = 0 \mid \mathbf{t}_0\}$

Isto é: $\log\{[p_1(t_0)][p_2(t_0)]^{-1}\}$ é o logaritmo da razão das probabilidades *a posteriori* de y pertencer à população 1 e à população 2, tendo sido observado t_0 .

A regra de classificação (2.4), que também é regra que minimiza a probabilidade total de classificação errônea (ver Seber(1984, cap. 6)), é chamada em Efron(1975) de Procedimento *Normal de Classificação*.

Conforme pode ser visto também em Seber(1984,cap.6), se as proporções de observações nos dois grupos são iguais, usar a regra (2.4) equivale a usar a Função Linear Discriminante de Fisher.

2.3 – O uso da taxa aparente de erro para definir C_0

Na tentativa de se encontrar um ponto de corte diferente de 0.5 é bastante comum um procedimento que consiste em se procurar dentre as probabilidades estimadas $\hat{\pi}_i$ aquela que, usada como ponto de corte, produza um menor valor para a taxa aparente de erro.

A justificativa para este procedimento reside no fato de que podemos interpretar a taxa aparente de erro como um estimador da probabilidade total de classificação errônea:

$$PTCE \equiv P(2 \mid 1)p_1 + P(1 \mid 2)p_2, \tag{2.5}$$

onde $P(i \mid j)$ com $i, j = 1, 2, i \neq j$, é a probabilidade de cometer o erro de classificar como pertencente ao grupo i uma observação do grupo j e p_i a proporção de observações no grupo i. Assim, esse estimador poderia ser escrito como:

$$\widehat{PTCE} = (r_1/n_1)\hat{p}_1 + (r_2/n_2)\hat{p}_2,$$

onde r_i seria o número de observações incorretamente classificadas no grupo i, n_i o número de observações no grupo i, $i = 1, 2$, e portanto r_i/n_i seria a proporção de observações do grupo i incorretamente classificadas. Tomando-se $\hat{p}_i = n_i/(n_1 + n_2)$ como os estimadores intuitivos de p_i , obtemos:

$$\widehat{PTCE} = (r_1 + r_2)/(n_1 + n_2) = \overline{err} \tag{2.6}$$

e portanto, minimizar (2.6) equivaleria a minimizar (2.5).

2.4 – A proposta alternativa para definir o ponto de corte

A taxa aparente de erro, procedimento conhecido como "contar os erros", é um estimador geralmente otimista para a verdadeira taxa de erro. Isto se deve ao fato de que os dados

foram usados duas vezes: a primeira para ajustar o modelo e a segunda para verificar sua adequacidade. Definimos o otimismo da taxa aparente de erro como:

$$op(y,\pi) \equiv Err - \overline{err}, \tag{2.7}$$

onde Err representa a verdadeira taxa de erro, isto é, o erro médio de predição para um novo vetor aleatório y^* com a mesma distribuição, porém independente do vetor y que gerou $\hat{\eta}$. A expressão (2.7) acima representa o vício da taxa aparente de erro para um determinado vetor de predições $\hat{\eta}$. Tomando-se a esperança de (2.7) obtemos o otimismo esperado para a regra de classificação $y \longrightarrow \hat{\eta}$, que denotaremos por:

$$\omega(\pi) = E_\pi \left[op(y,\pi)\right], \tag{2.8}$$

Vários estudos têm sido feitos visando-se obter um estimador não viciado para a verdadeira taxa de erro. Dentre as várias propostas existentes destacamos:

1º estimador obtido através do método *cross-validatio*, (Para um exemplo do método veja Efron (1986))

2º aquele obtido após a divisão da amostra em duas partes (Para uma discussão sobre o assunto veja Lachenbruch (1975, cap. 2)) e

3º estimador de máxima verossimilhança proposto em Efron(1986).

A nossa nova proposta de ponto de corte surgiu do estudo desse último estimador. Na verdade, Efron obtém o estimador de máxima verossimilhanca da verdadeira taxa de erro, denotado por $\widehat{Err}$, a partir de:

$$\widehat{Err} = \overline{err} + \widehat{\omega(\pi)},$$

onde $\widehat{\omega(\pi)}$ representa o estimador de máxima verossimilhança de $\omega(\pi)$.

Conforme pode ser visto em Efron(1986), desenvolvendo a expressão (2.8) acima e particularizando-a para o caso de dados binários, regra de classificação (2.1), e medida de erro de predição (2.3), chegamos à expressão:

$$\omega(\pi) = \frac{2}{n}\sum_{i=1}^{n} \pi_i(1-\pi_i)\phi\left(c_i d_i^{-1/2}\right) d_i^{1/2} \tag{2.9}$$

onde ϕ é a densidade da Normal(0,1),

$$c_i = \log\left[C_0(1-C_0)^{-1}\right] - t_i'\alpha, \tag{2.10}$$

$$d_i = var(t_i'\alpha) = t_i'\Sigma^{-1}t_i \tag{2.11}$$

e

$$\Sigma^{-1} = \sum_{i=1}^{n} \pi_i(1-\pi_i)t_i t_i' \tag{2.12}$$

Substituindo-se os estimadores de máxima verossimilhança $\hat{\alpha}$, obtidos através da regressão logística, em (1.1) e onde for necessário nas expressões (2.10), (2.11) e (2.12) acima, obtemos:

$$\widehat{\omega(\pi)} = \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i)\phi\left(\hat{c}_i \hat{d}_i^{-1/2}\right)\hat{d}_i^{1/2}, \tag{2.13}$$

o estimador de máxima verossimilhança de $\omega(\pi)$, com as usuais propriedades assintóticas ótimas dos estimadores de máxima verossimilhança.

É fácil ver que, se um conjunto de dados está bem classificado, a taxa aparente de erro será pequena e, como $\overline{err}$ estima a verdadeira taxa de erro, minimizar $\overline{err}$ implicaria maximizar $\widehat{\omega(\pi)}$. Realmente, durante o desenvolvimento de (2.9) temos a seguinte passagem:

$$\omega(\pi) = \frac{2}{n}\sum_{i=1}^{n} cov(y_i, \hat{\eta}_i)$$

que mostra que quanto mais corretas as predições maior o vício esperado.

Para termos então este novo valor para C_0 basta procurar dentre as probabilidades estimadas $\hat{\pi}_i$ aquela que, usada como ponto de corte em (2.10), maximiza (2.13). A idéia desse novo procedimento para determinar o valor de C_0 surgiu da observação do comportamento do estimador $\widehat{\omega(\pi)}$ durante simulações efetuadas para verificar a sua *performance* como estimador do vício de $\overline{err}$ em amostras pequenas, conforme veremos na próxima seção.

3 – AS SIMULAÇÕES

3.1 – O Primeiro Experimento Amostral

Conforme já dissemos, o processo de simulação que descreveremos a seguir foi montado com o propósito de verificar a performance de $\widehat{\omega(\pi)}$ em amostras pequenas.

Para tanto, suponha que o vetor aleatório x possa pertencer a uma dentre duas populações normais p-dimensionais diferindo em média mas não em covariâncias:

$$\begin{aligned} x &\sim \mathbf{N}_p(\mu_1, \Sigma) \quad \text{com probabilidade} \quad p_1 \\ x &\sim \mathbf{N}_p(\mu_2, \Sigma) \quad \text{com probabilidade} \quad p_2 = 1 - p_1 \end{aligned} \tag{3.1}$$

Se os parâmetros $p_1, p_2 = 1 - p_1, \mu_1, \mu_2$ e Σ são conhecidos, podemos usar alguma regra de classificação (como, por exemplo, a Função Linear Discriminante de Fisher) para alocar um novo vetor x_0 a uma das duas populações.

Na prática, entretanto, os parâmetros $p_1, p_2 = 1 - p_1, \mu_1 \mu_2$ e Σ são desconhecidos e usamos um conjunto, que chamamos de amostra de treinamento : $(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)$ independentes, onde y_i indica de qual população x_i é proveniente, de tal forma que:

$$\begin{aligned} y_i &= 1 \quad \text{com probabilidade} \quad p_1 \\ y_i &= 0 \quad \text{com probabilidade} \quad p_2 = 1 - p_1 \end{aligned} \tag{3.2}$$

e

$$x_i \mid y_i \sim \mathbf{N}_p \left(\mu(y_i), \Sigma\right) \tag{3.3}$$

onde $\mu(y_i) = \mu_1$ se $y_i = 1$ e $\mu(y_i) = \mu_2$ se $y_i = 0$. Assim, condicionalmente aos valores $x_1, x_2, \cdots, x_n$, as y_i são variáveis aleatórias binárias independentes e tais que:

$$Pr\left(y_i = 1 \mid x_i\right) \equiv p_1(x_i) \equiv \exp\left(\alpha_0 + \alpha' x_i\right) \left[1 + \exp\left(\alpha_0 + \alpha' x_i\right)\right]^{-1} \equiv \pi_i \tag{3.4}$$

Como vimos na Seção 1, a regressão logística estima α_0 eα a partir dos valores $(y_i,\ x_i)$, $i = = 1, \cdots, n$, assumindo o modelo (3.4) e maximizando a função de verossimilhança. Conforme provado em Mendes Luz(1988), neste caso, o verdadeiro valor de α é:

$$\begin{aligned} \alpha_0 &= -\log \frac{p_2}{p_1} - \frac{1}{2}\left(\mu_1 \Sigma^{-1} \mu_1 - \mu_2 \Sigma^{-1} \mu_2\right) \\ \alpha' &= \left(\mu_1 - \mu_2\right)' \Sigma^{-1} \end{aligned} \tag{3.5}$$

Para obtermos os nossos dados realizamos 100 ensaios onde, em cada um deles, eram gerados 20 vetores independentes (y_i, x_i), tais que:

$$\begin{aligned} y_i &= 1 \quad \text{com probabilidade} \quad p_1 \\ y_i &= 0 \quad \text{com probabilidade} \quad 1 - p_1 \end{aligned}$$

e

$$x_i | y_i \sim \mathbf{N}_2 \left[(y_i - \frac{1}{2}, 0), I_2\right] \tag{3.6}$$

Temos assim um caso particular de (3.3) onde $p = 2, \mu_1 \in \mathbf{R}^2$, $\mu_2 \in \mathbf{R}^2$, Σ é I_2. De acordo com a nossa notação anterior o modelo (3.4) fica:

$$p_1(x_i) = Pr(y_i = 1 | x_i) = \left[1 + \exp\left(-\alpha' t_i\right)\right]^{-1} \tag{3.7}$$

onde $t_i = (1, x_i)' \in \mathbf{R}^3$ e $\alpha' = (\alpha_0, \alpha_1, \alpha_2)$.

Em cada ciclo, cada um dos 20 vetores (y_i, x_i) era obtido gerando-se, primeiramente, um número H segundo uma distribuição uniforme em [0,1], e usando-se a sub-rotina GGUBFS do IMSL.. Em seguida são gerados dois números segundo uma distribuição Normal(0,1): X1 e X2 usando-se a sub-rotina GGNQF do IMSL. Define-se então, para cada $i = 1, 2, \cdots, 20$, valor da variável y_i segundo o critério:

$$y_i = 1 \quad \text{se} \quad H \in [0, p_1]$$
$$y_i = 0 \quad \text{se} \quad H \notin [0, p_1]$$

e os valores de $x_i = (x_{1i}, x_{2i})$

$$x_{1i} = X_1 + y_i - 0.5$$
$$x_{2i} = X_2$$

Pode-se mostrar, através do teorema de Bayes, que o valor da verdadeira probabilidade de sucesso é dado por:

$$\pi_i = Pr[y_i = 1 | x_i = (x_1, x_2)] = \left[1 + \left(\frac{1}{p_1} - 1\right) \exp(-x_1)\right]^{-1} \tag{3.8}$$

Nesta primeira simulação geramos os dados usando $p_1 = p_2 = 0.5$, assim, como feito em Efron(1986) (Mais tarde obteremos novos dados alterando o valor da *priori* p_1 para 0.3 e depois para 0.8). É fácil ver que o verdadeiro valor de α, segundo fórmula (3.5), é (0,1,0)'.

Usando o estimador de máxima verossimilhança $\hat{\alpha}$ obtido através do programa Logistic Regression do BMDP, a regra (2.2) e o ponto de corte $C_0 = 0.5$ calculamos, para cada um dos 100 ensaios, a taxa aparente de erro (2.3), o estimador $\omega(\hat{\pi})$ (2.13), o estimador de máxima verossimilhança $\hat{Err}$ e a verdadeira taxa de erro Err a partir de:

$$Err = \frac{1}{n}\sum_{i=1}^{n} \pi_i - \frac{2}{n}\sum_{i=1}^{n} \pi_i \hat{\eta}_i + \frac{1}{n}\sum_{i=1}^{n} \hat{\eta}_i$$

cuja dedução se enconta em Mendes Luz(1988).

A Tabela 1 abaixo mostra os resultados. Nas primeiras duas colunas estão a verdadeira taxa de erro e a taxa aparente de erro. A coluna 5 mostra o estimador de máxima verossimilhança $\widehat{\omega(\pi)}$ que demonstra ser um excelente estimador de $\omega(\pi) = 0.34856 - 0.27708 = 0.07148$.

Tabela 1

Primeiros 10 ensaios do Exp. Amostral 1 e resumo p/100

Ensaio	Err	$\overline{err}$	$\widehat{Err}$	$\widehat{\omega(\pi)}$
1	0.2808	0.40	0.4984	0.0984
2	0.3337	0.30	0.3998	0.0998
3	0.3140	0.35	0.4530	0.1030
4	0.3164	0.25	0.3165	0.0665
5	0.2885	0.05	0.0890	0.0390
6	0.3173	0.50	0.5974	0.0974
7	0.3113	0.20	0.2561	0.0561
8	0.3286	0.30	0.4071	0.1071
9	0.3887	0.35	0.4452	0.0952
10	0.3422	0.35	0.4650	0.1150
100 ensaios				
média	0.3486	0.2771	0.3569	0.07985
sd	(0.05523)	(0.09595)	(0.11237)	(0.02040)
coef. var.	0.158	0.346	0.322	0.255

Durante essa primeira fase do trabalho, nosso objetivo era apenas o de verificar se $\widehat{\omega(\pi)}$ é realmente bom para corrigir o otimismo de $\overline{err}$, o que, segundo Efron(1986), seria o uso mais óbvio do estimador. De fato, isto se verifica já que $\widehat{Err} = 0.3569$, apresentando inclusive menor coeficiente de variação.

Entretanto, observando os valores obtidos para o estimador, notamos que ele assumia sempre um valor máximo quando o valor de C_0 era próximo da probabilidade *a priori* de y ser sucesso: $p_1 = 0.5$. O gráfico 1 mostra o relacionamento de $\widehat{\omega(\pi)}$ e C_0 ,C_0 assumindo os valores $\hat{\pi}_i, i = 1, 2, \cdots, 20$, para um dos conjuntos de dados do experimento amostral 1.

Efetuamos então o procedimento descrito na Seção 2.4: em cada uma das 100 amostras e procuramos dentre as estimativas $\hat{\pi}_i$ aquela que, usada como ponto de corte, produzia um valor máximo para $\widehat{\omega(\pi)}$. O ponto de corte médio assim definido foi:

$$C_0 = 0.49735$$

com desvio padrão 0.1036, ao passo que aquele definido pela taxa aparente de erro mínima foi 0.49242 com maior desvio padrão, 0.11418, mas também muito próxima de 0.5.

Gostaríamos agora de analisar um pouco a expressão (2.13), procurando explicar a boa *performance* desse novo procedimento, a partir da pesquisa da observação j cuja estimativa $\hat{\pi}_j$, usada como ponto de corte, definiu o máximo de $\widehat{\omega(\pi)}$. Para tanto, consideremos a seguir a fórmula do estimador:

$$\widehat{\omega(\pi)} = 2/n \sum_{i=1}^{n} \hat{\chi}_i \phi\left(\hat{c}_i d_i^{-1/2}\right) \hat{d}_i^{1/2}$$

Supondo que as quantidades $\hat{d}_i = \hat{Var}(t_i'\hat{\alpha}) = t_i'\Sigma^{-1}t_i$ e $\hat{\chi}_i = \hat{\pi}_i(1-\hat{\pi}_i)$ não variem muito, e sabendo que as parcelas são todas positivas, é de se esperar o valor máximo da soma quando os argumentos da densidade da normal forem todos em torno de zero.

Vejamos a fórmula de $\hat{c}_i$:

$$\hat{c}_i = \log \frac{C_0}{1-C_o} - t_i'\hat{\alpha}$$

Para se determinar o máximo de $\widehat{\omega(\pi)}$ calcula-se para todo $j = 1, \cdots, n$ um $\widehat{\omega(\pi)} = \omega(\hat{\pi}, C_0)$ que usa como ponto de corte a probabilidade de sucesso estimada $\hat{\pi}_j$ correspondente. Assim, para cada $j = 1, \cdots, n$ os $\hat{c}_i$, $\quad i = 1, \cdots, n$, são:

$$\begin{aligned} \hat{c}_i &= \log \frac{\hat{\pi}_j}{1-\hat{\pi}_j} - t_i'\hat{\alpha} \\ &= t_j'\hat{\alpha} - t_i'\hat{\alpha} \end{aligned} \tag{3.9}$$

Sabemos que essas diferenças serão, em módulo, as menores possíveis quando $t_j'\ \hat{\alpha}$ for aquele mais próximo da média dos $t_j'\ \hat{\alpha}, i = 1, \cdots, n$.

Quando as x são normais p-variadas o teorema de Bayes nos dá:

$$\log \frac{p_1(x)}{p_2(x)} = -\log \frac{p_2}{p_1} - \frac{1}{2}\left[(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)\right] + (\mu_1-\mu_2)'\Sigma^{-1}x$$

$$t_j'\alpha \equiv \log \frac{\pi_j}{1-\pi_j} \equiv \log \frac{p_1(x_j)}{p_2(x_j)} = \alpha_0 + \alpha' x_j$$

$$= -\log \frac{p_2}{p_1} - \frac{1}{2}\left[(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)\right] + (\mu_1-\mu_2)'\Sigma^{-1}x_j$$

Então

$$\begin{aligned} c_i &= (\mu_1-\mu_2)'\Sigma^{-1}x_j - (\mu_1-\mu_2)'\Sigma^{-1}x_j \\ &= u_j - u_i, \end{aligned} \tag{3.10}$$

onde indentificamos $u_i = (\mu_1-\mu_2)'\Sigma^{-1} \quad x_i = \Lambda' x_i', \quad i = 1, \cdots, n$, como os valores da função linear discriminante de Fisher para as observações multivariadas x_i , conforme Johnson & Wichern(1982).

Vemos, assim, que os menores valores de $|t_j'\alpha - t_i'\hat{\alpha}|$ ocorrem quando a observação j for aquela que nos forneça um valor u_j o mais próximo possível de $\tilde{u}$:

$$\begin{aligned}\bar{u} &= \frac{1}{n}\left(\hat{\Lambda}'x_1 + \hat{\Lambda}'x_2 + \cdots + \hat{\Lambda}'x_n\right) \\ &= \frac{1}{n}\hat{\Lambda}'(x_1 + x_2 + \cdots + x_n) \\ &= \hat{\Lambda}' \quad \bar{x}\end{aligned}$$

É interessante notar que, mesmo quando x não é normal, mas pertence à família exponencial, $\log[p_1(x)/p_2(x)]$ ainda é uma função linear discriminante.

Concluímos, então, que essa observação j que "produziu" o máximo de $\widehat{\omega(\pi)}$ situa-se no centro de gravidade das duas populações univariadas u_1 e u_2 conforme Gráfico 2.

Fazendo uma comparação com o método de Fisher que usa como ponto de corte:

$$\hat{m} = \hat{\Lambda}'\frac{1}{2}(\bar{x}_1 + \bar{x}_2)$$

podemos dizer que temos procedimentos diferentes, pois existem as quantidades $\hat{d}_i$ e $\hat{\chi}_i$ e a função linear discriminante definida por $\log\{p_1(x)/p_2(x)\}$ depende da distribuição de x. Mas, para n_1 não muito diferente de n_2 e $x \sim$ Normal, os pontos de corte de Fisher e aquele definido pelo máximo de $\widehat{\omega(\pi)}$ serão muito próximos.

Observemos agora que o centro de gravidade das univariadas u_i depende da priori (p_1, p_2) e, portanto, a observação j que produziu o máximo de $\widehat{\omega(\pi)}$, sendo provavelmente definida por $\bar{u}$ é fortemente influenciada por elas. Notamos então que sendo $\hat{\pi}_j$ função crescente de $\hat{u}_j$ (3.10), teremos um ponto de corte mais "realista", isto é, próximo de 0.5 se $p_1 = 0.5$ mas, por exemplo, próximo de 0.3 se $p_1 = 0.3$. Para verificar estas afirmações realizamos mais dois experimentos amostrais tomando como valor para *a priori*: $p_1 = 0.3$ e depois $p_1 = 0.8$.

3.2 – Outras Simulações

O segundo experimento amostral, assim como o primeiro, consistiu em 100 ensaios onde, em cada ciclo, eram gerados 20 vetores (y_i, x_i) independentes, de acordo com (3.6) e com $p_1 = 0.3$. Agora o verdadeiro valor de α de acordo com (3.5) é (- 0.8473, 1, 0)' .

Novamente obtivemos os estimadores de máxima verossimilhança através da regressão logística e calculamos os 20 valores de $\widehat{\omega(\pi)}$ (a cada ensaio) usando como C_0, respectivamente, os 20 valores das probabilidades estimadas $\hat{\pi}_i$. Os resultados confirmaram que o ponto de corte médio definido pelo máximo valor de $\widehat{\omega(\pi)}$:

$$C_0 = 0.3564,$$

com desvio–padrão 0.09631 é bem mais interessante, de acordo com as considerações teóricas, do que o definido pela taxa aparente de erro mínima:

$$C_0 = 0.4338,$$

com desvio–padrão 0.0951.

Montamos então um terceiro experimento amostral usando *a priori* $p_1 = 0.8$.

Agora o verdadeiro valor de α é (1.3863,1,0)' e novamente o ponto de corte médio definido pelo máximo de $\widehat{\omega(\pi)}$:

$$C_0 = 0.7132$$

com desvio–padrão 0.10095, é bem melhor que 0.5, assim como é melhor do que o obtido com a $\overline{err}$ mínima:

$$C_0 = 0.5149,$$

que apresentou grande desvio–padrão 0.15397.

4 – UMA APLICAÇÃO EM MEDICINA

4.1 – O Problema

Existem vários motivos que levariam uma pessoa a um laboratório para fazer o exame "Espirografia" que consiste em uma série de testes da função respiratória. Ela pode ter sido encaminhada até ali por um pneumologista que gostaria de ver os resultados dos exames para

dar o seu diagnóstico, ou pode estar sofrendo uma crise asmática, ou por simples rotina pré-operatória, ou até mesmo por simples rotina para admissão em algum novo cargo ou concurso etc.

Essa população que procura o laboratório pode ser dividida em dois grupos: 1) doentes, 2) sadios. Neste caso, a variável aleatória y representa o laudo médico e será codificada como 1, sucesso, se o indivíduo for doente, e 0 (zero), caso contrário.

Feitos os exames, muitas vezes o médico não consegue decidir a qual dos dois grupos pertence o indivíduo. E isto pode acontecer porque ele está se iniciando no exercício da profissão e não tem a experiência necessária e/ou porque não existem fórmulas ou tabelas que o auxiliem nesta tarefa. Segundo a opinião do médico pneumologista que nos forneceu os dados, Dr. Cleonardo Augusto da Silva, isto é muito freqüente. Nosso propósito é então construir uma boa regra de classificação que o ajude a decidir.

Os dados são retirados de uma curva de expiração forçada, quando a pessoa enche o pulmão de ar e, em seguida, o solta com o máximo de força que puder. Esse volume total expirado é chamado Capacidade Vital Forçada e geralmente denotado por CVF. O tempo gasto para expirar 10% a 20% da CVF depende da força, o restante depende da permeabilidade das vias aéreas.

Na década de 40, Robert Tifeneau criou um índice, chamado índice de Tifeneau, que serve para medir a permeabilidade das vias aéreas de maior calibre. Para obtê-lo basta dividir o volume de ar expirado durante o segundo inicial do teste pela Capacidade Vital Forçada. Usaremos para essa variável a notação IT.

Na década de 60 começou-se a trabalhar com fluxos de expiração, que são dados de volume expirado pelo tempo.

Dentre esses fluxos, um dos mais usados é o FEF 25-75, definido como a razão entre a metade do volume total expirado e o tempo gasto para expirar os dois quartos centrais do volume total. Este número, normalizado pela CVF, dá origem à nossa segunda variável: FEF/CVF que mede o estado das vias aéreas da periferia, de pequeno calibre.

Além destas duas existe uma correção por idade: ID.

Nossos dados são portanto: 1)IT , 2)FEF/CVF, 3)ID e, de acordo com a notação usada:

$$t_i = (1, IT, FEF/CVF, ID)$$

Os dados foram colhidos no PULMOLAB, Laboratório de Fisiopatologia Pulmonar,RJ, e no Laboratório de prova de função do Hospital Pedro Ernesto,R.J. Uma cópia dos mesmos encontra-se em Mendes Luz(1988).

4.2 – O Melhor Ponto de Corte

Nossos objetivos são: em primeiro lugar, calcular numa primeira amostra, chamada amostra de treinamento, os valores dos pontos de corte definidos pela taxa aparente de erro mínima e pelo máximo de $\widehat{\omega(\pi)}$. Em segundo lugar testá-los em uma nova amostra da mesma população, que chamaremos de amostra de verificação. Iremos, também, testar o ponto de corte 0.5.

Assim, usando inicialmente uma amostra de treinamento com 254 observações, calculamos o valor do estimador de máxima verossimilhança do parâmetro α através do programa Logistic Regression do BMDP. A partir de $\hat{\alpha}$, e por substituição em (1.1), calculamos a estimativa da probabilidade de sucesso para cada paciente dessa amostra: $\hat{\pi}_i, i = 1, 2, \cdots, 254$.

Através do PROGRAMA 2 , dados em Mendes Luz(1988), foram feitos os cálculos necessários e a seguir temos um resumo dos resultados obtidos nesta primeira etapa.

Chamaremos de C1 o ponto de corte 0.5, C2 o ponto de corte definido pela taxa aparente de erro mínima, e C3 aquele definido pelo máximo de $\widehat{\omega(\pi)}$.

Resumo dos Resultados Obtidos na Amostra de Treinamento

C_0	$\overline{err}$	$\widehat{\omega(\pi)}$
$C1 = 0.5$	0.0945	0.0046
$C2 = 0.4442$	0.0866	0.0041
$C3 = 0.6582$	0.1063	0.0059

Observemos que o otimismo esperado da taxa aparente de erro é pequeno qualquer que seja o ponto de corte escolhido. Isto deve-se provavelmente ao fato de que temos uma amostra grande, n = 254, e apenas 4 parâmetros a estimar.

Foram então colhidos novos dados (69 observações) que usamos para verificar a adequacidade das regras de predição.

Para tanto queríamos determinar qual dentre os pontos de corte acima nos daria uma menor taxa verdadeira de erro, que definimos como a proporção de casos erroneamente classificados na amostra de verificação. Usamos o PROGRAMA 3 para fazer os cálculos necessários e a seguir estão os resultados:

C1 = 0.5000 $\Longrightarrow$ TAXA VERDADEIRA DE ERRO = 0.1159

C2 = 0.4442 $\Longrightarrow$ TAXA VERDADEIRA DE ERRO = 0.1304

C3 = 0.6582 $\Longrightarrow$ TAXA VERDADEIRA DE ERRO = 0.0870

Os resultados confirmam nossas expectativas e podemos então concluir que, para este problema, o melhor ponto de corte a ser usado na regra de classificação (2.1) é C3 = 0.6582, obtido a partir do máximo de $\widehat{\omega(\pi)}$.

Conforme dissemos na Introdução, os resultados obtidos aqui são apenas indicadores de um caminho a ser seguido, ficando para uma próxima etapa um estudo mais detalhado da variável aleatória C_0 , ponto de corte definido pelo máximo de $\widehat{\omega(\pi)}$.

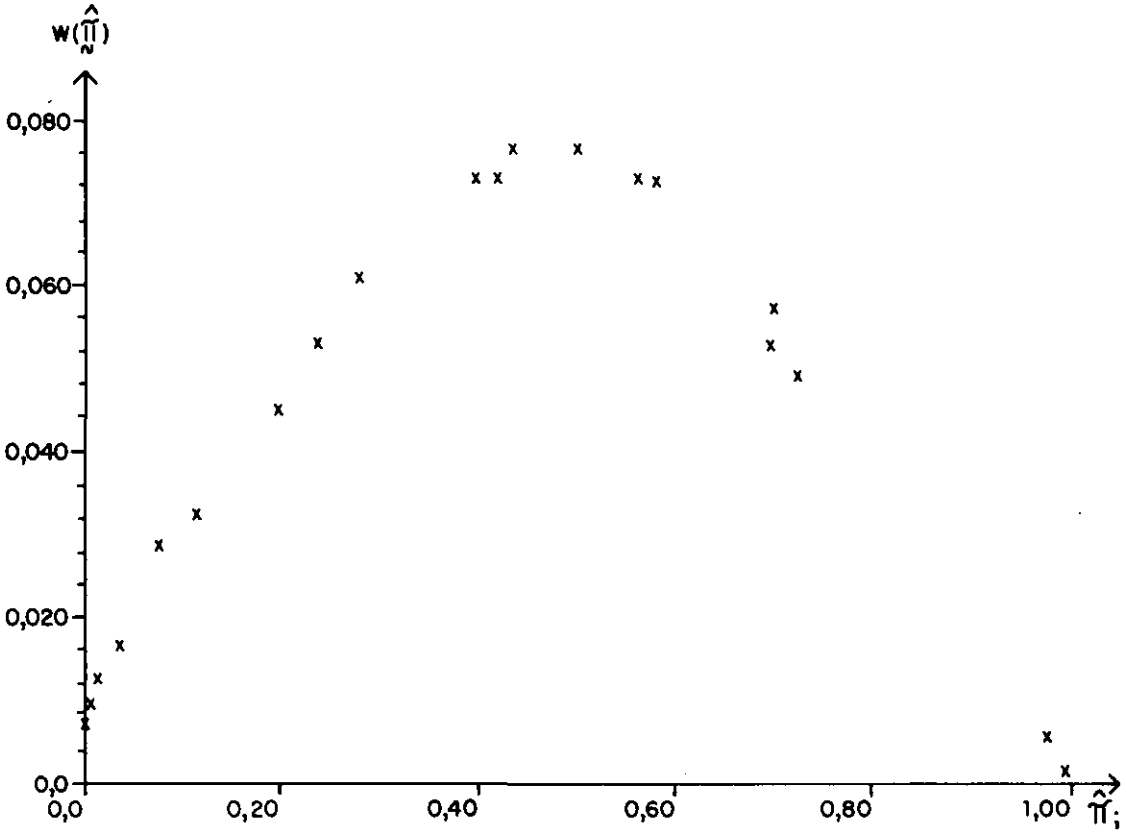


Gráfico 1

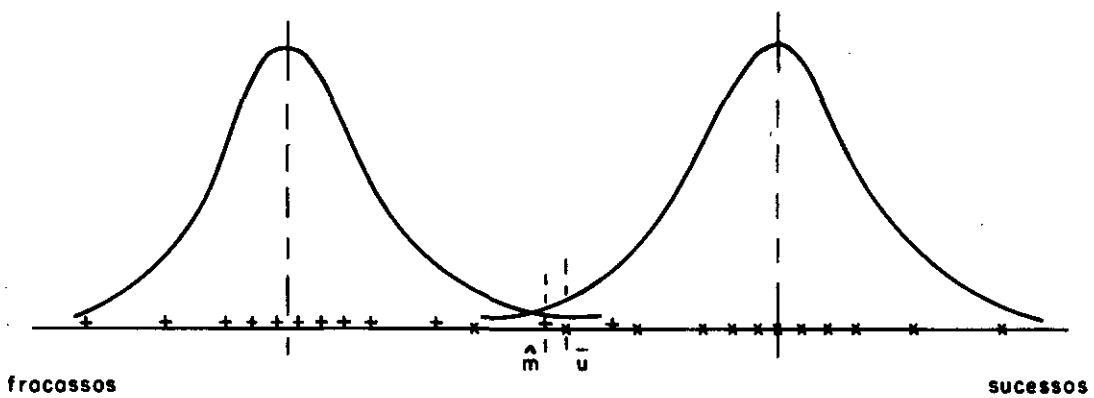


Gráfico 2

BIBLIOGRAFIA

COX, D. R. *The analysis de binary data.* Londres, Methuen, 1970.

EFRON, B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal Of the American Statistical Association*, (70): 892-8. 1975.

———— How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association*, (81): 460-71. 1986.

JOHNSON, R. A. & WICHERN, D. W. *Appplied multivariate statistical analysis.* New Jersey, Prentice-Hall, 1982.

LACHENBRUCH, P. A. *Discriminant analysis.* New York, Hafner Press, 1975

MENDES LUZ, B. *Um estimador para o vício da taxa aparente de erro na regressão logistica.* Tese (Mestrado)- Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1988.

SEBER, G. A. F. *Multivariate obervations.* New York, John Wiley, 1984.

RESUMO

O artigo trata do problema da classificação de dados binários. Em primeiro lugar, vemos qual o modelo adequado a este tipo de dados e que os parâmetros do modelo podem ser estimados através da Regressão Logística. Em seguida constrói-se uma regra de classificação e define-se uma medida de erro de predição: a taxa aparente de erro. Ficará claro que uma regra de classificação ótima dependerá da escolha de um ponto de corte ótimo. Apresentamos alternativas conhecidas para esse ponto de corte e propomos uma terceira opção obtida a partir do estudo de um estimador para o otimismo da taxa aparente de erro. Verifica-se então, através de simulações, a performance desse estimador ao definir um novo ponto de corte e, finalmente, aplicamos os resultados obtidos em um problema real com dados obtidos em um hospital do Rio de Janeiro.

ABSTRACT

This article is concerned with classification of binary data. First, an appropriate model to this kind of data is presented and we note that the estimated parameters may be obtained through the Logistic Regression. Second, a classification rule is formulated and a measure of prediction error is defined: the Apparent Error Rate. A good classification rule depends on the choice of the cutoff point. We present two well known possibilities for its cutoff point and propose a new one obtained from a study involving an estimator for the bias of the apparent error rate. We verify, through simulations, the performance of this new procedure and, finally, we apply the results to a real problem with data from a hospital in Rio de Janeiro.

DIFERENCIAIS DE RENDIMENTOS ENTRE CENSOS DEMOGRÁFICOS E PNADs – ASPECTOS CONCEITUAIS E OPERACIONAIS[1]

André Cezar Medici*

1 – INTRODUÇÃO

Censos Demográficos e Pesquisas por Amostras Domiciliares costumam ter diferentes objetivos. Os primeiros buscam obter informações capazes de delinear um quadro sócio-econômico e demográfico geral do país. Permitem que tais informações sejam fornecidas com o máximo de desagregação possível, seja em função de manter a representatividade de unidades espaciais menores ou de menor densidade demográfica, seja em decorrência de cruzamentos de variáveis que estabeleçam células ou agregados de baixa freqüência populacional.

Já as Pesquisas por Amostras Domiciliares contínuas podem ter objetivos múltiplos. No caso das PNADs podem ser identificados pelo menos três objetivos, a saber:

a) Acompanhar o comportamento conjuntural de algumas variáveis sócio-econômicas estruturais para o conjunto da população nacional ou para unidades espaciais de maior agregação;

b) Testar novos conceitos ou novas formas de captação de variáveis; e

c) Investigar temas especiais ou fenômenos sociais que tradicionalmente não são pesquisados, como sucedeu nos suplementos especiais das PNADs na década de 80.

*Economista, Professor e Pesquisador do IEI/UFRJ e da ENSP/FIOCRUZ.

A isto tudo somam-se ainda as diferenças trazidas pelos critérios de amostragem utilizados nas PNADs e no questionário da não-amostra do Censo Demográfico, que envolvem processos técnico-administrativos e operacionais distintos.

Dadas estas questões, é natural que pesquisas como as PNADs e os Censos Demográficos cheguem a resultados diferentes, mesmo quando observadas as mesmas variáveis segundo suas características.

O objetivo deste trabalho é estudar, de forma tópica, as diferenças existentes quanto às formas de captação dos rendimentos entre as PNADs e o Censo Demográfico de 1980[2]. Para tal serão consideradas apenas as duas PNADs relativas ao "em torno" do Censo Demográfico de 1980, ou seja, as realizadas em 1979 e 1981

2 - DIFERENÇAS DE ORDEM OPERACIONAL ENTRE O CENSO DEMOGRÁFICO DE 1980 E AS PNADs 1979 E 1981

Diferenças de critérios operacionais ou administrativos podem influenciar ou introduzir viés na comparabilidade entre distintas pesquisas voltadas a um mesmo objeto de conhecimento. No caso do Censo Demográfico de 1980 e das PNADs 1979 e 1981, podem ser enumeradas algumas diferenças desta natureza:

a) A comparação dos dados relativos ao Brasil, nestas duas pesquisas, fica prejudicada pelo fato de as PNADs não coletarem informações relacionadas à totalidade do espaço nacional, o que não ocorre com o Censo Demográfico de 1980. Assim, as informações contidas na PNAD 1979 excluem a área rural das Regiões Norte e Centro-Oeste. Em 1981, a PNAD passou a pesquisar a área rural da Região Centro-Oeste, embora tenha continuado a excluir a área rural da Região Norte. A recomposição da comparabilidade da totalidade do espaço nacional fica assim prejudicada quando se comparam o Censo Demográfico de 1980 e as PNADs 1979 e 1981;

b) A base cartográfica utilizada na PNAD 1979 foi a relativa ao Censo Demográfico de 1970, enquanto que a PNAD 1981 já se beneficiou da base cartográfica do Censo Demográfico de 1980. Assim, os setores censitários presentes na amostra da PNAD 1981 já incorporam a atualização da base operacional-geográfica permitida com as novas informações trazidas pelo Censo Demográfico de 1980;

c) A comparação dos valores absolutos destas pesquisas apresenta outras dificuldades, pois a expansão dos dados da amostra de 25% do Censo Demográfico de 1980 toma em consideração os próprios resultados obtidos no Censo, enquanto que a expansão dos dados da PNAD tem como base a utilização de uma projeção independente de população, construída a partir das

taxas de crescimento verificadas no espaço intercensitário das décadas anteriores[3]. Segundo informam as publicações das próprias PNADs, tal projeção foi "distribuída por sexo e 11 grupos de idade, segundo a composição etária obtida dos resultados da amostra. Os 22 fatores de expansão resultaram na divisão de cada grupo etário, assim calculado, pelo total de pessoas da amostra, nesses mesmos grupos"[4].

Os valores absolutos da PNAD 1979 foram expandidos com base na projeção de população calculada pela tendência de crescimento acumulada até 1970, enquanto que a PNAD 1981 incorporou, no cálculo da expansão dos valores absolutos, a tendência de crescimento populacional 1970/80. No entanto, uma publicação posterior[5] permitiu a correção de uma série de tabulações da PNAD 1979 pela tendência de crescimento populacional observada entre 1970 e 1980. Tal procedimento, no entanto, não se estendeu às fitas magnéticas da PNAD 1979, as quais continuam expandidas pela tendência da década anterior. Tais diferenças quanto aos valores absolutos não prejudicam, no entanto, a comparação dos valores "relativos". Assim, a "distribuição de rendimentos" não fica necessariamente alterada por este procedimento. Inversamente, o acompanhamento do crescimento e a comparação da magnitude da "massa salarial" tornam-se praticamente impossíveis de serem feitos em função das diferenças de critérios de expansão dos valores absolutos encontrados nas PNADs e nos Censos, bem como entre as PNADs 1979 e 1981;

d) As diferenças quanto ao período de referência podem, também, influenciar os dados, especialmente os de rendimentos, a partir da introdução de três efeitos de sazonalidade:

- variações nos reajustes salariais e períodos de referência de dissídio das distintas categorias profissionais, especialmente em conjunturas inflacionárias;
- variações no valor recebido por autônomos e assalariados com rendimento variável, cuja sazonalidade mensal pode aumentar ou reduzir o valor real dos rendimentos; e
- variações sazonais nos níveis de emprego.

O período de referência do Censo Demográfico de 1980 foi a última semana do mês de agosto para efeito de cômputo da condição de atividade, embora os rendimentos fossem computados como rendimentos médios mensais auferidos na ocasião da aplicação do questionário. Já os períodos de referência das PNADs 1979 e 1981 foram 22 a 28 de outubro e 8 a 14 de novembro, respectivamente; e

e) *Last but not least*,a rede de coleta da PNAD é composta por pesquisadores permanentes, com maior treinamento e experiência de campo, possibilitando uma maior "sintonia fina" na aplicação do questionário e permitindo, obviamente, a obtenção de respostas com maior fidedignidade do que as registradas nos Censos.

3 – DIFERENCIAIS RELACIONADOS AOS CONCEITOS

As diferenças conceituais na captação dos rendimentos entre as PNADs consideradas e o Censo Demográfico de 1980 podem ser estudadas, inicialmente, a partir da divisão entre "Rendimentos do Trabalho" e "Rendimentos de Outras Fontes".

3.1 – Rendimentos do Trabalho

Nas duas pesquisas em foco, a investigação sobre os rendimentos abrangeu todas as pessoas de dez anos e mais. O Censo Demográfico de 1980 investigou, em caráter excepcional, os rendimentos da ocupação habitual dos menores de 5 a 9 anos de idade[6] mas, para efeitos da agregação dos "rendimentos do trabalho", não foram consideradas tais fontes.

No que diz respeito aos rendimentos do trabalho, vale analisar a distinção existente entre "Rendimento da Ocupação Principal" e "Rendimento Obtido pelo Exercício de Outras Ocupações".

No caso do Censo Demográfico de 1980, os quesitos 37 e 38 dão conta da primeira situação, enquanto o quesito 39 computa o segundo caso. Para tal, é considerada "segunda ocupação" como aquela que caracteriza um novo conjunto de habilidades, e não o segundo trabalho. Um médico que trabalha no INAMPS e num consultório particular declara a soma desses rendimentos no quesito relativo à ocupação principal. Mas se ele trabalha no INAMPS e seu segundo emprego não é como médico, mas sim como professor, ele deveria declarar em separado os rendimentos como professor (enquanto oriundos da segunda ocupação).

No caso das PNADs, há uma diferença conceitual de grande importância. A PNAD 1979 investiga o rendimento do trabalho principal e de outros trabalhos (bloco 4, quesitos 6 e 7) no seio de uma mesma ocupação, além de investigar o rendimento de outras ocupações (bloco 4, quesito 8). Já a PNAD 1981 não investiga o rendimento segundo o critério ocupação principal/outras ocupações, mas sim através de outro critério: trabalho principal/outros trabalhos, o que pode levantar a hipótese de que nem mesmo os rendimentos da ocupação secundária estariam sendo levantados. Neste caso, a soma dos rendimentos do trabalho principal e dos outros trabalhos seria igual ao rendimento declarado como sendo da ocupação principal no questionário do Censo Demográfico de 1980.

Portanto, o critério de agregação de renda é distinto entre as duas pesquisas, dado que na ocupação principal podem ser computados (segundo o Censo Demográfico de 1980) os

rendimentos de mais de um trabalho na mesma ocupação, o que não acontece no cômputo do rendimento do trabalho principal na PNAD 1981. Presume-se, dessa forma, que, ao analisar o total dos rendimentos do trabalho declarado, há uma tendência ao dado do Censo Demográfico de 1980 estar superestimado com relação ao registrado na PNAD 1981. Nesse sentido, o rendimento agregado do trabalho do Censo Demográfico de 1980 estaria mais próximo do valor registrado, também de forma agregada na PNAD 1979.

O Censo Demográfico de 1980 pesquisou, em ambas as formas de rendimento, a existência de partes fixas e/ou variáveis. Para os rendimentos fixos, considerou-se, como período de referência, o mês de agosto de 1980 (mês de referência do Censo Demográfico de 1980), enquanto que, para os rendimentos variáveis, a média mensal dos rendimentos auferidos nos últimos 12 meses, até a data do Censo.

Nas PNADs , verifica-se que o critério de captação utilizado para a parte fixa dos rendimentos se mantém, o mesmo não ocorrendo para a parte variável. Neste caso, indaga-se apenas a média mensal (presumidamente) recebida e não a média mensal dos últimos 12 meses. Isto ocorre porque os rendimentos declarados na PNAD não têm como base o ano, mas sim o mês de referência da pesquisa.

Dois problemas se sobrepõem quanto a este ponto:

a) Primeiramente, a idéia de "média mensal" para os rendimentos variáveis deveria servir apenas como critério orientador para que o entrevistado fizesse uma estimativa. Mas não havia garantia, não só que o critério fosse aplicado, como também que a memória do entrevistado permitisse fornecer uma estimativa adequada; e

b) A existência de uma inflação que já atingia 120,02% entre agosto de 1979 e agosto de 1980 inviabilizava qualquer estimativa de média expressa em valores reais, para os rendimentos variáveis, a partir da informação prestada pela maioria dos entrevistados no Censo Demográfico de 1980. Os dados declarados representavam, provavelmente, um valor subestimado que se situava entre o nominalmente recebido pelo entrevistado e sua correção inflacionária. Isto ocorre inclusive nas PNADs, onde o recrudescimento inflacionário dificulta o cálculo econômico da correção de rendas nominais passadas em rendas reais presentes, por parte dos informantes.

Vale registrar, no entanto, que a discriminação do rendimento em parte fixa e parte variável é feita, no caso da PNAD 1979, tanto para ocupação principal como para as outras ocupações, o que não ocorre no Censo Demográfico de 1980 e na PNAD 1981 onde, no caso desta última, o rendimento das outras ocupações, a rigor, não é pesquisado.

Logicamente que existiam variações em torno desta questão. A primeira, no caso do Censo Demográfico de 1980, diz respeito às diferenças na captação dos rendimentos por posição na ocupação. Assim, no caso dos empregados (inclusive os trabalhadores agrícolas volantes), registrou-se a remuneração bruta em dinheiro recebida no mês de agosto de 1980. Se estes não

tivessem trabalhado naquele mês, seria registrado o rendimento bruto do último mês trabalhado. Este critério assemelha-se ao utilizado nas PNADs, excluindo-se a parte relacionada aos que não trabalharam no mês de referência, onde não são captadas informações sobre rendimentos.

Não seriam computados o 13º salário, a participação nos lucros ou qualquer outra forma de receita distribuída pela empresa ao assalariado. Este critério foi também utilizado nas PNADs em tela.

No caso dos trabalhadores autônomos e dos empregadores seria sempre considerada a média mensal do rendimento recebido nos últimos 12 meses. Portanto, o critério de mensuração dos rendimentos que os subdividia em "fixos" e "variáveis" era, na prática, válido somente para os empregados[7] tanto no Censo Demográfico de 1980 quanto nas PNADs, ressaltando-se que no caso destas últimas não se consideravam os 12 meses mas somente uma (vaga) estimativa média mensal. Isto ocorria dado que o rendimento de autônomos e empregadores era, por definição, considerado como variável.

Além do rendimento considerado "em dinheiro", o questionário do Censo Demográfico de 1980, bem como o das PNADs, buscou quantificar o rendimento médio mensal bruto auferido pelo recebimento de produtos ou mercadorias na ocupação declarada (quesito 38), desde que tais receitas fossem oriundas da comercialização desses produtos. Assim, não foi considerada a produção ou as mercadorias recebidas quando estavam voltadas para consumo do próprio informante. No caso das PNADs isso também ocorreu, sendo registrado não só para o trabalho principal, como também para os outros trabalhos (PNAD 1981) e para as outras ocupações, no caso da PNAD 1979. Isto difere do procedimento utilizado pelo Censo onde a pesquisa dessa forma de rendimento se restringe apenas à ocupação principal.

O Quadro I, em anexo, busca analisar as principais diferenças conceituais existentes na captação dos rendimentos do trabalho entre o Censo Demográfico de 1980 e as PNADs.

3.2 – Outros Rendimentos

O Censo Demográfico de 1980 pesquisa, ainda, outros rendimentos não provenientes do trabalho, para a totalidade das pessoas de 10 anos e mais, sejam ativas ou inativas. Tal investigação é feita através dos quesitos 46, 47, 48 e 49 do questionário do referido Censo Demográfico.

O quesito 46 investiga os rendimentos provenientes de aposentadoria, pensão, abono de permanência e ainda o equivalente a um duodécimo dos rendimentos retirados a título do PIS/PASEP. No caso do informante ter mais do que um desses rendimentos, deve ser

declarada, sempre, a soma dos rendimentos auferidos ao nível de todos esses itens. Nas PNADs, são pesquisados em separado (bloco 4, quesito 9 da PNAD 1979 e quesito 28 da PNAD 1981) os rendimentos relativos a aposentadorias e pensões (PNAD 1979) e aposentadorias, pensões e abonos de permanência (PNAD 1981). Isto não é feito, no entanto, para abonos de permanência (no caso da PNAD 1979) ou rendimentos do tipo PIS/PASEP. Portanto, a informação do Censo Demográfico de 1980 é mais completa do que a da PNAD 1979, embora a destas pesquisas permita visualizar, separadamente, o rendimento em aposentadorias e em pensões, o que não é possibilitado pelo questionário do Censo Demográfico de 1980.

O quesito 47 do questionário do Censo Demográfico de 1980 permite captar, também, o rendimento médio mensal (média dos últimos 12 meses) proveniente de aluguéis, arrendamentos de imóveis, móveis, veículos, máquinas e inclusive sublocações. O mesmo critério é utilizado no bloco 5, quesito 9 (1979) e no quesito 28 (1981) nas PNADs.

O quesito 48 do questionário do Censo Demográfico de 1980 investigou o rendimento médio mensal regularmente recebido, proveniente de doação em dinheiro, mesada de pessoa não moradora do domicílio ou pensão alimentícia. Considerou-se, adicionalmente, como doação o aluguel ou prestação habitacional mensal recebido de pessoa não moradora do domicílio. Portanto, este quesito permite investigar a renda de inúmeras pessoas como estudantes, filhos ou mulher desquitada ou divorciada etc. A PNAD 1979 pesquisa discriminadamente no quesito 9 do bloco 5 os rendimentos dessa natureza sob o título "doação ou mesada". Já a PNAD 1981 e as posteriores deixaram de investigar este item em separado, embora ele apareça na conceituação do subitem "outros rendimentos" (quesito 28).

Por fim, o Censo Demográfico de 1980 pesquisa a média mensal de rendimentos do emprego de capital recebidos pelo entrevistado nos últimos 12 meses, o que incluía lucros auferidos pelos proprietários (ou sócios) de negócios individuais e sociedades limitadas ou anônimas, mesmo que fossem pagos sob a forma de dividendos ou bonificações em ações. Incluía, também, os rendimentos derivados de aplicações financeiras dos recursos (cadernetas de poupança, letras de câmbio, letras imobiliárias, títulos da dívida pública, depósitos a prazo fixo etc.), computando, inclusive, como rendimentos, as correções monetárias, descontos, ágios etc. Por fim, incluía pensões derivadas de fundos de pensão, complementações salariais de entidades seguradoras etc.

Quanto ao registro desses dados, devem ser feitas algumas considerações:

a) Primeiramente, que os rendimentos de capital costumam ser subdeclarados em pesquisas domiciliares, não só no Brasil mas também em outros países, como demonstram as experiências internacionais. Ao que parece, as pesquisas domiciliares parecem adequadas para captar rendimentos do trabalho, embora o mesmo não possa ser dito com relação aos rendimentos de capital; e

b) Em segundo lugar, em conjunturas inflacionárias como a brasileira, torna-se difícil estabelecer critérios que possam mapear a real valorização financeira dos ativos mobiliários, distinguindo o que se pode atribuir a "valorização" ou lucro *stricto sensu* e o que deve ser creditado a título de "atualização ou correção monetária" destes valores. Assim, não parece correto considerar a correção monetária como "rendimento oriundo do emprego de capital", como sugere o manual do entrevistador do Censo Demográfico de 1980[8].

As PNADs, tanto a de 1979 como a de 1981, não registram, em separado, os "rendimentos de capital", englobando-os no item "outros", conjuntamente com diversas formas de rendas adicionais.

Da comparação do registro dos rendimentos não oriundos do trabalho, entre as duas pesquisas (Censo e PNAD), observa-se que a discriminação/agregação de itens, segundo os conceitos e prioridades utilizadas, pode trazer diferenças de quanto a magnitude dos valores registrados, principalmente quando se trata de fontes de renda sazonais ou esporádicas, que não costumam ser guardadas com fidedignidade na memória do entrevistado.

Mesmo assim, a maior parte dos rendimentos familiares são oriundos do trabalho e, portanto, as demais fontes entrariam como resíduo.

3.3 – Os Dados de Rendimento, Segundo Classes Especiais de Análise

No Censo Demográfico de 1980, os procedimentos administrativos de campo levaram o IBGE a adotar critérios de pagamento diferenciados para as entrevistas dos questionários do "Boletim da Amostra" e "da não-amostra"[9]. Com isso, na apuração dos resultados do Censo Demográfico de 1980, encontrou-se um número de pessoas por domicílio nos questionários da não-amostra superior ao verificado nos questionários da amostra. "Tal fato ocorreu de forma generalizada, ao nível do Brasil. Supõe-se que os entrevistados tenham burlado o critério estabelecido, mormente nas regiões que permitem menor controle, de forma a aplicar o questionário completo nos domicílios com menor número de pessoas, como meio de aumentar sua produtividade individual. Para isso ocorrer, sem ser notado pela supervisão, era necessário manter, em média, o critério de um questionário completo para cada 3 simplificados, mesmo não sendo seguida a sistemática de amostragem pré–estabelecida"[10]. É obvio que existem técnicas estatísticas que permitem corrigir o número médio de pessoas por domicílio, por ocasião da expansão da amostra de 25%; e isto foi feito no caso do Censo Demográfico de 1980. Mas tal correção não permitiu que a distorção na superestimativa dos indicadores sociais, particularmente os de renda, advinda dessa escolha "trágica" dos domicílios menores por parte dos entrevistadores, tenha sido corrigida.

O procedimento de pagamentos diferenciados por questionário preenchido foi, também, utilizado no Censo Demográfico de 1970, sem apresentar problemas como a diferença do número de pessoas por domicílios nos questionários da amostra e da não-amostra.

Assim, assume-se que os dados de renda domiciliar ou renda familiar, total e *per capita* em todos os casos do Censo Demográfico de 1980 podem estar divergentes dos encontrados nos Censos Demográficos anteriores e das PNADs em função destas diferenças.

4 – A COMPARAÇÃO DOS DADOS DE RENDA DO CENSO DEMOGRÁFICO DE 1980 E DAS PNADs 1979 E 1981

A demonstração empírica das diferenças conceituais e operacionais descritas anteriormente é muito difícil de ser feita. Necessitar-se-ia não só testar uma série de hipóteses relacionadas aos efeitos destas diferenças no comportamento dos dados, mas também construir um conjunto de tabulações especiais que possibilitassem a aceitação ou rejeição das mesmas. Além de tudo isso, a inexistência de PNADs realizadas em anos de Censo impossibilita a construção de comparações reais sobre os dados das duas pesquisas.

Quando se trata de comparar dados de anos distintos, tentando identificar diferenças de natureza conceitual ou metodológica, os riscos de existirem variáveis não controladas intervenientes no processo, ou ainda de se chegar a resultados aparentemente exitosos, mas baseados em correlações espúrias, são muito elevados. Por esses motivos, as comparações entre os dados de Censos e PNADs aqui apresentadas têm um objetivo meramente ilustrativo, ou seja, de apontar as diferenças sem ter a pretensão de explicá-las.

Por outro lado, estatísticas derivadas de distribuição de renda, tais como os índices e coeficientes de Gini, Theil, seus limites inferiores e superiores, bem como a distribuição da renda em decis e percentis de população, costumam apresentar valores distintos, mesmo quando se considera uma mesma base de dados relativa a um mesmo período de tempo. Tais diferenças podem estar relacionadas aos critérios de entrada de dados. Costuma-se dizer que quanto mais desagregadas as classes de renda utilizadas na entrada de dados, principalmente nos extremos da distribuição, maior a confiabilidade dos indicadores de concentração de renda.

Para evitar tais problemas, os dados de distribuição de renda aqui apresentados, tanto os do Censo Demográfico de 1980 quanto das PNADs 1979 e 1981, foram construídos a partir de oito classes de rendimento uniformes, tendo sido excluídas as pessoas com rendimento zero e as pessoas que não apresentaram declaração de rendimentos nessas pesquisas[11]. Mas como tais classes são dadas em salários mínimos, variações no valor real do salário mínimo de um ano para outro podem, também, introduzir pequenas distorções nas medidas de distribuição.

Excetuando-se a primeira tabela, os dados que se seguem indicam os rendimentos totais (do trabalho + outros rendimentos) da PEA com rendimentos, tanto no que diz respeito ao Censo Demográfico de 1980 quanto no que diz respeito às PNADs 1979 e 1981. Neste sentido, as próprias diferenças conceituais existentes entre estas pesquisas podem introduzir problemas de comparabilidade (mesmo que pequenos) entre os dados.

A Tabela 1 mostra que existem grandes discrepâncias entre os dados de rendimento dos ocupados do Censo Demográfico de 1980 e das PNADs, especialmente no que se refere às classes de mais baixo rendimento e as de mais alta renda (mais de 10 sm). À primeira vista, parece que a renda do Censo Demográfico de 1980 parece estar superestimada com relação à das PNADs. Apesar de os dados das PNADs excluírem do território as áreas rurais das Regiões Norte e Centro-Oeste (PNAD 1979) ou só da Região Norte (PNAD 1981), pode-se aventar a hipótese de que tal subestimação esteja associada aos procedimentos de campo do Censo Demográfico de 1980, já analisados, os quais podem ter introduzido um viés no questionário da amostra do referido Censo.

A Tabela 2 mostra a evolução do coeficiente de Gini[12] entre 1979 e 1981 e a comparabilidade deste indicador entre as PNADs relativas a estes anos e o Censo Demográfico de 1980. Nota-se que a concentração de renda, medida pelo Censo, parece ser menor do que a relativa às PNADs, para pelo menos dois segmentos da PEA: as mulheres e a PEA urbana. No primeiro caso, sabe-se que o Censo Demográfico de 1980, bem como os demais Censos Demográficos, apresenta uma tendência natural a subestimar o trabalho feminino, especialmente nas áreas rurais[13] onde o rendimento monetário oriundo do trabalho é seguramente inferior ao observado no meio urbano. Assim, pode-se atribuir a discrepância verificada nos níveis de concentração de renda feminina a subestimativa da PEA agrícola feminina.

No segundo caso, sabe-se, também, que as PNADs costumam ser mais precisas na captação do trabalho urbano do que os Censos. Assim, as discrepancias neste item devem-se a este fator. Quanto aos demais segmentos da PEA (homens e PEA rural), não se verifica, *a priori*, uma discrepância entre os coeficientes de Gini observados no Censo Demográfico de 1980 e nas PNADs 1979 e 1981. Mas deve-se ter cuidado ao analisar tal questão. Além de estas pesquisas não compreenderem a mesma cobertura das áreas rurais(Censos e PNADs), conforme já foi bastante ressaltado neste trabalho, o forte aumento da concentração verificado nas PNADs entre 1979 e 1981 pode ser o reflexo da introdução da área rural da Região Centro-Oeste na pesquisa, cujos rendimentos da PEA, em 1981, já eram bastante concentrados pela rápida penetração de métodos tecnificados de cultivo e da grande agricultura capitalista moderna na região. Para expurgar o efeito distorcivo que pode ser introduzido ao se comparar regiões que, mesmo residualmente, são diferentes, elaborou-se a Tabela 3 que compara os coeficientes de Gini do Censo Demográfico de 1980 e das PNADs no Estado de São Paulo e na Região Nordeste.

Verifica-se que, no caso de São Paulo, mesmo tratando-se de uma mesma região, as discrepâncias são bastante acentuadas quanto à magnitude destes indicadores, especialmente entre 1980 e 1981. Embora tal processo não se verifique nos dados relativos ao Nordeste, pode-se dizer que a comparabilidade dos dados de renda, entre o Censo Demográfico de 1980 e as PNADs, haja vista o exemplo de São Paulo, não é totalmente adequada.

Por fim, uma última comparação a ser feita diz respeito aos dados de distribuição de renda segundo decis e percentis. A Tabela 4 mostra que em cada corte populacional considerado existem diferenças na absorção da renda entre o Censo Demográfico de 1980 e as PNADs 1979 e 1981.

Outras diferenças merecem ser mencionadas. O percentual de trabalhadores sem remuneração, que chegava a 5,98% das pessoas ocupadas no Censo Demográfico de 1980, era muito inferior aos 11,64% encontrados para as pessoas ocupadas da PNAD 1979; percentual que se manteve semelhante na PNAD 1981. Esta gritante diferença pode estar ligada ao conceito de ocupação, que no Censo é investigado num período de 12 meses, enquanto que na PNAD está relacionado à semana de referência. O fato de a PNAD 1979 excluir as áreas rurais das Regiões Norte e Centro-Oeste atuaria no sentido inverso ao verificado, ou seja, tenderia a tornar o percentual de trabalhadores sem rendimento subestimado em relação ao Censo Demográfico de 1980.

De qualquer forma, as diferenças porventura existentes não invalidam de todo a comparabilidade entre os dados de rendimento do Censo Demográfico de 1980 e das PNADs, à medida que as estatísticas derivadas sobre o tema apresentam, no conjunto nacional, diferenças muito pequenas, em que pesem as discrepâncias existentes para algumas unidades espaciais em níveis menores de desagregação.

TABELA 1

DISTRIBUIÇÃO DAS PESSOAS OCUPADAS
POR CLASSES DE RENDIMENTO NO BRASIL: 1979-1981

CLASSES DE RENDIMENTO (em salários mínimos)	DISTRIBUIÇÃO DAS PESSOAS OCUPADAS (%)		
	1979	1980	1981
TOTAL	100,0	100,0	100,0
Até 1 Salário Mínimo	43,8	37,9	40,4
Mais de 1 a 2 Salários Mínimos	25,7	28,7	25,8
Mais de 2 a 5 Salários Mínimos	20,4	22,3	23,2
Mais de 5 a 10 Salários Mínimos	6,5	6,7	6,8
Mais de 10 Salários Mínimos	3,6	4,4	3,8

FONTE - IBGE, Censo Demográfico - 1980 e PNADs 1979 e 1981.

TABELA 2

COEFICIENTES DE GINI RELATIVOS AOS
RENDIMENTOS DA PEA COM RENDA MAIOR QUE ZERO
BRASIL: 1979-1981

CLASSES DA PEA	COEFICIENTES DE GINI		
	1979	1980	1981
TOTAL	0,5747	0,5700	0,5783
HOMENS	0,5674	0,5703	0,5716
MULHERES	0,5313	0,5194	0,5459
URBANO	0,5651	0,5571	0,5654
RURAL	0,4773	0,4906	0,5259

FONTE - IBGE, Censo Demográfico - 1980 e PNADs 1979 e 1981.

TABELA 3

COEFICIENTES DE GINI RELATIVOS AOS
RENDIMENTOS DA PEA COM RENDIMENTOS
BRASIL: 1979-1981

ANOS DE REFERÊNCIA	COEFICIENTES DE GINI POR REGIÕES	
	São Paulo	Nordeste
1979	0,5299	0,5672
1980	0,5179	0,5736
1981	0,5672	0,5707

FONTE - IBGE, Censo Demográfico - 1980 e PNADs 1979 e 1981.

TABELA 4

DISTRIBUIÇÃO DOS RENDIMENTOS DA PEA COM
RENDIMENTOS SEGUNDO DECIS E PERCENTIS
BRASIL: 1979-1981

DECIS E PERCENTIS	*DISTRIBUIÇÃO DOS RENDIMENTOS (%)*		
	1979	1980	1981
20^{-}	3,12	3,39	2,91
50^{-}	14,11	14,56	13,91
10^{+}	47,35	47,67	47,65
5^{+}	35,42	34,85	35,17
1^{+}	14,10	14,93	15,40

FONTE - IBGE, Censo Demográfico - 1980 e PNADs 1979 e 1981.

QUADRO I

COMPARAÇÃO ENTRE OS CONCEITOS E FORMAS DE CAPTAÇÃO DOS RENDIMENTOS DO TRABALHO ENTRE PNAD'S E O CD-1980

(continua)

CARACTERÍSTICAS	PESQUISAS	
	CENSO DEMOGRAFICO DE 1980	PNADS 1979 e 1981
1. Faixa Etária Abrangida	5 a 9 anos-rendimentos da ocupação habitual. 10 anos e mais-rendimentos da ocupação habitual e de outras ocupações.	10 anos e mais-rendimentos do trabalho principal e de outros trabalhos.
2. Período de Referência	Mês de referência (agosto de 1980) Obs.: para empregadores e autonomos foi considerada a média mensal dos rendimentos auferidos nos últimos 12 meses.	Mês de referência (outubro de 1979 e de 1980, respectivamente. Obs.: para empregadores e autonomos foi considerada uma estimativa do rendimento médio mensal auferido.
3. Forma de Rendimento Registrado	Fixo (para empregados) - considerado com o rendimento do mês de referência. Variavél - considerada com o rendimento médio mensal dos últimos 12 meses. Rendimento obtido pela venda de produtos ou mercadorias recibidos como pagamento no exercício da ocupação principal. Obs.: Os rendimentos fixos e variaveis são agregados num único quesito.	Fixo (para empregados) - considerado como o rendimento do mês de referência. Variavél - considerado como o rendimento estimado médio mensal. Rendimento obtido pela venda de produtos ou mercadorias recebidos como pagamento no exercício do trabalho principal e nos outros trabalhos. Obs.: Os rendimentos fixos e variaveis são agregados num único quesito. Existe um quesito que registra isoladamente o rendimento em produtos ou mercadorias dos outros trbalhos

QUADRO I (conclusão)

CARACTERÍSTICAS	PESQUISAS	
	CENSO DEMOGRÁFICO DE 1980	PNADS 1979 e 1981
4. Quesitos Envolvidos	37 - Rendimento da Ocupação Principal. 38 - Rendimento em Produtos ou Mercadorias da Ocupação Principal 39 - Rendimento Médio Mensal de Outras Ocupações.	PNAD 1979 Bloco 4 - Quesito 5: Rendimento do trabalho Principal. Bloco 4 - Quesito 6: Rendimento de Outros Trabalhos. Bloco 4 - Quesito 7: Rendimento de Outras Ocupações. Obs.: Em todos os tres quesitos são discriminados separadamente os rendimentos fixos, variaveis e em produtos ou mercadorias. PNAD 1981 Bloco 5 - Quesitos 7 e 9: Rendimento do trabalho principal e dos outros trabalhos, respectivamente
5. Conceitos Chave para Discriminação dos Rendimentos do Trabalho	Ocupação Principal - trabalho ou conjunto de trabalhos exercidos com com um mesmo ofício, habilidade ou profissão (que conferem maior rendimento ou ocupam a maior parte da jornada de trabalho do (entrevistado).	PNAD 1979 Trabalho Principal - aquele que propicia maior rendimento, ou em caso de empate, absorve a maior parte da jornada de trabalho. Ocupação Principal (ver definição ao lado) PNAD 1981 - IDENTICO AO CD-1980 Obs.: A PNAD 1979 permite recompor a forma de agregação dos dados de trabalho principal em ocupação principal, para efeitos de comparação com a PNAD 1981 e o CD-1980, mas a reciproca não e verdadeira.

ELABORAÇÃO: ANDRE CEZAR MEDICI

RESUMO

O propósito deste artigo é explicar as principais diferenças existentes entre os conceitos e a forma de captação dos dados de rendimento utilizados no Censo Demográfico de 1980 e nas PNADs situadas em seu em torno de (1979 e 1981). O artigo procura, também, comentar algumas diferenças entre os dados de distribuição de renda existentes entre estas pesquisas.

ABSTRACT

The proposal of this article is to explain the main differences between the concepts and the collect proceedings of income distribution data used in 1980 demographic census and in 1979 and 1981 household surveys in Brasil. The article also comments some differences about the data informations on income distribution produced by census and household surveys.

NOTAS

1 - Trabalho apresentado no VI Encontro Nacional de Estudos Populacionais, patrocinado pela ABEP, realizado em Olinda(PE), entre 16 e 20 de outubro de 1988.

2 - Grande parte das reflexões aqui presentes tomaram como ponto de partida os artigos "Notas Interpretativas sobre a variável Renda nos Censos Demográficos", publicado em "Censos, Consensos, Contra-Sensos" (III Seminário Metodológico dos Censos Demográficos, realizado em Ouro Preto em junho de 1984), ABEP, e "A Mensuração da Subjetividade: Notas sobre a Variável Renda nas PNADs", apresentado no Seminário de Avaliação das PNADs, patrocinado pela ABEP e realizado em Nova Friburgo (RJ) entre 13 e 15 de junho de 1988; ambos elaborados pelo autor do presente trabalho.

3 - Uma explicação aproximada das técnicas de projeção de população utilizadas pelo IBGE pode ser vista em LYRA MADEIRA, J. *et alii*, "A Dinâmica do Movimento Natural da População Brasileira", IBGE, Rio de Janeiro, 1979. 71p.

4 - Ver publicações do IBGE, relativas à PNAD 1981.

5 - "Metodologia das PNADs na Década de Setenta", IBGE, Rio de Janeiro, 1981.

6 - Tal informação, relacionada ao Bloco 3 do Boletim da Amostra do questionário do Censo Demográfico de 1980, deveria ser relativa apenas aos menores que nesta faixa (5 a 9 anos) detinham ocupação habitual, à semelhança do conceito definido no Manual do Recenseador para o quesito 30. De acordo com este conceito, ocupação habitual é aquela que é exercida durante a maior parte do período de referência.

7 - Ver "Manual do Recenseador CD 1.09", IBGE, 1980, p.61 e seguintes

8 - Sobre este ponto, ver comentários feitos por Waldomiro Pecht ao trabalho que apresentei no III Seminário Metodológico dos Censos Demográficos, patrocinado pela ABEP e realizado em Ouro Preto, em junho de 1984, citado na nota de rodapé no.1. Com relação aos rendimentos de capital, este autor afirma que, "no fundo, o critério do IBGE leva a confundir renda com riqueza, indicadores de fluxo com indicadores de estoque. O imposto de renda no Brasil, reconhecendo esse problema, considera a correção monetária das cadernetas de poupança e dos títulos em geral como não tributáveis, exatamente para não incorrer nos erros assinalados". *In* PECHT, W., "Algumas Observações Metodológicas sobre a Variável Renda tal como Aparece nos Censos Demográficos Brasileiros", *In* ABEP, "Censos, Consensos, Contra-Sensos", *op.cit.*, p. 138.

9 - Ver MEDICI, A.C., "Notas Interpretativas..." *op.cit.*, p. 91 e segs.

10- Idem, Idem, p.95-96.

11 - As classes de rendimento utilizadas para o cálculo das estatísticas derivadas e medidas de concentração de renda aqui apresentadas foram: mais de zero até 1/2 salário mínimo (sm); mais de 1/2 até 1 sm; mais de 1 até 2 sm; mais 2 até 3 sm; mais de 3 até 5 sm; mais de 5 até 10 sm; mais de 10 até 20 sm; mais de 20 sm.

12 - O Coeficiente de Gini é uma medida de desigualdade ou concentração que varia de zero a um, sendo zero a máxima igualdade e um a máxima desigualdade. Para maiores esclarecimentos ver, entre outros, o trabalho de COSTA, R. A., "Distribuição da Renda Pessoal no Brasil - 1970", IBGE, Rio de Janeiro, 1977.

13 - Ver PAIVA, P.T.A., "A Conceituação e a Enumeração da População Economicamente Ativa nos Censos Demográficos Brasileiros" In ABEP, "Censos, Consensos e Contra-Sensos", *op.cit.*, pp. 19-66.

REDUÇÃO DA AMOSTRA DA PESQUISA MENSAL DE EMPREGO; ESTRATÉGIA PARA REDUZIR O CUSTO DA PESQUISA[1]

Pedro Luis do Nascimento Silva*
Fernando Antonio da Silva Moura*

1 - INTRODUÇÃO

Durante todo o ano de 1987 foram apontados problemas com a coleta da Pesquisa Mensal de Emprego - PME em pelo menos três das seis regiões metropolitanas onde a pesquisa é realizada: Belo Horizonte, Recife e Porto Alegre. Esses problemas (aumento das taxas de não-entrevista, sobrecarga da equipe de campo, desvio de supervisores para atividade de coleta) vinham sendo apontados como provenientes do aumento exagerado da amostra coletada mensalmente ou por causa da redução da equipe de campo.

Motivada pela tentativa de solucionar esses problemas, e decartando a princípio a solução de contratação imediata de pessoal para a equipe de campo, decidiu-se estudar uma REDUÇÃO da amostra da PME, que desse conta de minimizar os problemas enfrentados durante a coleta.

Foi então iniciado um trabalho de avaliação da situação da carga de trabalho das equipes de campo da *PME*, descrito em *Mansoldo (1988)*, bem como de elaboração e avaliação de alternativas para a redução da amostra. Independentemente do resultado da avaliação da carga atual de trabalho das equipes de campo, a amostra atualmente pesquisada na PME,

*Analistas Especializados do Núcleo de Metodologia da Diretoria de Pesquisas, da Fundação Instituto Brasileiro de Geografia - IBGE.

com cerca de 45 000 domicílios a coletar cada mês, é considerada grande por comparação com a amostra da PNAD – Pesquisa Nacional por Amostra de Domicílios – que contempla cerca de 90 000 domicílios no país como um todo, e apenas cerca de 30 000 domicílios nas seis regiões metropolitanas abrangidas pela PME.

A comparação com a PNAD, embora inevitável, deve ser feita com cuidado, pois os objetivos dessa pesquisa quanto à produção de estimativas e quanto à precisão destas são distintos dos da PME, que procura estimar mensalmente taxas que dizem respeito a categorias rarefeitas, e cujas flutuações de um mês para o outro, mesmo pequenas, são alvo de interpretação nem sempre adequada por parte de seus usuários.

Além disso, a própria experiência de alguns técnicos com a recente redução por que passou a amostra da PNAD, levou a uma posição inicial de buscar alternativas para uma redução da amostra da PME antes mesmo de concluída a avaliação final da carga de trabalho do pessoal de campo. Havia uma impressão prévia de que o tamanho da amostra poderia ser reduzido sem impacto significativo na precisão das estimativas com possibilidades de ganho devido à redução de erros alheios à amostragem (redução das taxas de não–entrevista).

Os trabalhos foram então conduzidos em duas linhas paralelas: a avaliação da carga de trabalho por um lado e, por outro, a elaboração de alternativas para redução da amostra.

2 – AVALIAÇÃO DE CARGA DE TRABALHO DAS EQUIPES DE CAMPO

Os dados e informações utilizados nesta seção são provenientes de Mansoldo (1988), onde uma discussão mais detalhada do problema pode ser encontrada.

Inicialmente, na tentativa de identificar as causas dos problemas hoje verificados na coleta da PME, foi reunida uma série histórica que permite examinar o crescimento da amostra coletada desde o ano–base 1982, quando a pesquisa foi definitivamente implantada nas seis regiões metropolitanas atualmente investigadas. Os dados sobre a média dos tamanhos de amostra pesquisados a cada mês, que podem ser encontrados no Anexo 1, revelam que entre 1982 e 1987 a amostra teve um crescimento modesto, de apenas 9,3% no total, sendo que somente em Salvador (14,5%) e em Belo Horizonte (11,1%) as taxas de crescimento foram superiores a esse nível.

O crescimento da amostra da PME é explicado pela metodologia adotada, que compreende a realização periódica (pelo menos a cada dois anos) de operações de "listagem" para atualização dos cadastros de domicílios existentes nos setores selecionados para comporem a amostra, bem como pela adoção da amostra de "Novas Construções" utilizada para tentar incorporar na amostra parte das novas unidades domiciliares surgidas a cada ano. Para maiores detalhes

sobre a metodologia e esquema amostral da PME consultar IBGE (1983).

Em seguida, foi reunida outra série histórica, com o número de entrevistadores envolvidos na coleta da PME a cada mês, para cada uma das regiões metropolitanas abrangidas durante o período 1982 a 1987. Essa série encontra-se no Anexo 2, e sua análise revelou que somente em Belo Horizonte e em Porto Alegre houve redução expressiva no tamanho da equipe de entrevistadores durante o período considerado. Nas outras áreas, a equipe cresceu acompanhando o crescimento da amostra, à exceção de Recife onde permaneceu estável.

Isto levou à conclusão de que os problemas de coleta da PME atualmente enfrentados em Belo Horizonte e Porto Alegre podem ser explicados **principalmente** pela redução de suas equipes de coleta (23,6% em Belo Horizonte e 15,5% em Porto Alegre) que, combinada com o crescimento gradual da amostra, resultou em acréscimo substantivo da carga de trabalho dessas mesmas equipes.

Após essa análise da carga de trabalho das equipes de campo da PME nas diversas regiões metropolitanas, concluiu-se que a amostra precisaria ser reduzida apenas em Belo Horizonte (42%), Porto Alegre (33%) e Recife (12%) para eliminar os problemas que vêm sendo atribuídos à sobrecarga das equipes de campo. Esses percentuais de redução dariam conta de, mantidas as equipes de coleta atuais nessas regiões, reduzir a carga de trabalho de coleta para níveis considerados satisfatórios tanto pela rede de coleta quanto pelos técnicos responsáveis pela PME.

Embora não parecesse necessário reduzir a amostra da PME nas demais regiões, decidiu-se continuar os estudos em todas as regiões abrangidas pela pesquisa, já que era vista como necessária a implantação de novos procedimentos de campo. Os novos procedimentos demandarão mais tempo dos entrevistadores em atividades preparatórias e de registro/avaliação de ocorrências. Sendo assim, era considerada razoável a idéia de reduzir a amostra da PME também em Salvador, Rio de Janeiro e São Paulo, de modo a diminuir os custos de coleta e processamento, mas também de modo a viabilizar redução da carga de trabalho da equipe de campo nessas áreas, criando condições favoráveis à implementação dos novos procedimentos.

Os percentuais de redução da amostra sugeridos em função da carga de trabalho não foram assumidos como definitivos, mas serviram para orientar o encaminhamento dos estudos para escolha de uma proposta de redução.

3 - A AMOSTRA ATUAL DA PME

A amostra da PME cobre as Regiões Metropolitanas de Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo e Porto Alegre. A seleção da amostra em cada região foi feita

independentemente das demais, possibilitando a obtenção de resultados separadamente para cada uma das regiões.

Em cada uma das seis regiões pesquisadas, adotou-se um desenho com dois estágios de seleção. No primeiro estágio foram selecionados setores, previamente ordenados segundo os municípios, e dentro dos municípios segundo a situação (urbanos e depois rurais) com base no procedimento sistemático com probabilidades proporcionais ao número de domicílios do setor no Censo de 80. No segundo estágio, foi efetuada a seleção dos domicílios a serem pesquisados em cada um dos setores anteriormente escolhidos, usando sorteio sistemático simples.

A ordenação dos setores por município e situação e o sorteio de setores pelo método sistemático conferem à amostra uma pseudo-estratificação dos setores, a qual tentou assegurar que todos os municípios componentes de cada uma das regiões metropolitanas investigadas ficassem igualmente representados na amostra. Em alguns poucos casos específicos, isto não foi conseguido e há alguns casos de municípios sem nenhum setor selecionado para a amostra da PME. A intenção foi a mesma com a ordenação dos setores segundo a situação urbana x rural, dentro dos municípios.

Dentro de cada uma das regiões metropolitanas o desenho é autoponderado, isto é, todas as unidades domiciliares têm a mesma probabilidade de seleção (igual à fração de amostragem) e, conseqüentemente, o mesmo peso (inverso da fração de amostragem).

A tentativa de manter o desenho autoponderado ao longo do tempo é um dos fatores que determinam o crescimento da amostra, pois, cada vez que se procede nova operação de listagem dos setores componentes da amostra e se verifica que o número de unidades domiciliares num setor qualquer cresceu, o número de domicílios aí selecionados é aumentado na mesma proporção em que cresceu o número total de domicílios em relação ao período-base. Dentro de cada setor selecionado no primeiro estágio, os domicílios são selecionados por amostragem sistemática com eqüiprobabilidade. Em cada domicílio selecionado, todas as pessoas são arroladas, e são levantadas as informações requeridas para a pesquisa e que constam do questionário utilizado, como pode ser verificado em Metodologia da Pesquisa Mensal de Emprego 1980, (1983).

A PME possui também um mecanismo de atualização cadastral, denominado "Cadastro de Projetos de Novas Construções", que inclui as unidades domiciliares surgidas após o Censo de 1980 em unidades habitacionais com 30 ou mais domicílios. Esse universo é estratificado por municípios, sorteando-se em cada estrato uma amostra aleatória simples sem reposição dos domicílios, com a mesma fração amostral adotada para a amostra básica na região metropolitana correspondente.

Como é uma pesquisa mensal e repetitiva, a PME incorpora também um processo de rotação da amostra, destinado a minimizar os efeitos negativos que poderiam ser provocados pelo cansaço dos informantes, como a recusa em participar da pesquisa, por exemplo.

O procedimento de rotação consiste em substituir 25% das unidades domiciliares da amostra

a cada mês. Segundo o procedimento adotado, um certo domicílio será visitado durante oito meses ao todo, sendo que se ele entrou na amostra no mês t ele será visitado nos meses t, $t+1$, $t+2$ e $t+3$, depois ficará fora da amostra por oito meses, e retornará no mês $t+12$, sendo visitado também nos meses $t+13$, $t+14$ e $t+15$, quando então será definitivamente retirado da amostra.

Para assegurar melhor controle do trabalho de campo, a substituição dos domicílios da amostra se dá mediante a substituição dos **painéis** nos quais a amostra foi subdividida. Para maiores detalhes sobre o processo de rotação da amostra consultar Metodologia da Pesquisa Mensal de Emprego 1980, (1983).

Finalmente, devido ao "esgotamento" de alguns setores causado pelo sistema de rotação adotado, há um processo para substituição desses setores, que procura manter o espalhamento geográfico da amostra, bem como procura setores substitutos com tamanho similar ao dos substituídos.

4 – ALTERNATIVAS PARA A REDUÇÃO DA AMOSTRA

Durante as discussões iniciais em torno das alternativas possíveis para a redução da amostra da PME, foram explicitados alguns requisitos e limitações que a amostra reduzida deveria satisfazer.

A primeira limitação, e a mais séria de todas, foi a impossibilidade de se redesenhar a amostra. Esta hipótese foi descartada devido aos altos custos que implicaria (necessidade de realizar operação de listagem nos novos setores selecionados para a nova amostra), devido à exigüidade do tempo disponível para planejar e implementar em campo a amostra já reduzida e devido à crença de que o cadastro ou marco de referência fornecido pelo Censo Demográfico de 1980 não oferece uma base adequada para a redefinição do desenho, sendo conveniente para mudanças de grande porte aguardar a realização do Censo Demográfico de 1990. Além disto, havia uma grave restrição operacional impondo a manutenção do sistema computacional hoje usado para o cálculo das estimativas.

Em segundo lugar, embora estabelecida de maneira inexata, havia a intenção de manter um nível aceitável de precisão para as estimativas das taxas, totais ou médias divulgadas atualmente com base na PME. Isto significava dizer que os coeficientes de variação das estimativas não deveriam crescer muito em relação aos níveis atuais, embora essa afirmação nunca tenha sido traduzida em termos quantitativos pelos analistas, devido à natureza multivariada dos resultados e à complexidade do problema.

Em terceiro lugar, devido às restrições de natureza operacional, considerava-se desejável

a manutenção do esquema de autoponderação da amostra, com a finalidade de facilitar a estimação e o controle operacional.

As limitações impostas levaram ao exame de três alternativas para redução da amostra, todas baseadas na idéia de selecionar **subamostras** da amostra atual da PME.

A primeira alternativa consistia em selecionar uma subamostra de setores mantendo, nos setores da subamostra, a mesma amostra de domicílios. Esta alternativa, olhada sob a ótica da redução de custos, é considerada a de maior eficácia, devido a diminuir prioritariamente os custos de deslocamento entre os setores da amostra.

A segunda alternativa consistia em selecionar subamostras de domicílios, mantendo na amostra todos os setores atualmente pesquisados. Neste caso, embora haja redução nos custos devido à diminuição do número de domicílios a entrevistar, essa redução não afetará os custos de deslocamento, sendo considerada menos indicada que a primeira sob esse aspecto.

A terceira alternativa consistia numa combinação das alternativas anteriores, onde seriam selecionadas subamostras de setores e de domicílios. Essa alternativa nasceu devido a dois fatores: o primeiro foi a existência de municípios com reduzido número de setores na amostra, sendo contra-indicada a subamostragem de setores nesses municípios; o segundo foi a idéia de buscar um plano de redução ótimo, no sentido de minimização do impacto na precisão das estimativas. Mais tarde, a busca do ótimo se revelou impraticável devido à falta de informações sobre os custos de coleta da PME e sobre a precisão desejada das estimativas dos diversos indicadores e taxas divulgados ou calculados a partir da pesquisa.

Feita a definição das alternativas básicas que seriam testadas ou avaliadas, o trabalho prosseguiu com a tentativa de obtenção de instrumentos que permitissem avaliar essas alternativas, pelo menos quanto ao impacto a ser esperado na precisão das estimativas, já que do ponto de vista dos custos essa avaliação não seria possível devido à falta de informações para esse fim.

Decidiu-se, então, montar um arquivo com os dados individuais da PME referentes ao mês de outubro de 1987, excluindo os registros referentes a unidades domiciliares provenientes do "Cadastro de Projetos de Novas Construções". A escolha de um mês apenas para o trabalho de simulação e avaliação de alternativas deveu-se à exigüidade do tempo disponível para o trabalho, à complexidade dos cálculos envolvidos e à observação baseada na história da PME de que a precisão das estimativas não varia significativamente de um mês para o outro. O mês de outubro foi escolhido por ser considerado um mês não atípico em termos da série de resultados produzidos.

A exclusão das unidades domiciliares provenientes do "Cadastro de Projetos de Novas Construções" deveu-se à pequena importância relativa no total da amostra e ao aumento da complexidade dos cálculos que sua inclusão implicaria.

Finalmente, foi necessário definir um subconjunto das estimativas produzidas pela PME

que seria examinado – Almeida (1988) – já que se considerou impraticável trabalhar com o conjunto completo de indicadores, taxas e valores absolutos hoje elaborados a partir dessa pesquisa. Foram então escolhidas 35 "variáveis" (Anexo 3), isto é, taxas, indicadores ou valores absolutos cuja precisão seria analisada em detalhe para cada uma das alternativas de redução propostas. Essa escolha, intencional, levou em conta a relevância das variáveis, bem como tentou contemplar tanto as de maior como as de menor precisão, incluindo valores absolutos para evitar que somente estimativas baseadas em taxas fossem contempladas, e incluindo também dados de rendimento, para evitar que a análise se restringisse a estimativas baseadas em contagens.

5 – METODOLOGIA PARA AVALIAÇÃO ALTERNATIVAS DE REDUÇÃO

5.1 – Considerações Gerais

O primeiro passo na definição de uma metodologia para avaliação das diversas alternativas de redução consideradas para a amostra da PME foi a escolha de um método para avaliar a precisão da amostra atual da PME. Nessa pesquisa, cuja amostra de setores foi selecionada com base em procedimento sistemático, não há fórmulas exatas para o cálculo dos "erros de amostragem", isto é, da precisão das estimativas.

Optou-se por usar como aproximação a fórmula correspondente à de um desenho estratificado com dois estágios de seleção, onde no primeiro os setores são estratificados por município e sorteados com probabilidade proporcional ao número de domicílios com reposição e, no segundo, os domicílios são sorteados segundo amostragem aleatória simples sem reposição. Acredita-se que as fórmulas provenientes dessa aproximação fornecem uma cota superior para a variância das estimativas obtidas a partir da PME.

Além disso, essas fórmulas são as atualmente empregadas para o cálculo dos "erros de amostragem" da PME.

Assim, as fórmulas utilizadas para obtenção de estimativas de total, em cada uma das regiões metropolitanas, para uma variável y qualquer são dadas por:

$$\hat{Y}_r = P\frac{\hat{Y}}{\hat{P}} = P \cdot \hat{R} \qquad (1)$$

P população residente obtida por processo de projeção, independente da amostra [Frias

(1987)];

$\hat{Y}$ total da variável y estimada através da amostra;

$\hat{P}$ total de pessoas residentes estimado através da amostra;

$$\hat{Y} = \sum_{h=1}^{H} \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{1}{\Pi_{hi}} \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{n_{hi}} y_{hij} = \sum_{h=1}^{H} \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{\hat{Y}_{hi}}{\Pi_{hi}} \tag{2}$$

onde

y_{hij} é o total da característica y no j–ésimo domicílio na amostra do i–ésimo setor selecionado dentro do h–ésimo município;

$j \in \{1, 2, \cdots, n_{hi}\}$

$i \in \{1, 2, \cdots, m_h\}$

$h \in \{1, 2, \cdots, H\}$

n_{hi} é o número de domicílios na amostra do i–ésimo setor selecionado dentro do h–ésimo município;

m_h é o número de setores selecionados no h–ésimo município;

N_{hi} é o número total de domicílios do i–ésimo setor selecionado dentro do h–ésimo município de acordo com a última operação de listagem;

Π_{hi} é a probabilidade de seleção do i–ésimo setor selecionado dentro do h–ésimo município e corresponde à razão entre o número de domicílios existente nesse setor e o número total de domicílios do município a que pertencia à época do Censo Demográfico de 1980;

$\hat{Y}_{hi} = \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{n_{hi}} y_{hij}$

A variância relativa do estimador $\hat{Y}_r$ pode ser estimada, com base no processo denominado *Ultimate Cluster* – Hansen e outros (1953, p. 419) –, por:

$$\hat{V}_r(\hat{Y}_r) = \frac{1}{\hat{Y}^2} \sum_{h=1}^{H} \frac{1}{m_h} \left[s_{hy}^2 + \hat{R}^2 s_{hp}^2 - 2\hat{R} s_{hpy} \right] \tag{3}$$

onde

$$
\begin{aligned}
s_{hpy} &= \frac{1}{m_h - 1} \sum_{i=1}^{m_h} \left(\frac{\hat{P}_{hi}}{\Pi_{hi}} - \hat{P}_h \right) \left(\frac{\hat{Y}_{hi}}{\Pi_{hi}} - \hat{Y}_h \right) \\
\hat{Y}_h &= \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{\hat{Y}_{hi}}{\Pi_{hi}} \\
\hat{P}_h &= \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{\hat{P}_{hi}}{\Pi_{hi}} \\
s_{hy}^2 &= s_{hyy} \\
s_{hp}^2 &= s_{hpp}
\end{aligned}
\tag{4}
$$

sendo $\hat{P}_{hi}$ a estimativa da população residente no i-ésimo setor do h-ésimo município, com base na amostra.

Para estimativas de taxas utiliza-se a seguinte fórmula:

$$
\hat{T} = \frac{\hat{Y}_r}{\hat{X}_r} = \frac{P \cdot \frac{\hat{Y}}{\hat{P}}}{P \cdot \frac{\hat{X}}{\hat{P}}} = \frac{\hat{Y}}{\hat{X}} \tag{5}
$$

onde $\hat{Y}_r$ e $\hat{X}_r$ são, respectivamente, estimativas dos totais Y e X das variáveis y e x obtidas a partir da amostra.

A variância relativa do estimador $\hat{T}$ pode ser calculada fazendo-se as adaptações necessárias nas fórmulas (3) e (4), uma vez que $\hat{T}$ é também um estimador de razão.

Do que foi exposto na Seção 4, concluiu-se que a única alternativa viável para a redução seria o emprego de subamostragem. No entanto, a manutenção dos estimadores atuais é um dos requisitos desejáveis (embora não cruciais) que *a priori* não estavam garantidos com a alternativa de subamostragem.

Era preciso, então, avaliar o que implicaria a subamostragem no processo de estimação e, em particular, que método de seleção da subamostra deveria ser utilizado.

Decidiu-se empregar amostragem aleatória simples sem reposição para fazer subamostragem de setores, optando-se por fazer a seleção independentemente em cada município, a fim de manter o efeito de estratificação dos municípios. Quanto à subamostragem de domicílios, também seria feita aleatoriamente e sem reposição independentemente em cada um dos setores remanescentes.

Conforme demonstrado em Moura e Silva (1989), verificou-se que o emprego de amostragem aleatória simples sem reposição para fazer subamostragem permite que os estimadores atuais da PME sejam mantidos. Além disso, tornou-se possível estimar o impacto da redução da amostra sobre a precisão das estimativas usando apenas os dados da amostra atual e informações sobre as porcentagens de redução a serem aplicadas.

5.2 – Impacto das Alternativas de Redução sobre as Estimativas

Conforme já dito na Seção 4 é preciso avaliar, para as diversas alternativas de redução, o impacto esperado na precisão das estimativas, revelado pelo acréscimo que deverão sofrer os respectivos coeficientes de variação. Baseado nisso torna–se possível então escolher uma alternativa que atenda às necessidades de redução (no todo ou em parte) sem comprometer a qualidade dos resultados da pesquisa.

Para estimar os coeficientes de variação das 35 variáveis escolhidas para estudo havia dois métodos. O primeiro seria simular uma ou mais subamostras para cada alternativa de redução e com base nelas calcular as estimativas e seus respectivos coeficientes de variação. O segundo consistia em estimar as componentes de variância a partir dos dados da amostra completa de outubro de 1987.

Devido ao segundo método fazer uso de um conjunto maior de observações do que o primeiro, tem–se como conseqüência maior precisão estatística da estimativa dos coeficientes de variação.

Isto, aliado ao fato do primeiro método depender dos valores particulares gerados em cada simulação, e ao elevado custo de efetuar a simulação, o que limitaria drasticamente o número de replicações, levou à escolha do segundo método para estimar o impacto da redução na precisão das estimativas.

Conforme Hansen e outros (1953, p.422) e Moura e Silva (1989), tem–se que as variâncias relativas das taxas ou totais que serão obtidos a partir da amostra **reduzida** da PME, com base nos estimadores de razão descritos na Seção 5.1, podem ser estimadas a partir da amostra **atual** da PME usando:

$$(\widehat{CV}^*)^2 = \frac{1}{\hat{Y}^2}\left[\sum_{h=1}^{H}\frac{V_h^2}{m_h^*} + \sum_{h=1}^{H}\frac{1}{m_h^*}\frac{1}{m_h}\sum_{i=1}^{m_h}\frac{N_{hi}^2}{\Pi_{hi}^2}\left(\frac{1}{n_{hi}^*}-\frac{1}{N_{hi}}\right)s_{hi}^2\right] \tag{6}$$

onde

H é o número de municípios de cada Região Metropolitana – RM;

m_h^* é o número de setores na amostra reduzida da PME no h–ésimo município da RM;

$\hat{Y}$ é o mesmo definido em (2);

$$v_h^2 = v_{hy}^2 + \hat{T}^2 v_{hx}^2 - 2\hat{T}v_{hxy} \tag{7}$$

$\hat{T}$ como definido em (5);

$$v_{hy}^2 = s_{hy}^2 - \frac{1}{m_h}\sum_{i=1}^{m_h}\frac{N_{hi}^2}{\Pi_{hi}^2}\left(\frac{1}{n_{hi}}-\frac{1}{N_{hi}}\right)s_{hiy}^2 \tag{8}$$

$$v_{hx}^2 = s_{hx}^2 - \frac{1}{m_h}\sum_{i=1}^{m_h}\frac{N_{hi}^2}{\Pi_{hi}^2}\left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}}\right)s_{hix}^2 \tag{9}$$

$$v_{hxy}^2 = s_{hxy}^2 - \frac{1}{m_h}\sum_{i=1}^{m_h}\frac{N_{hi}^2}{\Pi_{hi}^2}\left(\frac{1}{n_{hi}} - \frac{1}{N_{hi}}\right)s_{hixy}^2 \tag{10}$$

s_{hy}^2, s_{hx}^2 e s_{hxy} como definidos em (4)

$$S_{hixy} = \frac{1}{(n_{hi}-1)}\sum_{j=1}^{n_{hi}}(x_{hij} - \bar{x}_{hij})(y_{hij} - \bar{y}_{hij}) \tag{11}$$

$s_{hix}^2 = s_{hixx}$

$s_{hiy}^2 = s_{hiyy}$

n_{hi}^* é o número de domicílios na amostra reduzida no i–ésimo setor do h–ésimo município;

$$s_{hi}^2 = s_{hiy}^2 + \hat{T}^2 \cdot s_{hix}^2 - 2 \cdot \hat{T} \cdot s_{hixy} \tag{12}$$

Reescrevendo a expressão (6) de forma conveniente, obtém-se:

$$(\widehat{CV}^*)^2 = \frac{\sum_{h=1}^{H}\frac{v_h^2}{m_h^*}}{\hat{Y}^2} + \frac{\sum_{h=1}^{H}\frac{1}{m_h^*}\frac{1}{m_h}\sum_{i=1}^{m_h}\frac{N_{hi}^2}{\Pi_{hi}^2}\cdot\frac{s_{hi}^2}{n_{hi}^*}}{\hat{Y}^2} - \frac{\sum_{h=1}^{H}\frac{1}{m_h^*}\sum_{i=1}^{m_h}\frac{N_{hi}}{\Pi_{hi}^2}s_{hi}^2}{\hat{Y}^2} \tag{13}$$

Definindo–se as proporções de retenção da amostra por:

$$\alpha_h = \frac{m_h^*}{m_h} \quad \forall \quad h \in \{1,2,\cdots,H\} \tag{14}$$

$$\gamma_{hi} = \frac{n_{hi}^*}{n_{hi}} \quad \forall \quad i \in \{1,2,\cdots,m_h\} \quad \text{e} \quad \forall \quad h \in \{1,2,\cdots,H\} \tag{15}$$

e adotando–se a hipótese simplificadora

$$\gamma_{hi} = \gamma_h \quad \forall \quad i \in \{1,2,\cdots,m_h\} \quad \text{e} \quad \forall \quad h \in \{1,2,\cdots,H\}$$

verifica–se que as proporções de redução da amostra são dadas por

$(1-\alpha_h)$ - redução da amostra de setores no h–ésimo município;

$(1-\gamma_h)$ - redução da amostra de domicílios no h–ésimo município.

Além disso, a expressão (13) pode se reescrita como:

$$(\widehat{CV}^*)^2 = \frac{\sum_{h=1}^{H} \frac{v_h^2}{\alpha_h \cdot m_h}}{\hat{Y}^2} + \frac{\sum_{h=1}^{H} \frac{1}{\alpha_h \cdot \gamma_h \cdot m_h^2} \sum_{i=1}^{m_h} \frac{N_{hi}^2}{\Pi_{hi}^2} \cdot \frac{s_{hi}^2}{n_{hi}}}{\hat{Y}^2} -$$

$$- \frac{\sum_{h=1}^{H} \frac{1}{\alpha_h m_h^2} \sum_{i=1}^{m_h} \frac{N_{hi}}{\Pi_{hi}^2} s_{hi}^2}{\hat{Y}^2} \tag{16}$$

Ou ainda, definindo $\hat{C}_{1h}, \hat{C}_{21h}$ e $\hat{C}_{22h}$ de forma apropriada, tem-se:

$$(\widehat{CV}^*)^2 = \frac{\sum_{h=1}^{H} \frac{\hat{C}_{1h}}{\alpha_h}}{\hat{Y}^2} + \frac{\sum_{h=1}^{H} \frac{1}{\alpha_h \gamma_h} \hat{C}_{21h}}{\hat{Y}^2} - \frac{\sum_{h=1}^{H} \frac{\hat{C}_{22h}}{\alpha_h}}{\hat{Y}^2} \tag{17}$$

onde

$$\hat{C}_{1h} = \frac{v_h^2}{m_h} \quad \forall \quad h \in \{1, 2, \cdots, H\} \tag{18}$$

$$\hat{C}_{21h} = \frac{1}{m_h^2} \sum_{i=1}^{m_h} \frac{N_{hi}^2}{\Pi_{hi}^2} \frac{s_{hi}^2}{n_{hi}} \quad \forall \quad h \in \{1, 2, \cdots, H\} \tag{19}$$

$$\hat{C}_{22h} = \frac{1}{m_h^2} \sum_{i=1}^{m_h} \frac{N_{hi}}{\Pi_{hi}^2} \cdot s_{hi}^2 \quad \forall \quad h \in \{1, 2, \cdots, H\} \tag{20}$$

Dessas definições, observa-se que $\hat{C}_{1h}$ é um estimador da primeira componente de variância, isto é, da variância devida ao primeiro estágio de seleção (sorteio de setores), em cada um dos municípios. Já $\hat{C}_{21h}$ e $\hat{C}_{22h}$, podem ser combinadas de forma a produzir um estimador para a segunda componente de variância, isto é, para a variância devida ao segundo estágio (seleção de domicílios dentro dos setores). Essa combinação é obviamente dada por:

$$\hat{C}_{2h} = \hat{C}_{21h} - \hat{C}_{22h} \tag{21}$$

Deve ser notado que a fórmula (17) é exatamente a fórmula usada para estimar as variâncias relativas com a amostra atual da PME quando se tem $\alpha_h = \gamma_h = 1 \quad \forall \quad h \in \{1, 2, \cdots, H\}$, isto é:

$$\widehat{CV}^2 = \frac{\sum_{h=1}^{H} \hat{C}_{1h} + \sum_{h=1}^{H} \hat{C}_{21h} - \sum_{h=1}^{H} \hat{C}_{22h}}{\hat{Y}^2} \tag{22}$$

Agora, definindo-se:

$$\hat{V}_h = \hat{C}_{1h} + \hat{C}_{21h} - \hat{C}_{22h} \quad \forall \quad h \in \{1, 2, \cdots, H\} \tag{23}$$

$$\hat{\rho}_h = \frac{\hat{C}_{1h} - \hat{C}_{22h}}{\hat{V}_h} \quad \forall \quad h \in \{1, 2, \cdots, H\} \tag{24}$$

e substituindo–se os valores de (23) e (24) na fórmula (17), pode–se reescrevê-la como:

$$(\widehat{CV}^*)^2 = \frac{\sum_{h=1}^{H} \frac{\hat{V}_h}{\alpha_h \gamma_h}[1 + (\gamma_h - 1)\hat{\rho}_h]}{\hat{Y}^2} \tag{25}$$

Finalmente, pode–se então obter um estimador para o acréscimo que a redução da amostra acarretará sobre as variâncias relativas da amostra atual por:

$$\hat{A}^2 = \frac{(\widehat{CV}^*)^2}{\widehat{CV}^2} = \frac{\sum_{h=1}^{H} \frac{\hat{V}_h}{\alpha_h \gamma_h} \cdot [1 + (\gamma_h - 1)\hat{\rho}_h]}{\sum_{h=1}^{H} \hat{V}_h} \tag{26}$$

Ou ainda, definindo $\hat{Z}_h = \frac{\hat{V}_h}{\sum_{h=1}^{H} \hat{V}_h} \quad \forall \quad h \in \{1, 2, \cdots, H\}$ tem–se

$$\hat{A}^2 = \sum_{h=1}^{H} \frac{\hat{Z}_h}{\alpha_h \gamma_h}[1 + (\gamma_h - 1)\hat{\rho}_h] \tag{27}$$

5.3 – Estratégias para Redução da Amostra

Uma vez calculadas as estimativas $\hat{V}_h$ e $\hat{\rho}_h$ em cada município de uma determinada região metropolitana a fórmula (27) fornece estimativas dos acréscimos sofridos pelas variâncias relativas das 35 variáveis consideradas para qualquer conjunto de valores de $(\alpha_h, \gamma_h) \in \{(0;1) \times (0;1)\}, \quad \forall \quad h \in \{1, 2, \cdots, H\}$, que definem as alternativas de redução, desde que a proporção de redução de domicílios seja igual para todos os setores dentro de cada município, isto é, desde que $\gamma_{hi} = \gamma_h \quad \forall \quad i \in \{1, 2, \cdots, m_h\}$ e $\forall \quad h \in \{1, 2, \cdots, H\}$. A fórmula (27) não contempla o caso mais geral em que se permitem taxas diferenciadas de redução de domicílios nos setores de um mesmo município. Esta possibilidade foi descartada de imediato devido aos problemas operacionais que sua implantação em campo traria.

Diante deste conjunto de possibilidades de redução é natural então perguntar-se se é possível para cada RM encontrar um conjunto de alternativas de redução que minimizasse os acréscimos dos coeficientes de variação (na melhor das hipóteses para todas as variáveis ou em um conjunto apreciável ou "importante" destas), quando fixada uma proporção geral de retenção na RM. Estas proporções poderiam ser as sugeridas por Mansoldo (1988), no caso das regiões metropolitanas onde ocorrem problemas de sobrecarga, por exemplo.

Para as outras regiões, poderiam ser testadas alternativas de redução que mantivessem em níveis aceitáveis a precisão das estimativas. Aqui surgiram alguns problemas, já que os "níveis aceitáveis de precisão das estimativas" não estavam claros e objetivamente definidos, já que diferentes usuários, pesquisadores e mesmo os técnicos envolvidos com a PME podem ter opiniões diferentes sobre o assunto.

Examinando-se a fórmula (27) pode-se eliminar imediatamente as alternativas $(\alpha_h; \gamma_h)$ em que ambos $\alpha_h < 1$ e $\gamma_h < 1$, isto é, as alternativas em que num mesmo município é aplicada subamostragem de setores e de domicílios simultaneamente. Isto é possível porque sempre há alternativas melhores, pois se $\hat{\rho}_h > 0$ então é melhor fazer a redução no 2^o estágio, isto é, $\alpha_h = 1$ e $\gamma_h =$ (redução total), ou se $\hat{\rho}_h < 0$ é melhor fazer a redução no 1^o estágio, isto é, $\gamma_h = 1$ e $\alpha_h =$ (redução total).

Isto mostra que é sempre mais eficiente, para uma dada variável, reduzir a amostra apenas por meio da subamostragem de setores (ou alternativamente da subamostragem de domicílios) em cada município. Cabe registrar que, mesmo do ponto de vista operacional, a combinação de subamostragem de setores e domicílios num mesmo município é indesejável.

Devido à natureza multivariada do problema e à diversidade do conjunto das variáveis é praticamente impossível encontrar uma solução que seja ótima para todo o elenco de variáveis. Uma forma de contornar este problema seria construir uma "função perda" dos acréscimos dos coeficientes de variação, cujos pesos seriam atribuídos de acordo com uma escala de prioridade das variáveis. O conjunto de valores $S = \{(\alpha_h, \gamma_h), h \in \{1, \cdots, H\}\}$ que minimiza esta função seria então uma solução de compromisso, privilegiando as variáveis para as quais é necessário ter um maior controle sobre os acréscimos dos coeficientes de variação.

O problema de otimização pode ser visto sob outra ótica, determinando-se uma solução que minimizasse a taxa geral de retenção, tendo sido estipulado para cada variável um acréscimo do coeficiente de variação aceitável e prefixado.

A aplicabilidade de ambas as formas de busca de uma solução ótima depende do comportamento de cada variável em cada estrato quanto aos valores de $\hat{\rho}_h$. Quando, num mesmo município, não há nem mesmo predominância de sinal para os valores do $\hat{\rho}_h$ referentes às diversas variáveis consideradas no estudo, fica difícil determinar uma estratégia para redução da amostra nesse município. A estratégia que seria ótima para algumas variáveis seria ruim para as outras.

Os valores de $\hat{\rho}_h$ foram calculados com base na amostra de outubro de 1987 da PME para o elenco de 35 variáveis consideradas no estudo, e o exame de sua distribuição relevou que dentro dos municípios não há predominância de valores positivos ou negativos para todas as variáveis. Este resultado, aliado ao fato de que as estimativas $\hat{\rho}_h$ são pouco confiáveis no caso dos municípios que têm poucos setores na amostra, fez com que as idéias sobre otimização fossem abandonadas.

Decidiu-se, entao, por motivo de simplicidade operacional (manutenção da autoponderação da amostra), considerar apenas alternativas de redução que fossem uniformes em todos os municípios de uma mesma região metropolitana, isto é, alternativas nas quais:

$$\alpha_h = \alpha \quad \forall \quad h \in \{1, \cdots, H\}$$
$$\gamma_h = \gamma \quad \forall \quad h \in \{1, \cdots, H\}$$

Uma vez que a redução de setores era a que mais contribuía para a diminuição da sobrecarga de trabalho, devido reduzir as necessidades de deslocamento entre setores, resolveu-se analisar apenas duas famílias de alternativas:

- subamostragem apenas de domicílios;
- subamostragem apenas de setores, exceto nos municípios com menos de dez setores na amostra atual, onde seria feita subamostragem de domicílios.

Em relação à segunda alternativa, devido ao pequeno número de setores na amostra em alguns municípios (sendo que em alguns havia apenas um setor), resolveu-se que para os municípios com menos de dez setores na amostra atual a redução seria feita sempre por subamostragem de domicílios. Nos outros municípios, a subamostragem seria de setores, mantendo-se em cada RM a mesma taxa de retenção em todos os municípios. Com estas duas estratégias assegurava-se a manutenção da autoponderação, permitindo aproveitamento do sistema computacional de estimação hoje em uso.

A fim de eleger uma dessas alternativas de redução, foram calculados pelo "método das componentes" os acréscimos dos coeficientes de variação para cada uma das duas famílias de alternativas consideradas acima, com taxas de retenção variando de 0.95 a 0.50, em intervalos de 0.05, para cada região metropolitana.

Devido aos problemas encontrados com as estimativas de $\hat{\rho}_h$ para os municípios com poucos setores na amostra atual, foram feitas simulações das subamostras para avaliar a adequação do "método das componentes". Os resultados das simulações não diferiram muito daqueles obtidos pelo método das componentes para as taxas de retenção consideradas em estudo, flutuando pouco em torno dos valores encontrados por aquele método.

Após análise dos resultados dos acréscimos dos CVs para as 35 variáveis em cada uma das seis regiões metropolitanas, e das estimativas dos parâmetros de interesse, decidiu-se eleger a segunda alternativa de redução: subamostragem de setores com dez ou mais setores na amostra atual, e subamostragem de domicílios nos demais municípios, fixando-se a porcentagem de redução da amostra em 30%.

As razões para essa escolha podem ser mais bem compreendidas após análise dos resultados descrita na Seção 6 a seguir.

6 - ANÁLISE DOS RESULTADOS

Seguindo a metodologia descrita na Seção 5.2 foram calculadas estimativas de $\hat{\rho}_h$ para cada uma das 35 variáveis consideradas no estudo, em cada um dos municípios componentes da amostra da PME nas seis RMs que são cobertas por essa pesquisa. Os valores de $\hat{\rho}_h$ referentes à taxa de desemprego aberto podem ser encontrados no Anexo 5, a título de ilustração.

O Anexo 5 contém informações que resumem propriedades da distribuição de freqüências dos $\hat{\rho}_h$ calculados para o conjunto das 35 variáveis examinadas. Essas informações estão organizadas de forma a ilustrar as principais conclusões alcançadas na análise dos valores dos $\hat{\rho}_h$, que foram as seguintes:

- em cada município (ou estrato) não se pode apontar predominância de sinal, ao examinar os valores de $\hat{\rho}_h$ de todas as 35 variáveis;

- as estimativas de $\hat{\rho}_h$ para os municípios com poucos setores na amostra atual apresentaram problemas de estabilidade;

- as diversas variáveis apresentaram comportamento diferenciado com respeito ao sinal de $\hat{\rho}_h$.

As conclusões anteriores sobre os valores dos $\hat{\rho}_h$, aliadas a restrições de ordem operacional, levaram ao abandono da idéia de buscar alternativas de redução ótimas, conforme já relatado. Para orientar a escolha entre as duas famílias de alternativas que restaram, foram calculados os valores dos coeficientes de variação das 35 estimativas para cada RM, supondo taxas de retenção da amostra de 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55% e 50%. Esses coeficientes de variação referentes ao total Brasil podem ser encontrados no Anexo 6.

A análise dos coeficientes de variação obtidos com as duas famílias de alternativas para a redução da amostra revelou que:

- os acréscimos dos CVs em relação ao CV atual das estimativas crescem suavemente à medida que as taxas de retenção diminuem; de fato, parece razoável supor que $A = \frac{1}{\sqrt{k}}$ onde k é a taxa de retenção;

- os CVs obtidos com as formas alternativas de subamostragem para um mesmo nível da taxa de retenção são praticamente iguais, indicando que ambas são praticamente indiferentes do ponto de vista do impacto na precisão das estimativas;

- para variáveis cujos CVs são pequenos, a situação não piora muito, mesmo que a redução seja de 5% – por exemplo, um CV de 1% hoje em dia deveria crescer para algo como 1,4%

após uma redução de 50% da amostra; para variáveis cujo CV já não é pequeno, a situação é a mesma; e

- uma redução de 30% provocaria um acréscimo de cerca de 20% apenas nos coeficientes de variação atuais.

Com base nesses resultados, decidiu-se por uma redução de 30% da amostra em todas as seis regiões metropolitanas cobertas pela PME, utilizando a subamostragem de setores nos municípios com dez ou mais setores na amostra atual, e a subamostragem de domicílios nos demais municípios.

A opção de reduzir a amostra também nas Regiões Metropolitanas de São Paulo, Rio de Janeiro e Salvador, onde não foram verificados problemas de sobrecarga da rede de coleta, deveu-se à implantação de novos procedimentos de campo prevista para ocorrer junto da implantação da redução, à tentativa de diminuição dos custos da pesquisa (inclusive com a possibilidade de liberação de parte da equipe da PME nessas RMs para outras tarefas), e ao fato de que o tamanho absoluto da amostra no Rio de Janeiro e São Paulo, mesmo após a redução, ainda será próximo a 7 000 domicílios por mês.

Para Belo Horizonte e Porto Alegre, o percentual de redução proposto não dá conta de eliminar a sobrecarga de trabalho da equipe de campo, tendo sido sugerida a contratação de pessoal para completar a solução do problema.

7 - CONCLUSÕES E RECOMENDAÇÕES

A principal conclusão alcançada foi a proposta de metodologia para redução da amostra da PME, que na opinião dos autores atende aos requisitos impostos ao projeto:

- baixo custo de implementação;
- possibilidade de implementação a curto prazo (efetivada em agosto/88);
- manutenção do nível de precisão aceitável para as estimativas;
- eliminação total (ou quase total) da sobrecarga de trabalho das equipes de campo;
- redução do custo da pesquisa; e
- manutenção do sistema computacional.

Além disso, se a redução da amostra não pode ser vista como um ganho de qualidade sob a ótica do "erro amostral", certamente possibilitará a implantação de procedimentos de campo mais eficazes que deverão reduzir substancialmente as taxas de não-entrevista observadas na

PME. E também a liberação dos supervisores eventualmente deslocados para atividade de coleta permitirá aumentar o controle sobre o trabalho de campo, atividade que foi bastante prejudicada pela situação de sobrecarga vivida pelas equipes de algumas regiões metropolitanas.

É meta já fixada pelo corpo técnico e acertada com os coordenadores das equipes de campo a redução das taxas de não–entrevista para níveis nunca superiores a 5%.

Além do resultado pretendido, a execução do projeto de redução da amostra da PME evidenciou uma série de problemas cujo tratamento está fora do escopo deste trabalho, porém cuja solução poderá contribuir de forma significativa para a melhoria da qualidade dos resultados obtidos a partir dessa pesquisa.

O primeiro problema encontrado foi a precária documentação disponível sobre o desenho amostral e sobre a própria amostra selecionada, chegando mesmo a ser necessário **recalcular** as probabilidades de seleção dos setores selecionados já que as mesmas não se encontravam disponíveis. Este problema só não foi mais grave porque foi possível contar com o auxílio de algumas pessoas que já trabalham há vários anos com a PME.

O segundo problema foi a situação do sistema computacional, que não permite sua utilização para efetuar os cálculos necessários ao processo de estudo das alternativas para redução, além de calcular estimativas de taxas por um método, de valores absolutos (totais) por outro, e os respectivos CVs por outro método. Isto obrigou ao desenvolvimento de programas alternativos para expansão dos resultados e cálculo dos CVs, o que atrasou a conclusão dos trabalhos.

Com vistas à solução desse problema recomenda–se que, quando da abertura do projeto de Reformulação da PME, desenvolva–se novo sistema computacional, para o qual os programas de expansão elaborados podem servir de subsídio.

O terceiro problema está associado à metodologia adotada para expansão dos resultados, que ora incorpora correção para não–entrevista no cálculo das estimativas de valores absolutos, ora não incorpora essa correção no cálculo de taxas e dos CVs.

Numa pesquisa cujas taxas de não–entrevista chegaram a níveis tão elevados como a PME, esta situação pode provocar inconsistência entre os resultados. Ainda mais, se a ocorrência de não–entrevista não for uniforme entre os setores da amostra, parece inoportuno o uso da "autoponderação" para a expansão dos resultados.

Neste caso, recomenda–se fortemente que, além das medidas destinadas a diminuir o nível das taxas de não–entrevista, se procure unificar o procedimento de expansão, se possível incorporando (caso julgue ser adequado) procedimentos para correção de não–entrevista, e também abandonando de vez a **autoponderação** enquanto ferramenta para obtenção de pesos para a expansão da amostra.

Ainda mais, o processo de expansão atual da PME não faz qualquer uso do fato de que 75% da amostra permanece de um mês para o outro. É preciso investigar métodos de expansão que,

fazendo melhor uso dos dados disponíveis, permitam reduzir o "erro amostral" das estimativas obtidas a partir da PME.

Recomenda-se, ainda, que seja executado *monitoramento constante do trabalho de campo* a fim de assegurar a queda esperada nas taxas de não-entrevista, bem como para impedir que situações similares à atual se repitam.

Finalmente, recomenda-se, tão logo seja possível, que se abra linha de estudos visando à **reformulação** do desenho amostral da PME com vistas à sua implementação pós-Censo de 90, a fim de evitar que limitações de tempo imponham limites muito estreitos à capacidade de reformular a pesquisa.

8 - AVALIAÇÃO DOS EFEITOS DA REDUÇÃO SOBRE AS TAXAS DE NÃO-ENTREVISTA

Quando do momento da publicação deste trabalho, já haviam decorrido seis meses desde a implantação da redução da amostra, o que permitiu incluir um exame, ainda que breve, dos impactos dessa redução sobre as taxas mensais de não-entrevista, isto é, sobre a porcentagem de entrevistas não realizadas em relação ao total de entrevistas programadas mensalmente para a PME. O Anexo 7 contém gráficos que ilustram o comportamento dessas taxas mensais de não-entrevista para cada uma das seis regiões metropolitanas cobertas pela PME, no período que vai de janeiro de 1987 a fevereiro de 1989.

Analisando-se os gráficos apresentados no Anexo 7, observa-se que em todas as regiões metropolitanas houve, durante o ano de 1988, aumentos significativos no nível das taxas de não-entrevista, exceção feita para Belo Horizonte e Porto Alegre, onde essa elevação ocorreu antes e durante o ano de 1987.

Em todas as áreas, exceto São Paulo e Rio de Janeiro, os níveis das taxas de não-entrevista tornaram-se alarmantes e insustentáveis, chegando a um pico de mais 24% em Recife, no mês de maio de 1988. Níveis tão elevados de não-entrevista são capazes de comprometer seriamente a qualidade das estimativas, e não são refletidos adequadamente nas medidas de precisão amostral calculadas. Justificam, no entanto, a estratégia de buscar uma redução da amostra com a qual se pudesse buscar maior controle sobre a execução da pesquisa no campo, que permitisse reduzir os erros não-amostrais das estimativas.

Essa estratégia parece ter logrado bom êxito, depois da implantação da redução em agosto/88. Como se pode observar nos gráficos do Anexo 7, houve quedas significativas nas taxas de não-entrevista em todas as regiões metropolitanas desde então.

Depois da implantação da redução da amostra, as taxas de não-entrevista caíram para

níveis inferiores a 5%, exceção feita para Salvador e Belo Horizonte onde ainda superaram este patamar no período dezembro–88 a fevereiro–89, habitualmente caracterizado por uma elevação das taxas de não–entrevista devida a fatores de caráter sazonal tais como as festas de fim de ano e a grande incidência de férias do pessoal de coleta.

ANEXO 1

NÚMERO MÉDIO MENSAL DE DOMICÍLIOS PESQUISADOS NA PESQUISA MENSAL DE EMPREGO POR REGIÃO METROPOLITANA SEGUNDO O ANO DA PESQUISA - 1982-87

ANOS E VARIAÇÃO PERCENTUAL	Total	São Paulo	Rio de Janeiro	Belo Horizonte	Recife	Porto Alegre	Salvador
1982 (1)	42 739	8 884	9 070	6 795	5 846	7 432	4 712
1983	42 973	8 943	9 099	6 813	5 866	7 521	4 731
1984	45 315	9 411	9 502	7 213	6 181	7 806	5 202
1985	46 645	9 616	9 758	7 343	6 389	8 104	5 434
1986	46 739	9 703	9 838	7 517	6 256	8 056	5 395
1987 (2)	46 739	9 713	9 843	7 547	6 213	8 028	5 395
Variação Percentual 1982-87	9,3	9,3	8,5	11,1	6,3	8,0	14,5

FONTE - Mansoldo (1988).
(1) Média calculada com os valores mensais de maio a dezembro de 1982; (2) Média calculada com os valores mensais de janeiro a outubro de 1987.

ANEXO 2

NÚMERO DE ENTREVISTADORES DA PESQUISA MENSAL DE EMPREGO, POR REGIÃO METROPOLITANA, SEGUNDO O ANO DA PESQUISA - 1982-87

ANOS	NÚMERO DE ENTREVISTADORES DA PME (1)					
	São Paulo	Rio de Janeiro	Belo Horizonte	Recife	Porto Alegre	Salvador
1982	97	72	55	53	58	40
1983	101	72	55	53	54	42
1984	108	95	55	55	56	44
1985	105	89	58	49	50	44
1986	106	113	59	51	49	50
1987	106	114	42	52	49	50

FONTE - Mansoldo (1988).
(1) Dados obtidos a partir da variável código do entrevistador nos arquivos de dados da Pesquisa Mensal de Empregos.

ANEXO 3

RELAÇÃO DAS "VARIAVEIS" ESCOLHIDAS PARA AVALIAÇÃO DAS ALTERNATIVAS DE REDUÇÃO DA AMOSTRA DA PME

1 - Taxa de desemprego aberto.
2 - Taxa de desemprego aberto das pessoas que nunca trabalharam.
3 - Taxa de desemprego no setor da Indústria de Transformação.
4 - Taxa de desemprego no setor da Construção Civil.
5 - Taxa de desemprego no setor do Comércio.
6 - Taxa de desemprego no setor de Serviços.
7 - Taxa de desemprego no setor das Outras Atividades.
8 - Taxa de atividade.
9 - Taxa dos ocupados na Indústria de Transformação.
10 - Taxa dos ocupados na Construção Civil.
11 - Taxa dos ocupados no Comércio.
12 - Taxa dos ocupados em Serviços.
13 - Taxa dos ocupados em Outras Atividades.
14 - Taxa dos conta-própria sem rendimentos.
15 - Taxa dos desempregados e ocupados com menos de um salário-mínimo.
16 - Rendimento médio dos conta-própria.
17 - Rendimento médio dos empregados com carteira.
18 - Rendimento médio dos empregados sem carteira.
19 - Rendimento médio dos empregadores.
20 - População Economicamente Ativa.
21 - Pessoas ocupadas.
22 - Pessoas procurando trabalho.
23 - Pessoas procurando trabalho que nunca trabalharam antes.
24 - Pessoas procurando trabalho que já trabalharam anteriormente.
25 - Pessoas ocupadas na Indústria de Transformação.
26 - Pessoas ocupadas na Construção Civil.
27 - Pessoas ocupadas no Comércio.
28 - Pessoas ocupadas em Outras Atividades.
29 - Empregados com carteira.
30 - Empregados sem carteira.
31 - Conta-própria.
32 - Empregadores.
33 - Conta-própria com rendimento de 3 a menos de 5 Salários-mínimos.
34 - Pessoas ocupadas com 30 a 39 horas efetivamente trabalhadas em todos os trabalhos.
35 - Empregados com carteira que efetivamente receberam rendimento de 1 a menos de 2 Salários-mínimos.

ANEXO 4

VALORES DE ($) POR MUNICIPIO, PARA A TAXA DE DESEMPREGO ABERTO
POR REGIÃO METROPOLITANA

TABELA 1

MUNICIPIOS	REGIÃO METROPOLITANA DO RIO DE JANEIRO	
	Número de Setores na Amostra	Valores de ($)
Duque de Caxias	27	- 0,65
Itaboraí	5	- 0,45
Itaguaí	4	- 0,34
Magé	9	0,13
Mangaratiba	2	0,96
Maricá	2	0,90
Nilópolis	7	- 0,58
Niterói	19	0,31
Nova Iguaçu	51	- 0,38
Paracambi	1	-
Petrópolis	12	0,06
Rio de Janeiro	261	0,06
São Gonçalo	30	0,40
São João de Meriti	18	0,12

TABELA 2

MUNICIPIOS	REGIÃO METROPOLITANA DE RECIFE	
	Número de Setores na Amostra	Valores de ($)
Cabo	13	- 0,40
Igarassu	9	- 0,55
Itamaracá	2	- 3,46
Jaboatão	39	- 0,28
Moreno	4	- 0,03
Olinda	32	- 0,46
Paulista	23	0,11
Recife	144	0,11
São Lourenço da Mata	17	- 0,20

TABELA 3

MUNICIPIOS	REGIÃO METROPOLITANA DE BELO HORIZONTE	
	Número de Setores na Amostra	Valores de ($)
Belo Horizonte	231	- 0,05
Betim	11	- 1,83
Caeté	4	- 0,17
Contagem	37	- 0,07
Ibirité	5	0,13
Lagoa Santa	3	- 3,27
Nova Lima	6	0,23
Pedro Leopoldo	3	0,52
Raposos	1	-
Ribeirão das Neves	9	- 0,49
Sabará	8	- 1,18
Santa Luzia	8	- 0,86
Vespasiano	4	- 3,64

TABELA 4

MUNICIPIOS	REGIÃO METROPOLITANA DE SÃO PAULO	
	Número de Setores na Amostra	Valores de ($)
Barueri	3	- 6,66
Carapicuíba	6	0,05
Cotia	3	0,68
Diadema	8	0,25
Embu	3	- 4,44
Ferraz de Vasconcelos	2	- 0,58
Francisco Morato	1	-
Franco da Rocha	2	0,75
Guarulhos	19	- 0,05
Itapecerica da Serra	2	- 1,29
Itapevi	2	0,36
Itaquaquecetuba	2	- 3,34
Jandira	1	-
Mairiporã	2	0,68
Mauá	7	- 0,84
Mogi das Cruzes	7	- 0,29
Osasco	16	0,26
Poá	1	-
Ribeirão Pires	2	- 0,13
Santa Isabel	1	-
Santo André	19	0,15
São Bernardo do Campo	14	- 1,02
São Caetano do Sul	6	0,09
São Paulo	308	0,07
Susano	4	- 0,90
Taboão da Serra	3	0,57

TABELA 5

MUNICIPIOS	REGIÃO METROPOLITANA DE SALVADOR	
	Número de Setores na Amostra	Valores de ($)
Camaçari	14	- 0,09
Candeias	7	0,43
Itaparica	1	-
Lauro de Freitas	5	- 2,41
Salvador	183	- 0,05
São Francisco do Conde	3	- 18,23
Simões Filho	6	0,43
Vera Cruz	3	0,09

TABELA 6

MUNICIPIOS	REGIÃO METROPOLITANA DE PORTO ALEGRE	
	Número de Setores na Amostra	Valores de ($)
Alvorada	13	-
Cachoeirinha	9	- 0,15
Campo Bom	5	- 0,37
Canoas	33	0,23
Estância Velha	2	- 36,12
Esteio	8	0,39
Gravataí	19	0,27
Guaíba	10	0,18
Novo Hamburgo	20	- 0,27
Porto Alegre	187	- 0,01
São Leopoldo	15	- 0,60
Sapiranga	5	0,05
Sapucaia do Sul	12	- 1,42
Viamão	18	- 0,91

ANEXO 5

DISTRIBUIÇÃO DE FREQÜÊNCIA DOS SINAIS DOS ($)
CALCULADAS PARA AS 36 VARIÁVEIS

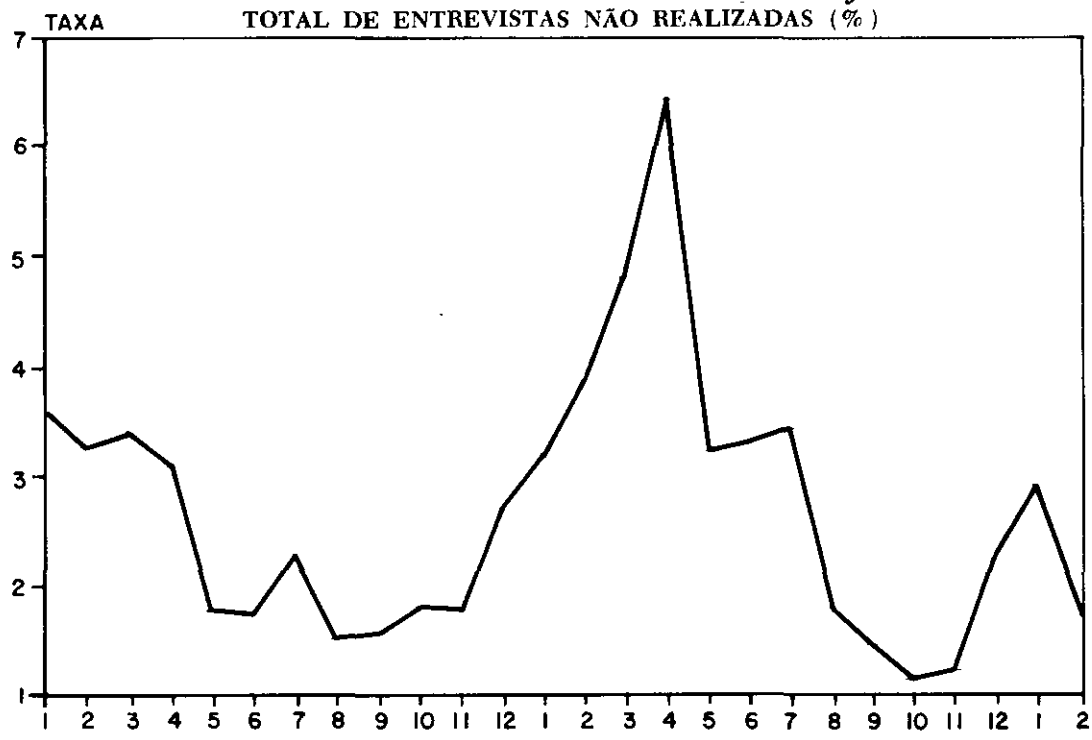
NÚMERO DE VALORES NEGATIVOS DE ($)	NÚMERO DE MUNICÍPIOS	PORCENTAGEM DO TOTAL
Total	84	100,0
8	1	1,2
9	2	2,4
11	4	4,8
12	2	2,4
13	1	1,2
14	4	4,8
15	4	4,8
16	6	7,1
17	7	8,3
18	8	9,5
19	8	9,5
20	4	4,8
21	3	3,6
22	4	4,8
23	6	7,1
24	4	4,8
25	3	3,6
26	1	1,2
28	4	4,8
31	1	1,2
36	7	8,3

ANEXO 6

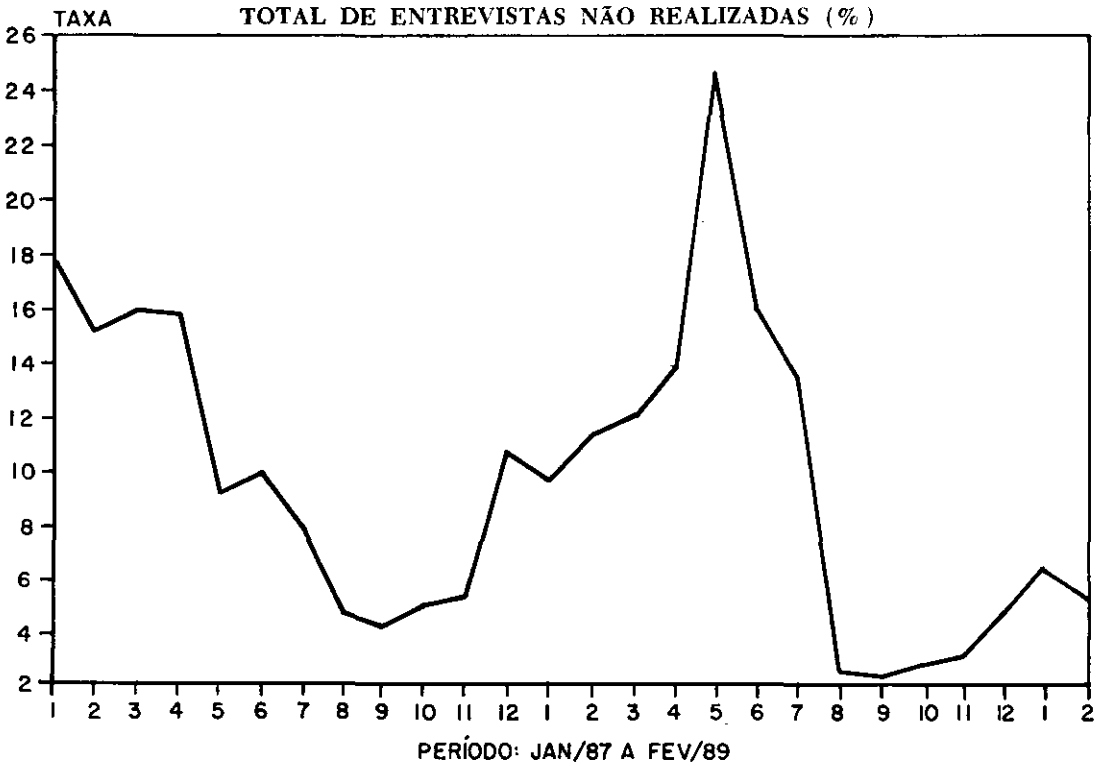
VALORES ESTIMADOS DO COEFICIENTE DE VARIAÇÃO DA TAXA DE DESEMPREGO ABERTO, NO BRASIL, SEGUNDO HIPÓTESES DE REDUÇÃO

PORCENTAGEM DE REDUÇÃO	REDUÇÃO APENAS NO NÚMERO DE SETORES	REDUÇÃO APENAS NO NÚMERO DE DOMICÍLIOS
Amostra Atual	2,68	2,68
5	2,75	2,75
10	2,83	2,83
15	2,91	2,91
20	3,00	3,00
25	3,10	3,09
30	3,21	3,20
35	3,32	3,32
40	3,47	3,46
45	3,63	3,61
50	3,80	3,79

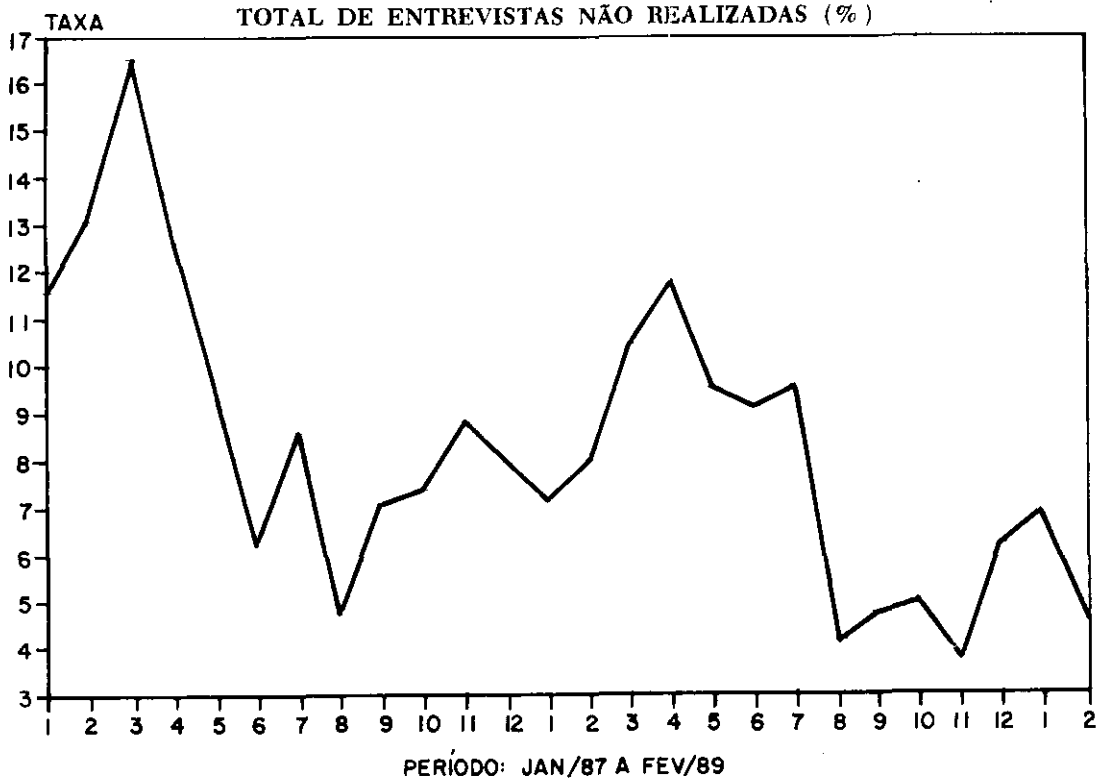
REGIÃO METROPOLITANA DO RIO DE JANEIRO



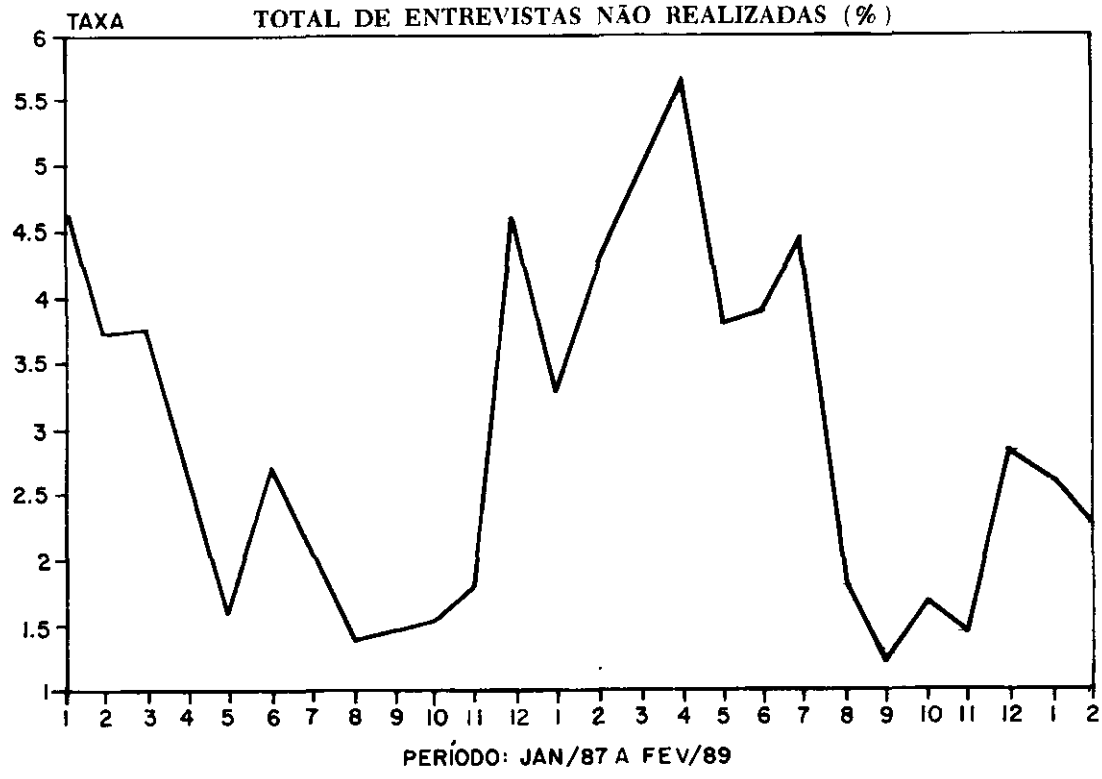
REGIÃO METROPOLITANA DE RECIFE



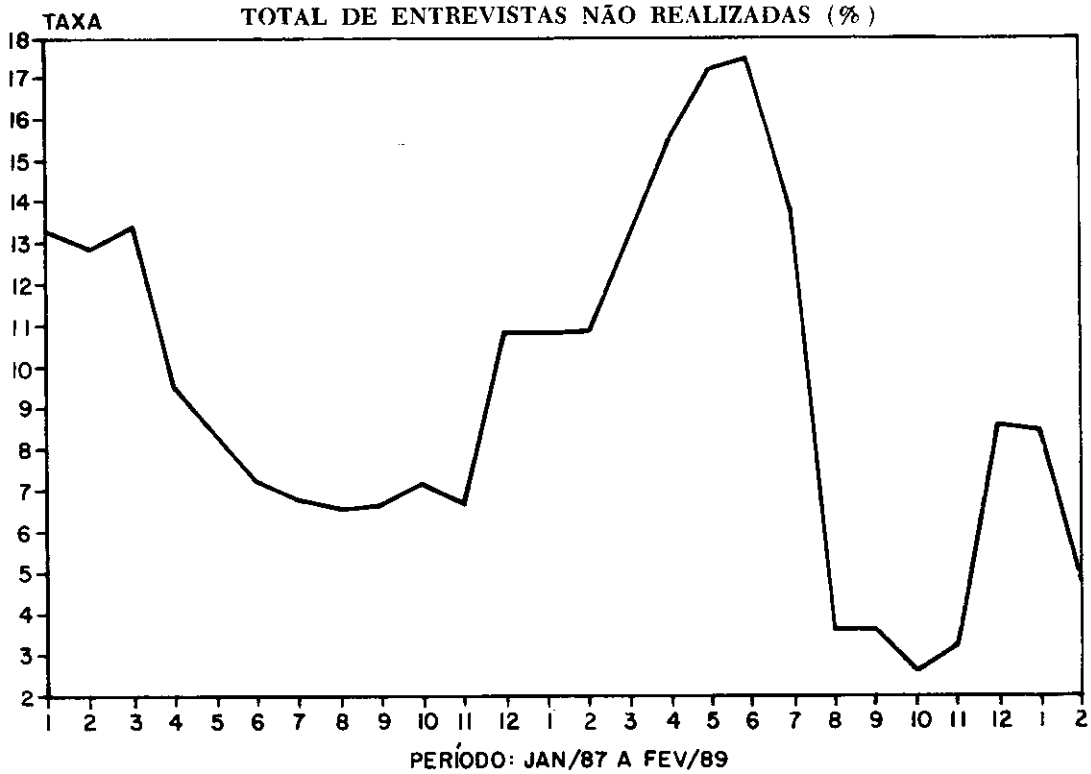
REGIÃO METROPOLITANA DE BELO HORIZONTE



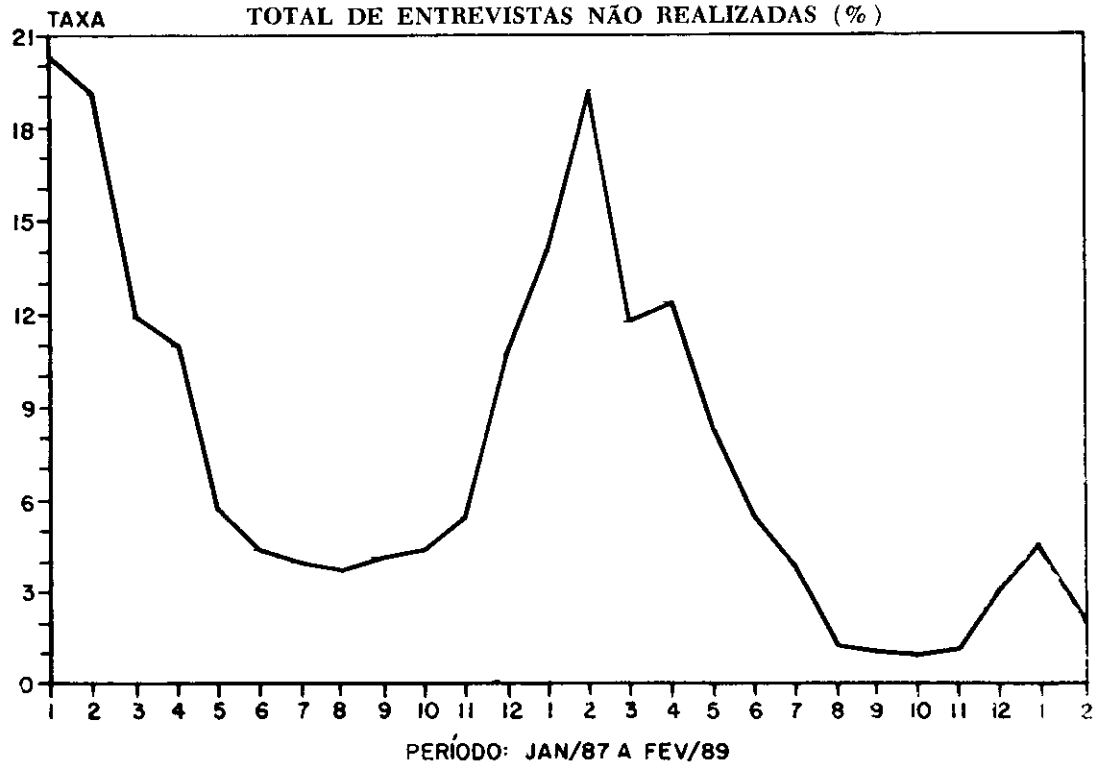
REGIÃO METROPOLITANA DE SÃO PAULO



REGIÃO METROPOLITANA DE SALVADOR



REGIÃO METROPOLITANA DE PORTO ALEGRE



BIBLIOGRAFIA

ALMEIDA, R. A. P. *Redução da amostra da pesquisa mensal de emprego.* Rio de Janeiro, IBGE, 1988.

COCHRAN, W. G. *Sampling techniques.* 3rd ed. New York, John Wiley, 1987. 428p.

FRIAS, L. A. M. *Determinação do limite superior ou inferior de curvas logísticas em projetos de população com base na tendência passada.* Rio de Janeiro, IBGE, 1987 (mimeo).

HANSEN, M. H.; HURWITZ, M. *Sample survey methods and theory.* New York, John Wiley, 1953. 2v.

MANSOLDO, M. S. *Caracterização dos fatores que contribuiram para a atual "sobrecarga" com os trabalhos de campo hoje executados pela pesquisa mensal de emprego (PME).* Rio de Janeiro, IBGE, 1988.

Metodologia da Pesquisa Mensal de Emprego 1980 – Rio de Janeiro, IBGE, 1983. v.2 (Série Relatórios Metodológicos.)

REIS, J. G. A. *Ordem de Serviço n. 01/88 da Diretoria de Pesquisa e Inquéritos,* Rio de Janeiro, IBGE, 1988.

NOTA

O presente texto foi elaborado com vistas a documentar ou registrar a metodologia adotada para a redução da amostra da PME. Ele incorpora idéias formadas em discussões sobre como reduzir o custo da coleta da PME das quais tomaram parte os integrantes da equipe do projeto, mas cuja expressão nesse texto baseia-se na interpretação pessoal dos autores.

Nesse sentido não devem os integrantes da equipe ser responsabilizados por omissões aqui identificadas.

Ao mesmo tempo, é fundamental registrar nossa gratidão para com toda a equipe que participou do projeto, pela dedicação, seriedade e competência demonstradas ao longo do trabalho.

RESUMO

Tendo em vista diminuir, principalmente, os encargos de coleta e aumentar o controle sobre os erros não amostrais da Pesquisa Mensal de Emprego do IBGE, decidiu-se estudar uma redução da amostra pesquisada mensalmente (cerca de 45 000 domicílios em 6 regiões metropolitanas: Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo e Porto Alegre).

Apresenta-se a metodologia utilizada para avaliar as diversas hipóteses consideradas para redução da amostra, os impactos esperados sobre as estimativas e sua precisão, e as conclusões quanto à alternativa a ser adotada.

ABSTRACT

In order to reduce costs (mainly in data collection) and improve control over nonsampling errors, it was decided to study a reduction in the sample of the Brazilian Monthly Survey of Employment, which actually includes about 45,000 households every month in 6 Metropolitan Areas, namely: Recife, Salvador, Belo Horizonte, Rio de Janeiro e São Paulo. The methods used to evaluate the various alternatives considered to reduce the sample are presented, together with estimates of the future impact of the proposed sample reduction over the precision of estimates from the reduced sample, including some concluding remarks about the final alternative adopted and finally some recommendations and directions for future work in the impprovement of the sample design used in the Brazilian Monthly Survey of Employment.

DISTRIBUIÇÃO ESPACIAL DA POPULAÇÃO BRASILEIRA E ALGUMAS CARACTERÍSTICAS SÓCIO-ECONÔMICAS ENTRE 1960-1980[1]

Antonio de Ponte Jardim*

1 – INTRODUÇÃO

O presente trabalho tem por objetivo descrever e regionalizar a ocupação demográfica do Território Nacional, com base em técnicas desenvolvidas por Meót[3], cuja finalidade é de melhor avaliar os aspectos gerais da ocupação demográfica de um determinado espaço territorial, seja este um país ou uma região geográfica. Essas técnicas permitem a construção de um índice demográfico que possibilita analisar a evolução e as modalidades da ocupação demográfica, não só do espaço territorial como um todo, mas também de suas características urbana e rural. Possibilitam, ainda, a aplicação de outros indicadores demográficos que facilitam a regionalização do espaço geográfico em zonas de ocupação conforme o tipo de ocupação, permitindo, assim, que se estabeleçam relações inter e intra cada zona de ocupação e as características sócio-econômicas e demográficas das respectivas populações. O índice tem uma construção simples cuja base é a densidade demográfica de cada unidade geográfica, que associado a critérios sócio-econômicos e demográficos pode contribuir para uma melhor compreensão do processo de distribuição e redistribuição da ocupação demográfica do Território Nacional.

* Sociólogo e Analista Especializado do Departamento de População - DEPOP-IBGE[2]

Tendo em vista a natureza, vantagens e limitações do índice, supõe-se que este seja útil para avaliar diferenças espaciais no processo de ocupação territorial, como também poderá servir de base para a tomada de algumas estratégias do desenvolvimento regional a partir do quadro detectado pela relação **População** e **Desenvolvimento**

No caso da distribuição e ocupação do espaço territorial brasileiro, utilizou-se a técnica para as Unidades da Federação[4] com base na informação dos três últimos Censos Demográficos (1960, 1970, 1980), período em que houve um maior ritmo de concentração da população brasileira em áreas urbanas e em especial nas regiões metropolitanas. Este trabalho divide-se em cinco partes, onde apresenta, em primeiro lugar, sumariamente os objetivos gerais e específicos; em segundo lugar, os antecedentes metodológicos, em terceiro lugar, analisam-se as características de ocupação da população brasileira entre 1960-1980, enfatizando-se os aspectos de concentração da população urbana, especificamente aquela residente em centros urbanos de 10 000 e mais habitantes. Em quarto lugar, analisam-se, para as zonas de ocupação, as características sócio-econômicas da população em função da distribuição da PEA por setores de atividade, dos índices de alfabetização e da superfície aproveitada com agropecuária. Finalmente, conclui-se que a dinâmica do desenvolvimento sócio-econômico está intimamente ligada a formas desiguais de ocupação demográfica regional e que, de acordo com as características e modalidades desse desenvolvimento, a tendência predominante é de concentrar-se cada vez mais a população em poucos grandes centros urbanos, fazendo-se com que não só aumentem as divergências regionais e intra-regionais, como também a subocupação relativa do Território Nacional.

2 – ANTECEDENTES METODOLÓGICOS

As condições de ocupação demográfica do território podem ser estudadas a partir de indicadores que se referem tanto à população total como à população urbana e rural. Entretanto, o indicador tradicional de ocupação demográfica é a densidade demográfica, que por ser um indicador bruto não oferece muitas vantagens devido à enorme dispersão que este possa apresentar. Isto é, varia enormemente em função da relação população/território. Assim que, devido a esse problema e como uma forma de se obterem indicadores com menor variabilidade, optou-se pela utilização do índice relativo de ocupação demográfica do território (I_i) que possui vantagem de, além de permitir a comparação relativa da ocupação demográfica entre unidades estudadas, possibilita, também, agrupar essas Unidades Geográficas em classes mais homogêneas[5]. Desse modo teremos:

$$I_i = \frac{\text{Densidade da Unidade Geográfica "}i''}{\Sigma\text{Densidade das Unidades Geográficas "}i''} = \frac{D_i}{\Sigma D_i}$$

Observa-se que a soma dos valores de I_i, cuja soma é igual a 1 ou 100 por cento, mostra os graus relativos de homogeneização da população no espaço analisado em cada uma das Unidades Geográficas.

Este indicador permite, ainda, observar o grau de ocupação relativa da população que vive nas distintas Unidades Geográficas de análise em relação à população total. Aplicando-se sucessivamente aos espaços urbano e rural, tem-se uma medida que fornece uma imagem quantificada da repartição territorial da população nesses espaços.

Com efeito, é preciso ter presente que os indicadores de ocupação demográfica dependem, por definição, do somatório das densidades altas, sendo importante estudar o efeito dos valores altos sobre a distribuição. Se existe, no conjunto, uma unidade espacial com densidade acentuadamente mais alta que as demais, o seu I_i será superior a $1/n$, onde n é o número de Unidades Geográficas e os demais serão, provavelmente, inferiores a $1/n$.

Na prática, considera-se que uma densidade é muito alta quando esta é igual ou superior a 10% do somatório das densidades (ΣD_i) das respectivas Unidades de Observação. Considera-se ainda a existência de unidades superconcentradas quando, e somente, ocorram, no máximo, dois valores altos de densidade. Com efeito, se existirem pelo menos três valores altos, esses não poderão ser considerados casos especiais, e passariam a constituir um grupo de valores altos no conjunto das unidades.

Nestas condições o método requer:

a) eliminação provisória das unidades urbanas ou rurais, com densidades excepcionalmente mais altas que as demais, constituem áreas de superconcentração demográfica, portanto, devem ser descartadas provisoriamente. Estima-se a informação dessa Unidade Geográfica ou área quando tem um valor alto em relação às demais densidades, ou quando o valor do indicador I_i (rural ou urbano) for superior a 10% do valor de $\sum D_i$, efetua-se novamente, sem as respectivas densidades. Uma vez aplicado este critério, obtém-se uma maior homogeneização entre os índices de ocupação demográfica[6].

b) grau de heterogeneidade de ocupação do espaço rural:

$$_nI_{\gamma_i} = \frac{D_{\gamma_i}}{\frac{\Sigma D_{\gamma_i}}{n}} \quad \text{ou} \quad n \cdot \frac{D_{\gamma_i}}{\Sigma D_{\gamma_i}}$$

onde: D_γ = densidade do espaço rural total = $\frac{\text{População rural total}}{\text{Superfície total}}$

c) grau de heterogeneidade de ocupação do espaço urbano:

$$_nI_{i_u} = \frac{D_u}{\Sigma D_{u_i}}$$

onde: $D_{u_i} = \frac{\text{População urbana total}}{\text{Superfície total}}$

ΣD_{u_i} = Somatório das densidades urbanas das distintas unidades territoriais

d) coeficiente urbano-rural de heterogeneidade de ocupação:

$$K = \frac{\bar{D}_u}{\bar{D}_\gamma} \quad \text{onde} \quad \bar{D}_u \quad \text{e} \quad \bar{D}_\gamma \quad \text{são as médias aritméticas de} \quad D_{u_i} \quad \text{e} \quad D_{\gamma_i}.$$

ou seja:

$$\bar{D}_u = \frac{\Sigma D_{u_i}}{n} \quad \text{e} \quad \bar{D}_\gamma = \frac{\Sigma D_{\gamma_i}}{n}$$

k - representa o grau médio de heterogeneidade que existe entre as condições de povoamento dos espaços urbano e rural.

Significado:

a) um valor de NI expressa as disparidades nas densidades das unidades territoriais, e essas serão maiores à medida que o valor de NI se distancia de 1;

$nI < 1$: significa que a heterogeneidade provém de uma subocupação da maior parte do território.

b) classificação das Unidades Geográficas em função dos tipos de ocupação do espaço urbano e rural

A análise no interior do país:

i) - Se $I_i < 1/n$ (sendo n o número de Unidades Consideradas) a Unidade Geográfica apresenta uma condição de subocupação relativa:

ii) - Se $I_i > 1/n$ existe uma sub-reocupação relativa.

Observa-se que o valor de $1/n$ indicaria o valor de I, se o território fosse ocupado homogeneamente com densidades iguais em todas as Unidades Geográficas.

Esse índice de ocupação demográfica, aplicado aos espaços urbano e rural, nos permite classificar as Unidades Geográficas em quatro zonas de ocupação, segundo os respectivos valores dos índices:

a) **Zonas de Ocupação Total**: onde os índices urbano e rural são superiores ou iguais a $1/n$, o que significa que existe uma sobreocupação relativa dos dois espaços e, portanto, de população total;

b) Zona de Ocupação Parcial: quando um dos dois indicadores é superior ou igual a $1/n$ e o outro é inferior a $1/n$;

c) Zonas de Subocupação: onde os dois indicadores são inferiores a $1/n$, condição de subpovoamento relativo total. Esta zona pode ser denominada de zona de subocupação parcial; e

d) finalmente, existe um quarto tipo de zona de ocupação, derivada da zona de subocupação, quando a densidade da população rural é inferior a 1 hab./km^2. Trata-se de zonas vazias, onde os dois indicadores sejam $1/n$, mas praticamente não existe população rural.

2.1 – Outras Medidas Resultantes (para o País como um todo)

São medidas que permitem comparações entre países ou para o mesmo país, em épocas distintas.

a) *grau de heterogeneidade de ocupação do espaço total:*

$$nI = \frac{nD_i}{\Sigma D_i} \quad \text{onde} \quad D_i = \frac{\text{População total}_i}{\text{Superfície total}_i}$$

$nI > 1.$ *: trata-se do oposto, isto é, o território está sobreocupado na maior parte da sua superfície.*

b) $\mathbf{k = 1.}$ *: existe homogeneidade nas condições de povoamento dos espaços urbano e rural;*

$$k > 1: \quad \text{a heterogeneidade de que} \quad \bar{D}_u > D_\gamma$$
$$k < 1: \quad \text{a heterogeneidade de que} \quad \bar{D}_\gamma > \bar{D}_u$$

2.2 – Limitações

Ainda que o índice relativo de ocupação demográfica compare o nível de ocupação populacional do território, há algumas limitações gerais que devem ser consideradas:

1 - Como o indicador de ocupação demográfica do território se baseia nas densidades de população urbana e rural em relação à superfície da Unidade da Federação correspondente, uma análise comparativa dos índices de ocupação demográfica do território (I_i) por áreas

independentes tem a desvantagem de se relacionarem, em cada caso, diferentes superfícies e populações. Seria recomendável não calcular o I_{u_i} (índice de ocupação demográfica urbana) e o I_{γ_i} (índice de ocupação demográfica rural) em relação às respectivas Unidades de Observação mas, sim, em relação à mesma Unidade da Federação e comparar esses resultados com as demais Unidades Federativas (UFs);

2 - Quando se supõe que em uma distribuição eqüitativa os valores de I_i deveriam ser iguais a $1/n$, estamos apenas considerando a população com relação a uma superfície, mas podemos encontrar casos em que a população parece estar distribuída eqüitativamente nas áreas urbanas e rurais. Porém, por ter cada área diferentes superfícies, resulta, oculta, verdadeira densidade, isto porque o valor de I_i para cada área pode ser parecido, mas não se sabe se a superfície territorial de cada uma possui a população concentrada ou dispersa; e

3 - Pelo fato da existência das múltiplas diversidades demográficas e sócio-econômicas existentes no país, seria recomendável calcular o índice de ocupação demográfica do território (I_i) para municípios e agrupá-los por Grandes Regiões geográficas, pois estampariam com maior detalhe a heterogeneidade demográfica e sócio-econômica existente no interior do país.

3 – OCUPAÇÃO ESPACIAL DA POPULAÇÃO BRASILEIRA

Nesta parte, analisaremos as modalidades de ocupação do Território Nacional através dos indicadores de heterogeneidade demográfica. Esses são indicadores sugestivos quando analisados ao nível das situações urbana e rural, tendo em vista a medição dos graus de heterogeneidade relativa da ocupação demográfica dos espaços urbano e rural. Outrossim, servem para medir os graus de concentração da população ao longo do período em questão. Nesse sentido, se intentará inicialmente descrever o desenvolvimento da ocupação espacial da população brasileira como um todo, em seguida para a zona de ocupação, e finalmente tratar da concentração e dispersão da população urbana, principalmente daquela que vive em cidades de 10 000 e mais habitantes.

Os dados relativos ao Censo de 1960 constituem um marco de referência, período em que se intensifica o processo de crescimento urbano no Brasil.

Nos últimos 20 anos, a relação entre população e território praticamente dobrou, passando de 8 para 14 habitantes por km^2 em 1980. Embora essa relação esteja ligada ao crescimento populacional que se distribuiu diferencialmente ao longo do Território Nacional, o Brasil ainda continua apresentando características de subpovoamento na maior parte do seu território.

O processo de concentração da população urbana alterou a dinâmica populacional com novas modalidades de ocupação e distribuição espacial da população, seja no campo ou na

cidade, como podemos observar na Tabela 1, que nos mostra a crescente divergência entre as respectivas densidades médias.

Esse aumento relativo da densidade média urbana mostra não só as modalidades do desenvolvimento nacional nesse período, como também é o reflexo direto das variações na estrutura do mercado de trabalho que se tornou, em grande parte, urbano a partir da década de 60. Tais condições passaram a atrair e a "exigir" crescentes contingentes de força de trabalho para as atividades econômicas nos grandes centros urbanos, enquanto no campo registrava-se um decréscimo relativo da participação dos economicamente ativos. Para a ocupação do território, o aumento de densidade média urbana nos leva a inferir que houve um aumento da heterogeneidade urbana em relação à rural. Como ainda podemos constatar na Tabela 1, já em 1960, quando mais da metade da população brasileira ainda se concentrava na área rural, o coeficiente de heterogeneidade urbano/rural era maior do que 1. Isto significa que a densidade média urbana era bem superior à rural. Por outro lado, o aumento da heterogeneidade do espaço urbano está associado às características e às diferenciações no processo de concentração da população urbana, o qual possivelmente reflete padrões de desenvolvimento econômico regionalmente distintos.

Atribui-se a influência desses padrões ao processo acelerado do desenvolvimento econômico na década de 50 que atraiu com maior intensidade para os principais centros industriais e administrativos grande parte da população nacional. A título de exemplo, São Paulo e Rio de Janeiro concentravam, em 1960, 27,8% da população brasileira, enquanto que em 1980 essas duas UFs passaram a deter 30,5% da população nacional.

Essas Unidades da Federação eqüidistavam 19 e 6,5 vezes da densidade média nacional em 1960, passando a representar, em 1980, 18,6 e 7,2 vezes a média do país, respectivamente. Esses resultados são bastante significativos para ilustrar, a nível global, aspectos da heterogeneidade da ocupação demográfica do Território Nacional, sua manutenção e até mesmo a sua evolução ao longo das duas últimas décadas. Como veremos, esse processo de concentração populacional tem conduzido o país a um crescente processo de "metropolização hipertrofiada" onde, em contrapartida, o espaço ainda se apresenta parcialmente constituído por vastas zonas subpovoadas, com estruturas espaciais que não permitem uma integração rural-urbana mais equilibrada. Esta situação contribuiu para que aumentassem progressivamente as divergências entre as condições de vida e os níveis de desenvolvimento sócio-econômico, reproduzindo segmentos populacionais apenas marginalmente vinculados à dinâmica econômica modernizante, segmentos esses que passaram a ocupar espaços menos integrados pelo referido processo.

Sabe-se também que a análise das condições de ocupação de um território nos permite conhecer a evolução recente e passada da distribuição espacial da população anteriormente descrita através dos resultados censitários. Desse modo, essa metodologia permite analisar a evolução das modalidades da ocupação espacial não só no território como um todo, mas

também nas suas subáreas urbana e rural.

A análise metodológica dos índices de Meót[7] permite ainda aplicar outros indicadores de ocupação demográfica do território, a partir de cujos valores é possível regionalizar o país em zonas de ocupação de acordo com o tipo de ocupação. Estas zonas, por sua vez, estão subdivididas em subzonas, mediante o grau de dispersão (urbano e rural) que apresenta cada Unidade da Federação (UF). Seguindo esses passos, torna-se possível estabelecer uma relação entre cada zona e as características sócio-econômicas da população que a habita.

3.1 – Índices de Ocupação Demográfica do Território

O índice de ocupação demográfica do território permite-nos medir graus relativos de concentração de população em cada Unidade Geográfica[8]. Possibilita, ainda, fazer uma análise da repartição das populações nesses espaços, mediante a classificação do espaço nacional em grandes zonas de ocupação demográfica que, como vimos, subdividem-se em: total e subocupada, sendo que esta pode ser subdividida em parcial e vazia.

Esse índice permite mensurar de forma mais clara e objetiva a distribuição da ocupação demográfica da população por Unidade da Federação, dado que não considera a superfície territorial fisicamente inaproveitável e não depende dos limites arbitrários das classes de densidades demográficas. Portanto, pode-se aplicá-lo tanto ao total das UFs quanto às suas respectivas áreas (urbana e rural), permitindo, ainda, medir e analisar o grau de heterogeneidade relativa das mesmas em relação ao espaço total.

É importante ressaltar que as condições de ocupação demográfica do território estão determinadas pela evolução histórica dos processos econômicos e sociais que condicionam a concentração da população, acarretando a existência de fortes fluxos migratórios no sentido rural-urbano e de cidades ou localidades urbanas menores para cidades maiores, a ritmos cada vez mais acelerados.

Essa situação determina algumas conseqüências que podem ser resumidas em 3 aspectos[9] principais:

a) deserção do campo e subaproveitamento das potencialidades do espaço rural;

b) existência de vazios demográficos, como as zonas de fronteiras, ao lado de outras zonas saturadas; e

c) na satisfação das necessidades básicas da maioria da população, assim como implicações sobre a existência e crescimento de setores marginais urbanos.

A configuração da distribuição da população brasileira ao longo dos últimos 20 anos ilustra essa concentração populacional. Assim, já em 1960, quando mais da metade da popu-

lação brasileira ainda estava na área rural, somente 11 Unidades da Federação (UFs), com características de sobreocupação em seus espaços urbano e rural (zona de ocupação total), detinham 58,2% da população em apenas 13,1% do Território Nacional. Enquanto isso, a superfície restante apresentava características de subpovoamento, predominando a existência das chamadas zonas vazias (56,6% da superfície), ou seja, aquelas com menos de 1 hab./km^2 (Tabela 2).

Duas décadas depois essa situação apresenta-se da seguinte maneira:

• Permaneciam praticamente as mesmas UFs que apresentavam sobreocupação total nos espaços urbano e rural em 1960, com exceção do Paraná;

- Em 1980, as UFs com ocupação total passaram a concentrar 57,8% da população total em apenas 12,2% do Território Nacional, ou seja, com um espaço relativo menor, o qual passou a deter 63,7% da população urbana brasileira;

- Embora tenha decrescido ligeiramente a porcentagem de população residente na zona de ocupação total em relação a 1960 (58,2% para 57,8%), o incremento populacional absoluto atingiu a magnitude de 28 milhões de habitantes entre 1960-1980, ou seja, com o incremento da ordem de 69,0%; e

- A população urbana passou de 21,4 milhões em 1960 para 31,3 milhões em 1980, representando um incremento relativo de ordem de 46,3%.

Como podemos observar, essa situação teve seus reflexos diretos sobre o aumento do grau relativo de concentração da população urbana[10] que, na zona de ocupação total, em 1960, já era de 1,19 e passou para 1,88 em 1980, enquanto que nas zonas de subocupação o grau de concentração urbana evoluiu de 0,19 em 1960 para 0,80 em 1980, o que revela também um considerável incremento relativo (vide Tabela 2).

Os dados anteriores podem ser considerados ilustrativos para explicar os níveis de concentração relativa da população brasileira, assinalando tanto a concentração urbana como a regional, destacando-se nesse sentido o fato de que o ritmo relativo de concentração foi mais intenso nas chamadas zonas subocupadas (parcial e vazia) do que na própria zona de ocupação total.

Para o país, o aumento da concentração relativa da população está intimamente ligada aos seguintes fatores:

a) *aumento progressivo da sobreocupação relativa das seguintes Unidades da Federação: Rio* de Janeiro, São Paulo, Pernambuco e Distrito Federal, que contribuíram para aumentar não só o índice de heterogeneidade entre elas, como também o grau de heterogeneidade demográfica do espaço nacional;

b) progressivo grau de urbanização decorrente dos fluxos migratórios para os centros ur-

banos, que contribuíram para que aumentasse a sobreocupação do espaço urbano, e, conseqüentemente, contribuíram com um maior grau de heterogeneidade urbana do país; e

c) crescente metropolização que, intensificada pelos fenômenos anteriormente citados, passou a concentrar nas metrópoles 25,6% e 29,0% da população brasileira em 1970 e 1980.

Através dos aspectos ressaltados nesses três itens, sintetizam-se as características básicas das zonas de ocupação total e parcial. Em relação à zona de ocupação total, observa-se que a mesma vem apresentando, desde 1960, um alto índice de heterogeneidade relativa e um grau de heterogeneidade demográfica dos respectivos espaços urbano e rural em relação às demais zonas (Tabelas 3 e 4).

Dentre as outras zonas de ocupação, a zona de ocupação parcial foi a que apresentou maiores mudanças na estrutura espacial da população entre 1960 e 1980. Em 1960, essa apresentava sobreocupação na área rural na maioria das UFs que a compunha; em 1980, essa sobreocupação passou a ser preponderantemente urbana. Esse aspecto mostra não só o processo de expulsão da população das áreas rurais, como também a influência dos padrões genéricos das zonas de ocupação total. No primeiro caso, destacam-se os Estados do Paraná e Minas Gerais, sendo que as mudanças no campo se deram com maior intensidade no Paraná que passou para zona de ocupação total a partir de 1970. No segundo caso, a transferência do Distrito Federal para Brasília contribuiu, em parte, para que houvesse uma maior sobreocupação urbana nessa zona de ocupação, causada principalmente pelos grandes contingentes populacionais que para ela se dirigiram. Essa situação fez com que, inclusive, em 1970, Brasília passasse para zona de ocupação total, embora a área rural apresentasse uma grande heterogeneidade. È importante destacar a influência das "Cidades-satélites" que contribuíram para que a Capital Federal retornasse em 1980 para zona de ocupação parcial. O traço diferencial pode ser observado no fato de que ela se tornou a UF que apresentou, em 1980, o mais alto grau de heterogeneidade relativa do país, influenciada principalmente pelo crescimento das periferias urbanas ("Cidades-satélites").

Por outro lado, as zonas de subocupação e vazia, apesar de apresentarem o mais baixo grau de heterogeneidade urbana e rural ao longo do período em questão, sofreram grandes mudanças na distribuição espacial, principalmente quando comparadas às zonas de ocupação total e parcial. O grau de heterogeneidade demográfica aumentou entre as zonas de ocupação, principalmente em relação aos respectivos espaços urbanos e, para as UFs, destacaram-se os Estados de Goiás e do Acre em 1980, que foram os que apresentaram maior grau de heterogeneidade dos espaços urbano e rural. Cabe aqui ressaltar que as UFs que compõem as zonas de subocupação parcial e vazia, apesar de possuírem o mais baixo grau de heterogeneidade demográfica em seus respectivos espaços urbano e rural em relação às demais UFs do país, registraram um grau relativo de urbanização (população urbana sobre o total) crescente, ex-

ceto para Rondônia e Pará, que em 1980 apresentaram população nas áreas rurais superior à de 1970.

Outro aspecto importante derivado da observação anterior é que a zona vazia apresentou em 1980 um grau de urbanização superior à zona de ocupação parcial, com 58,9% contra 48,6%. Isto significa que, embora não apresentassem diferenciação no baixo grau relativo de concentração urbana como um todo, as UFs de Amazonas e Roraima passaram a concentrar maior percentagem de suas respectivas populações nas áreas urbanas. Estas, que em 1960 apresentavam mais de 60% de sua população urbana residindo em centros urbanos de 10 000 e mais habitantes (66,1% e 83,8% respectivamente), passaram a concentrar, em 1980, mais de 80%, ou seja, 86,2% respectivamente.

Outro aspecto importante a considerar é a crescente metropolização da população brasileira. Em 1970, 25,6% da população nacional viviam nas nove regiões metropolitanas, percentual que evoluiu para 29,0% da população do país em 1980. Esse incremento respondeu por mais de 40% do crescimento da população urbana entre 1970 e 1980. Tais dados contribuem efetivamente não só para situar os fatores que explicam o aumento da sobreocupação do espaço urbano brasileiro, como também espelham as condições assumidas pela dinâmica de distribuição e concentração da população no Território Nacional.

Como podemos constatar na Tabela 5 a taxa de crescimento médio geométrico anual da população residente nas regiões metropolitanas entre 1970-1980 foi superior à taxa de crescimento da população total e urbana das zonas de ocupação consideradas, sendo que as regiões metropolitanas da zona de subocupação (urbana) foram as que apresentaram as maiores taxas de crescimento populacional. Entretanto, apesar da zona de ocupação total ter crescido a uma taxa inferior à do crescimento médio anual brasileiro, ainda detinha, em 1980, 57,8% da população nacional. Dessa forma, durante a década de 70, revelou-se ainda um razoável crescimento populacional nas regiões metropolitanas, principalmente nas de subocupação e manteve-se, ainda, altos valores absolutos nas de ocupação total.

Em resumo, esses resultados são também importantes para uma reflexão acerca de um fenômeno complementar, ou seja, a relativa subocupação demográfica do espaço rural, com seus reflexos sobre o nível de ocupação do Território Nacional.

Na parte seguinte, mostrar-se-ão alguns aspectos específicos da concentração da população urbana nos últimos 20 anos, como forma de reunir outros elementos que ajudem a interpretar as modalidades internas de sobreocupação do espaço urbano nacional.

3.2 – Índices de Concentração da População Urbana Residente em Centros Urbanos de 10 000 e mais Habitantes

Para medir a concentração da população de um território, em determinado momento, é necessária a aplicação de indicadores operacionais para estimar o grau de desigualdade relativa entre a população e o território.

Como forma de medir essa desigualdade relativa, utilizam-se, neste trabalho, os índices de similaridade e de Gini, para dados agrupados. Ambos medem diferenças relativas entre duas variáveis (que no caso é o no. de centros urbanos)[11] refletindo as desigualdades entre elas.

À medida que o índice de Gini se afasta de zero (0), pode-se dizer que o ganho relativo em população (x) é muito mais rápido que o avanço relativo do número de centros urbanos (y). Por outro lado, convém ter presente que, se a população estivesse igualmente repartida em todos os centros urbanos (o que, equivale dizer, se fossem de igual tamanho), as variáveis população (x) e no. de centros urbanos (y) variariam em igual proporção e, portanto, não existiria nenhuma concentração da população.

O índice de Gini para dados agrupados permite visualizar como é variável a tendência à concentração da população em distintos momentos para uma unidade espacial (país, região), ou, em um mesmo momento, para distintas unidades espaciais. A interpretação desse índice é que, se os valores de ambas variáveis (y, x) aumentam uniformemente, o índice vale zero (0), o que significa que a população se distribui uniformemente em todos os centros urbanos e, portanto, não existe concentração alguma. Por outro lado, se a população cresce mais rapidamente que o número de centros urbanos, o índice é maior ou igual a zero (0) e o seu limite máximo é 1.

Para finalizar, o índice de Gini reflete em que medida a população dos centros urbanos difere do caso hipotético, onde todas as localidades teriam o mesmo tamanho de população. À medida que aumenta o índice, maior é a concentração da população nos grandes centros urbanos.

A Tabela 6 fornece uma idéia da concentração da população brasileira residente em centros urbanos de 10 000 e mais habitantes nas duas últimas décadas.

Esses índices mostram o alto grau de concentração da população brasileira residente em centros urbanos de 10 000 e mais habitantes, assim como altos índices de similaridade entre o número de centros urbanos e a população residente entre 1960 e 1980.

Para se ter uma idéia, já em 1960, quando ainda mais da metade da população brasileira residia na área rural, a população residente em centros urbanos com mais de 10 000 habitantes representava 1/3 da população total e 73% da população urbana do país, sendo que, desta, 58% concentravam-se em cidades com mais de 100 000 habitantes.

Em 1970, acentuou-se ainda mais essa concentração populacional, principalmente nos cen-

tros urbanos de 10 000 e mais habitantes. A percentagem da população brasileira residente nesses centros urbanos passou de 32,6% em 1960 para 40,6% em 1970. Em 1980, essa proporção alcançou 50,2% da população brasileira e 74% da população urbana, sendo que, desta, 64% vivia em cidades com mais de 100 000 habitantes. Nota-se, contudo, que as medidas derivadas da aplicação dos dois índices revelam um ligeiro decréscimo na concentração das pessoas residentes em cidades de 10 000 e mais habitantes entre 1970 e 1980.

Resultados interessantes podem ser observados na Tabela 6 quando se analisam os mesmos índices por zonas de ocupação.

A desagregação desses índices revela a mesma tendência observada para o conjunto da população brasileira, refletindo o fato de que, independente da zona de ocupação, é alto o grau de concentração da população urbana residente em centros urbanos de 10 000 e mais habitantes. Nas zonas de ocupação, sabe-se que é a zona de ocupação total aquela que desde 1960 vem concentrando a maior percentagem da população brasileira (58,2% em 1960 e 57,8% em 1980). Entretanto, nota-se que, apesar da zona de ocupação total ter apresentado índices de Gini e de similaridade superiores à média dos anos considerados, não foi a que apresentou as maiores variações nos graus relativos de concentração da população, haja vista que as zonas de ocupação parcial e vazia registraram, durante o período, incrementos relativos de concentração mais elevados.

Em síntese, as alterações progressivas nos índices de Gini e de similaridade indicam que, por um lado, o número de centros urbanos de 10 000 e mais habitantes não acompanha a crescente concentração da população ou, dito de outra forma, é inferior ao número crescente de população em cada zona de ocupação, o que explica o padrão de concentração populacional. Por outro lado, a concentração crescente da população em centros urbanos ocorre independentemente do tipo de zona de ocupação e inclusive mostra-se, nas duas últimas décadas, mais considerável nas zonas de ocupação parcial e vazia. Tal processo contribui para caracterizar a forma específica de crescimento de urbanização no Brasil que é crescente e concentradora.

Para melhor situar os padrões de concentração da população residente em cidades de 10 000 habitantes, considerou-se que uma forma complementar de caracterização consiste na mensuração dos graus de dispersão dessa população, na medida em que através da dispersão pode-se, de uma outra maneira, visualizar o seu oposto, ou seja, a concentração.

3.3 – Dispersão da População Residente em Centros Urbanos de 10 000 e mais habitantes[12]

Para medir a dispersão da população urbana, tomaram-se como critério de aplicação os municípios de cada Unidade da Federação cuja população urbana da sede municipal somava

10 000 e mais habitantes. O restante dos municípios corresponde àqueles que têm a população dispersa. Consideram-se Unidades da Federação com população dispersa aquelas em que a maior parte da população urbana está residindo em municípios com população de menos de 10 000 habitantes. Em caso contrário, ou seja, aquelas em que a maior parte da população reside em cidades com 10 000 e mais habitantes, o que se verifica é uma baixa dispersão.

Para qualificar as Unidades da Federação segundo o grau de dispersão, adotaram-se as seguintes escalas de variação a partir das porcentagens da população urbana localizada nas sedes dos municípios de 10 000 e mais habitantes:

de zero a 49 por cento - alta
de 50 a 74 por cento - média
de 75 a 100 por cento - baixa

Este critério foi seguido, segundo as zonas de ocupação, para os anos de 1960, 1970 e 1980.

Como havíamos dito anteriormente, a população urbana brasileira tende a se concentrar, cada vez mais, nos grandes centros urbanos, fazendo com que diminua o grau de dispersão como um todo e que cada vez mais passe a existir um maior número de centros urbanos com 10 000 e mais habitantes.

Essa situação é facilmente constatada através das Tabelas 7 e 8, com ênfase na última, a qual contém a subdivisão das zonas de ocupação e Unidades da Federação, segundo o grau de dispersão. Assim, enquanto em 1960, de acordo com o critério de qualificação adotado, 6 Unidades da Federação encontravam-se em baixo grau de dispersão (mais 75% da população urbana residindo em cidades de 10 000 e mais habitantes), em 1980 o número de UFs nessa situação evoluiu para 9. I interessante notar que esse crescimento ocorreu quase que exclusivamente nas UFs da zona de ocupação vazia, o que chama atenção para o fato de que o processo de urbanização nessas áreas vem ocorrendo com um grau de concentração relativamente mais acelerado do que nas demais áreas, onde a ocupação demográfica, historicamente, havia resultado em níveis mais elevados de densidade demográfica.

Com efeito, as características de urbanização nas áreas de ocupação vazia, que compreendem basicamente UFs das Regiões Norte e Centro-Oeste e onde se observa um baixo grau de dispersão em 1980, parecem refletir os efeitos atuais dos movimentos populacionais nessas áreas que são comumente chamadas áreas de fronteira agrícola. A dinâmica dessa ocupação tem sido caracterizada pelo "fechamento" das terras por parte dos grandes grupos econômicos e pela expulsão e redirecionamento de fluxos migratórios anteriormente atraídos para essas áreas. Esse redirecionamento tem como destino os principais centros urbanos regionais e seus bairros periféricos. Desse modo, a baixa dispersão e seu complemento, a concentração urbana, que estão visivelmente ocorrendo nas zonas de ocupação vazia, não significam necessariamente um processo de desenvolvimento econômico ou social via urbanização, mas, sim, a periferização e proletarização urbana de centenas de milhares de pessoas sem acesso às terras "livres" ou

prometidas nas áreas de fronteira agrícola.

Cabe assinalar que nas demais zonas de ocupação prevalecem os níveis de dispersão médios, ou seja, o grau relativo de população residindo em cidades de 10 000 e mais habitantes, durante o período de 1960-1980, mantém-se dentro da faixa intermediária, revelando um padrão mais ou menos estável que não parece estar afetado decisivamente por características de cada zona e ocupação demográfica. O nível médio de dispersão é, inclusive, coerente com o conjunto do Brasil, sendo que a tendência, embora gradual, parece ser a passagem de algumas unidades do nível médio para o nível de baixa dispersão. Tal fato corresponde ao processo geral de urbanização no país, o qual, como vimos, até certo ponto é independente das zonas de ocupação, na medida em que a tendência à concentração em diversos núcleos regionais urbanos é mais ou menos generalizada para o conjunto do país.

Resumindo, o que se pode inferir é que há uma tendência genérica da população brasileira de situar-se em três níveis de dispersão:

1° - No primeiro nível, a baixa dispersão é mantida pelos grandes centros urbanos, que cada vez mais contribuem para que haja uma menor dispersão da população urbana;

2° - No segundo nível, pela tendência em aumentar o número de Unidades da Federação que apresentam um grau de dispersão médio, ou seja, com uma taxa de população urbana entre 50 e 75% residindo nas localidades com 100 000 e mais habitantes, aumento esse que ocorre em áreas onde anteriormente a dispersão era alta; e

3° - No terceiro nível, prevalece um alto grau de dispersão da população urbana em um número muito pequeno de UFs (Maranhão e Piauí, em 1960, Ceará e Distrito Federal, em 1980).

Esses três níveis de dispersão podem ser, também, visualizados através das zonas de ocupação, mas de tal modo que praticamente em todas essas zonas prevalecem os níveis de dispersão baixa e média, ou seja, os graus de dispersão da população urbana residindo em cidades de 10 000 e mais habitantes não estavam preponderantemente afetados pelas grandes zonas de ocupação demográfica do Território Nacional (*vide* Tabela 8).

4 – TENDÊNCIAS E DIFERENÇAS QUE APRESENTAM AS CARACTERÍSTICAS SÓCIO-ECONÔMICAS DA POPULAÇÃO BRASILEIRA, SEGUNDO AS ZONAS DE OCUPAÇÃO ENTRE 1960-1980

Sabe-se que os desequilíbrios no processo de povoamento de um território estão intimamente ligados à dinâmica do desenvolvimento sócio-econômico, tendo em vista que os desníveis porventura existentes na estrutura sócio-econômica, em um determinado momento, afetam direta ou indiretamente as condições de dinâmica populacional e de sua ocupação territorial.

Em outras palavras, as disparidades sócio-econômicas existentes em um momento no espaço são frutos do processo histórico e social de desenvolvimento que, em última instância, condiciona uma série de características sócio-econômicas e demográficas de um determinado território.

Assim, como uma forma de descrever essas características, procura-se utilizar indicadores sócio-econômicos globais para o conjunto da população brasileira e para as zonas de ocupação que possam refletir e explicar, de certo modo, as disparidades encontradas anteriormente no processo de ocupação e concentração de população brasileira nas duas últimas décadas.

Tais indicadores seriam basicamente os seguintes:

- População Economicamente Ativa (PEA) por setores de atividades;
- Índice de Alfabetização da População de 5 anos e mais;
- Grau de Urbanização; e
- Índice de Superfície Aproveitada com agropecuária.

De modo geral, a análise desses indicadores mostra diferenças bastante acentuadas no transcorrer do período considerado (1960-1980)[13], tanto nas zonas de ocupação como no interior destas. Pode-se observar, também, que o alto grau de heterogeneidade demográfica do espaço brasileiro nas duas últimas décadas vem sendo acompanhado de grandes disparidades sócio-econômicas. Essa situação tende a manter-se e a diferenciar-se entre si, pois vem se acentuando desde 1960, tanto no Brasil como no interior das zonas de ocupação.

O aumento e a diversificação da PEA urbana (PEA no Secundário e no Terciário), do grau de urbanização e da superfície aproveitada com agropecuária explicam e refletem as mudanças ocorridas na estrutura econômica brasileira nos últimos 20 anos. Ainda mais, mostram, por um lado, as características e as modalidades do desenvolvimento nacional e, por outro, funcionam como indicadores explicativos do processo de distribuição e concentração espacial da

população brasileira. No conjunto, observou-se que a dinâmica da economia urbana, setorial e espacialmente diversificada contribuiu direta e indiretamente para uma maior diferenciação regional (para ocupação) nos índices de alfabetização, participação da PEA no Secundário e Terciário, assim como realimentou as condições de crescimento dos centros urbanos. A síntese geral desses processos contribuiu, de modo significativo, não só para a elevação do grau relativo de heterogeneidade demográfica do Território Nacional e, em especial, no que tange às áreas urbanas, como, também, reflete a concentração e as desigualdades sócio-econômicas do Território Nacional.

Essa situação pode ser visualizada através das zonas de ocupação, principalmente no que se refere às mudanças ocorridas na zona de ocupação total e em relação às demais. Como podemos observar na Tabela 11 e seguintes (Anexo 1), essa zona vem apresentando, desde 1960, os mais altos índices de concentração e mobilidade da força de trabalho, principalmente em São Paulo e Rio de Janeiro. Estas UFs possuíam, em 1960, 30,0% da população economicamente ativa do Brasil e 49,8% da PEA da referida zona de ocupação. Em 1980, passaram a deter 34,1% da força de trabalho[14] do país. A participação da PEA de ambas as UFs na respectiva zona de ocupação foi, então, da ordem de 57,1%. Esses valores por si só são consideravelmente elevados e bastante ilustrativos para revelar o processo de concentração da economia e da população brasileira, cujos reflexos podem ser, também, observados através do aumento das desigualdades sócio-econômicas regionais e da elevação do grau de heterogeneidade urbana do país. A análise dessas diferenças nos permite avaliar, de certo modo, as características da dinâmica entre as zonas de ocupação.

Para o conjunto das demais zonas de ocupação (parcial, subocupada e vazia) destaca-se a zona subocupada, onde se processou a maior dinamização na estrutura econômica e social, haja vista que foi a que apresentou maiores índices de crescimento da força de trabalho, em todos os setores de atividade, durante os anos 70. Em contrapartida, os índices da superfície aproveitada com agropecuária passaram de 19,7% a 26,8% entre 1960 e 1970, o que representou um incremento relativo da ordem de 185,4%, ou seja, 650 mil km^2 a mais em relação a 1960. Com base em informações contidas na Tabela 14, observa-se que, no período 70-80, diminuiu, consideravelmente, o ritmo de crescimento da superfície utilizada com agropecuária no país e, em especial, nas zonas de ocupação total e vazia. Acrescenta-se a esse aspecto o aumento do peso da superfície utilizada com pecuária que passou de 69,6%, em 1970, para 75,3%, em 1980. Este dado reflete, por si só, a perda de importância da agricultura na década de 70. Não é por acaso que se deu o esvaziamento do campo das zonas de ocupação total e vazia. Na primeira, mais de 1,5 milhões economicamente ativos ligados à atividade do Setor Primário deixaram o campo, especialmente no Centro-Sul do país (no Paraná, São Paulo e Rio Grande do Sul). Esse abandono representou uma perda da ordem de 20,9% da PEA no Primário em relação a 1970. A zona de ocupação vazia sofreu uma evasão de mais de 130 mil ativos das atividades

primárias, montante que representou uma perda de 26,9% da PEA.

Assim, os níveis reduzidos de crescimento da superfície aproveitada com agropecuária durante a década de 70, associados ao decréscimo relativo da participação da PEA no Setor Primário, refletem as mudanças que vão ocorrendo na organização demográfica e econômica do espaço rural, onde a penetração de novas relações de produção, de tipo capitalista, concorre para que as transformações no tipo e na escala da produção agrícola possibilite, concomitantemente, uma acelerada liberação de mão-de-obra. Essa população, posta em movimento, vai redimensionar os setores de atividades urbanas e, sobretudo, intensificar o grau de heterogeneidade relativa do espaço, com o aumento da concentração, principalmente nos centros urbanos de 10 000 e mais habitantes.

Entretanto, é importante ter presente que, embora as zonas de ocupação mostrem comportamentos diferenciados na ocupação de seus espaços, não podemos desvinculá-los do todo, pois sabe-se que maior ou menor incremento relativo das variáveis sócio-econômicas representam não só os diferentes " estímulos" no processo de desenvolvimento nacional, a partir do Centro-Sul do país, e em especial de São Paulo e Rio de Janeiro, como também das respectivas modalidades em cada zona de ocupação.

Uma forma de constatar isso é ver como se dá a locação da PEA por setores de atividade, por zonas de ocupação entre as duas últimas décadas. Na década de 60, podemos observar que quase a totalidade do incremento da PEA ocupada foi alocada na zona de ocupação total, ficando o restante desse incremento para a zona subocupada, já que a zona vazia apresentou, no mesmo período, um incremento negativo na taxa de atividade (Tabela 9).

Diante desse panorama, dois aspectos são importantes em relação aos setores de atividade, começando pelo Setor Primário, através do aumento dos índices de aproveitamento da superfície com agropecuária, e em seguida, tratando das taxas de atividade no Setor Secundário e Terciário. Em relação ao primeiro aspecto, essas mudanças estão intimamente vinculadas à "dinamização" da penetração de capital na agropecuária, da qual resultou um aumento da subocupação demográfica do espaço rural em função da expulsão da mão-de-obra tradicionalmente ocupada no campo. A título de exemplo, pode-se observar que a zona vazia expulsou entre 1960 e 1970 mais de 200 mil indivíduos que estavam vinculados às atividades primárias, dos quais apenas uma pequena parcela supostamente absorvida pelo Setor Secundário da zona considerada. Em relação ao segundo aspecto, a dinamização da economia fez-se sentir mais a nível regional, principalmente na zona de ocupação total, e em especial no Centro-sul do país. Essa situação contribuiu ainda mais para o aumento da heterogeneidade espacial e sócio-econômica do país, que tornou-se tipicamente urbana a partir da década de 60. Nesse sentido, o incremento nos setores de atividades urbanas (Secundário e Terciário) está naturalmente vinculado às profundas alterações que vão sendo registradas durante a década na distribuição espacial e na organização econômica regional do conjunto da sociedade brasileira. Convém não

esquecer que esse processo já era parcialmente visualizado através das tendências históricas que começaram a se impor no Brasil após 1930.

Na década de 70, essa situação continuou a acentuar-se, observando-se, contudo, que a redistribuição espacial das atividades econômicas ocasionou uma maior mobilidade absoluta e relativa da força de trabalho, não só nas regiões como também entre elas. Entretanto, a maior alocação dessa força de trabalho continuou dando-se em função da zona de ocupação total (conforme podemos observar na Tabela 10), embora a zona subocupada tenha igualmente absorvido uma parte expressiva do crescimento relativo do conjunto da população ativa.

É importante notar-se que, para as demais zonas de ocupação, processaram-se ilustrativas mudanças na elaboração da força de trabalho na última década. A zona subocupada, por exemplo, que tinha absorvido somente 9,6% da PEA ocupada entre 1960-1970, passou a concentrar mais de 50,0% do incremento geral da PEA entre 1970-1980. Outro dado importante encontra-se nas mudanças ocorridas na zona vazia que na década de 60 expulsou mais de 200 mil pessoas economicamente ativas, mais que na década passada; reteve por volta de 70 mil ativos que se repartiram principalmente entre o Secundário e Terciário. Em contrapartida, as atividades do Setor Primário nas zonas de ocupação total e vazia expulsaram 1,7 milhão de economicamente ativos, o que representou mais de 12,0% do crescimento da força de trabalho na década de 70.

Em síntese, pode-se concluir que a dinâmica do desenvolvimento brasileiro está intimamente ligada às formas desiguais de ocupação regional, desigualdades estas que induzem a uma concentração cada vez maior dos recursos e da população em detrimento de uma distribuição "eqüitativa" entre população e território, ou seja, de um "aproveitamento equilibrado" do espaço nacional. De acordo com as modalidades atuais de desenvolvimento, a tendência predominante é de concentrar-se cada vez mais a população nos grandes centros urbanos, contribuindo para que não só aumentem as divergências regionais como também intra-regionais, aumentando, assim, a sobreocupação demográfica dos respectivos espaços urbanos e conseqüentemente a subocupação dos espaços rurais. Por outro lado, incremento nas taxas de industrialização e de terciarização da economia a nível regional e os decréscimos da PEA ocupada no Setor Primário, principalmente na zona vazia, nos mostram o fechamento da fronteira agrícola e nos dão a impressão de que o país permanecerá ainda, por algum tempo, em condições de subocupação demográfica na maior parte de seu território.

5 – CONCLUSÕES

As condições de ocupação vistas através do grau relativo de *heterogeneidade demográfica*

do espaço, seja este total, urbano ou rural, não devem ser interpretadas como uma simples razão entre quantidades (área e população, cuja dimensão se traduz no grau de concentração e dispersão da população) em determinado momento. Na realidade, devem ser vistas como uma relação variável ao longo do tempo, cuja dimensão quantitativa está determinada pelo processo histórico de ocupação do espaço. Este por sua vez está subordinado à lógica interna da reprodução do capital que em sua dimensão territorial passa a ser função de um processo de concentração e centralização dos desenvolvimentos produtivo e técnico, o que contribui para que o espaço apresente-se cada vez mais heterogêneo, tanto do ponto de vista econômico quanto demográfico, de tal maneira que as conseqüências manifestam-se através de um processo crescente de concentração populacional nos grandes centros urbanos. Nesse sentido, pode-se dizer que o aumento e a variação no grau relativo de heterogeneidade de uma determinada área estão intimamente vinculados à mobilidade espacial da força de trabalho, seja esta proveniente da área rural e/ou outras áreas urbanas menores para os grandes centros metropolitanos. Tal mobilidade espacial da força de trabalho espelha a nível populacional os padrões da dinâmica do desenvolvimento capitalista nacional, sendo seus movimentos perfeitamente visíveis nos períodos históricos de formação e consolidação do capitalismo no Brasil e com especial ênfase nas duas últimas décadas.

Não é por acaso que, em 1970, 25,6% da população nacional estava residindo nas regiões metropolitanas, razão essa que subiu para 29,0% em 1980. Este aumento relativo representou mais de 40% do crescimento da população urbana entre 1970-1980. Além do mais, foi responsável por um aumento no grau de heterogeneidade urbana brasileira de 16,5% entre 1970 e 1980. Se excluirmos a população urbana metropolitana, verifica-se que o aumento do grau de heterogeneidade urbana foi de apenas 2,4% nesse período. Esses resultados mostram não só o peso do processo de metropolização da população urbana do país, como também traduzem o aumento progressivo da sobreocupação demográfica de áreas urbanas que possivelmente já se apresentavam saturadas demograficamente. Contudo, não podemos deixar de levar em consideração a capacidade de absorção da mão-de-obra em setores produtivos da economia dessas áreas, visto que a existência inequívoca de uma tendência de expansão e diversificação do Setor Terciário que passa a "absorver" uma massa de trabalhadores potencialmente em condições de subemprego ou de desemprego. Deste modo, a expansão demográfica das regiões metropolitanas reflete, em grande parte, a aglutinação de segmentos populacionais provenientes de outros centros urbanos e/ou de áreas rurais que vão preencher os requisitos da demanda por força de trabalho na dinâmica dos mercados metropolitanos.

Esse processo de metropolização foi bastante intensificado a partir da década de 50, visto ter sido observado que, em 1960, quando mais da metade da população brasileira ainda residia em áreas rurais, a população urbana residente em centros urbanos de 10 000 e mais habitantes representava pouco mais de 30% da população total e 73% da população urbana do país, sendo

que, desta, 58% residia em cidades de 100 000 habitantes e mais. Em 1980, essa porcentagem alcançou 50,2% da população total e 74% da população urbana brasileira, e a residente em cidades de 100 000 e mais habitantes passou a representar 64% da população urbana nacional.

Essa mesma tendência é observada nas zonas de ocupação, pois as alterações progressivas no grau de concentração, vistas através dos índices de Gini e similaridade, nos mostram que os centros urbanos tornaram-se mais iguais entre si, não pela homogeneização entre eles, mas pelas crescentes desigualdades. Isto significa que o número de cidades não vem acompanhando o ritmo crescente de concentração populacional e a igualdade se fez mais pelo elevado grau de concentração populacional do que em função das respectivas zonas de ocupação. Entretanto, é importante destacar que o processo de concentração populacional se deu com maior intensidade em cidades das zonas de ocupação vazias e parcial, regiões onde houve maior expansão da superfície aproveitada com agropecuária. Esta situação levou a um maior aumento relativo no grau de urbanização de zona vazia do que nas demais zonas de ocupação em 1980.

De modo geral, esse quadro reflete a diminuição no grau de dispersão da população brasileira, o qual se faz sentir com maior peso nas UFs da zona de ocupação vazia, onde o processo de urbanização, via terciarização, vem ocorrendo em um grau relativo de concentração mais elevado, em relação ao restante do país, como assinalamos anteriormente.

No caso específico dessa zona de ocupação, o processo atual é condicionado pelos movimentos populacionais recentes para essa região, seja através da chamada expansão da fronteira agrícola, seja pelo subseqüente "fechamento" desta, isto é, pela apropriação capitalista para terras "disponíveis" para serem ocupadas. Estes aspectos redimensionaram os fluxos migratórios, anteriormente atraídos pela busca de vastas áreas vazias e possíveis de serem ocupadas, em direção aos centros urbanos regionais e, desses, para bairros periféricos e/ou áreas empobrecidas do espaço urbano. Esta trajetória pode ser exemplificada quando se assinala o crescimento de Manaus, centro amazônico e cidade que mais concentrou, em termos relativos, na última década. Ainda uma vez, esses aspectos nos levam a concluir que a baixa dispersão, assim como a baixa ocupação relativa da zona de ocupação vazia, não significam exatamente a contrapartida de um processo de desenvolvimento econômico e social, via urbanização, mas que pode ser o resultado de um processo distorcido que se revela através da periferização e proletarização de centenas de milhares de pessoas sem acesso às terras "prometidas" na fronteira agrícola.

Outro aspecto que contribuiu sensivelmente para a concentração da população em determinados centros urbanos foi a dinamização da economia urbana, vista através da diversificação da PEA urbana, que, conjugada com o grau de urbanização e o aumento da superfície aproveitada com agropecuária, reflete direta e indiretamente as mudanças ocorridas na estrutura econômica brasileira nas respectivas zonas de ocupação. Esses dados nos mostram, ainda que de modo genérico, a expansão das relações capitalistas nas UFs da zona de ocupação total, localizada

no Centro-Sul do país (São Paulo e Rio de Janeiro, em particular), que permitiu, de certo modo, que a PEA se apresentasse setorial e especialmente diversificada, refletindo não só uma maior diferenciação entre os setores de atividade como também entre as respectivas zonas de ocupação. Esses aspectos sintetizam as desigualdades sociais e econômicas, tanto para o Brasil como, e principalmente, inter e intra respectivas zonas de ocupação demográfica.

Em síntese, todos esses resultados em conjunto nos levam a concluir que o aumento e o elevado grau de heterogeneidade demográfica do Território Nacional estão mais intimamente vinculados à estrutura econômica e social do que à estrutura geográfica de cada zona de ocupação. Assim, há uma tendência da população a concentrar-se cada vez mais em grandes centros que cumprem as exigências do processo de desenvolvimento capitalista brasileiro. Com efeito, são esses espaços relativamente pequenos territorialmente que vão condicionar, em grande parte, os padrões de heterogeneidade demográfica do Território Nacional. Desse modo, além da variedade de superfície, essas UFs apresentam maiores variações em suas estruturas espaciais, situação que supostamente está vinculada ao próprio padrão de heterogeneidade demográfica do Território Nacional.

Esses aspectos gerais, associados aos indicadores sócio-econômicos, nos permitem inferir que o tamanho da superfície se correlaciona mais com a estrutura de distribuição espacial global da população do país, do que com as respectivas zonas de ocupação. Essas apresentam distintas modalidades de distribuição espacial da população e são o resultado de uma dinâmica muito mais ampla, cujo processo de desenvolvimento capitalista experimentado no período em questão contribuiu não só para aumentar o grau de heterogeneidade demográfica dos respectivos espaços territoriais (total, urbano e rural), como também para intensificar as desigualdades econômicas e sociais inter e intra cada zona de ocupação demográfica.

ANEXO 1

1 – Indicadores Sócio-econômicos Básicos

Esses indicadores correspondem aos Censos Demográficos de 1960, 1970 e 1980. Os indicadores selecionados foram:

- taxa de alfabetização da população de 5 anos e mais;
- porcentagem da População Economicamente Ativa (PEA) de 10 anos e mais na população total;
- taxa refinada de atividade;

- porcentagem da PEA no Setor Primário;
- porcentagem da PEA no Setor Secundário;
- porcentagem da PEA no Setor Terciário;
- taxa de crescimento médio geométrico anual da PEA (1960-1970); e
- taxa de terciarização (taxa hiperbólica de crescimento 1960-1970).

2 – Indicadores de Urbanização

Os indicadores de urbanização foram os seguintes:

- grau de urbanização (porcentagem da população urbana na população total de cada zona);
- índice de Gini para os centros urbanos[15] de 10 000 e mais habitantes (1960, 1970, 1980); e
- índice de similaridade para os centros urbanos[16] de 10 000 e mais habitantes (1960, 1970, 1980).

3 – Conceitos Utilizados

Nos últimos três Censos Demográficos o conceito de população urbana abrange a população recenseada nas cidades e vilas, sendo que em 1980 foram consideradas áreas urbanas isoladas delimitadas pelas respectivas posições municipais vigentes em 31 de agosto de 1980. **O conceito de cidade** englobava a população urbana residente na sede municipal (distrito-sede) mais as vilas, ou seja, população urbana residente na sede dos demais distritos do município.

Como podemos notar, o critério de população urbana está intimamente ligado ao aspecto administrativo municipal, visto ser de competência das Prefeituras estabelecer os limites urbanos. **O conceito de cidade** utilizado neste trabalho parte da população urbana residente na sede municipal dos municípios de 10 000 e mais habitantes, pois supõe-se que, a partir desse número, já existe uma organizaçao sócio-econômica e cultural que configura um padrão de organização urbana. Entretanto, ressaltamos que o objetivo deste trabalho não é a análise do papel das cidades, mas sim o estudo do crescimento e concentração demográfica, vinculado, entre outros, a padrões de crescimento urbano:

- O crescimento de **população rural** foi o da população recenseada fora dos limites de cidades e vilas; e

• **Pessoas alfabetizadas** - Os Censos Demográficos de 1960, 1970 e 1980 consideram alfabetizadas as pessoas de 5 anos e mais capazes de ler e escrever um bilhete simples, em um idioma qualquer. Os que assinavam apenas o próprio nome foram considerados analfabetos.

4 – População Economicamente Ativa (PEA)

Foram consideradas nos últimos censos as pessoas de 10 anos e mais que declararam estar exercendo uma atividade econômica em maior parte do ano anterior ou procurando trabalho à data do censo. Em 1960, não foi considerado o grupo procurando trabalho pela primeira vez.

Essas pessoas foram agrupadas em setores de atividade de acordo com as principais características dos ramos de atividade exercida por essa população. Os critérios de agrupamento são mais ou menos semelhantes entre 1960 e 1980:

• **Setor Primário**: inclui os seguintes ramos de atividade: agricultura, pecuária, silvicultura, atividades extrativas (extração vegetal, caça e pesca e extração mineral).

• **Setor Secundário**:atividades industriais (indústria de transformação, indústria da construção e serviços industriais de utilidade pública).

• **Setor Terciário**: Comércio de mercadorias, Prestação de Serviços (incluindo pedreiros, carpinteiros, etc., trabalhando por conta própria); Transportes, comunicações e armazenagem; Atividades Sociais (Ensino, Assistência médico-hospitalar, Previdência Social, etc.); Administração Pública (serviços administrativos governamentais), Legislativo, Justiça, Defesa Nacional e Segurança Pública; e Outras Atividades (comércio de imóveis e valores imobiliários, instituições de crédito, seguros e capitalização e profissões liberais) que incluem, também, as pessoas que procuravam trabalho pela primeira vez.

TABELA 1

DENSIDADE MÉDIA, TOTAL, URBANA E RURAL E COEFICIENTE URBANO-RURAL DE HETEROGENEIDADE - BRASIL - 1960, 1970 E 1980

ANOS	DENSIDADE MÉDIA (1)			COEFICIENTE URBANO-RURAL DE HETEROGENEIDADE
	Total	Urbana (A)	Rural (B)	(A/B)
1960	22,77	11,56	11,21	1,03
1970	32,41	20,95	11,46	1,83
1980	42,49	31,95	10,54	3,03

FONTE - IBGE. Sinopse dos Censos Demográficos de 1960, 1970 e 1980.

(1) A densidade média refere-se à média aritmética simples das densidades estaduais.

TABELA 2

INDICADORES DE DISTRIBUIÇÃO ESPACIAL, SEGUNDO ZONAS DE OCUPAÇÃO 1960-1980

(continua)

INDICADORES E ANO	ZONAS DE OCUPAÇÃO				
	Total	Parcial		Subocupada	Vazia
		Urbana	Rural		
1960					
1 - Porcentagem da população total.	58,16	0,20	20,24	16,47	4,93
2 - Porcentagem da população urbana	68,46	0,28	16,47	10,64	4,15
3 - Porcentagem da população rural.	49,85	0,13	23,28	21,19	5,55
4 - Grau relativo de concentração urbana (1)	1,1967	0,09	0,08	0,04	0,02
5 - Grau relativo de concentração rural (2)	1,43	0,04	0,10	0,08	0,01
6 - Grau de urbanização (3)	48,65	63,02	69,02	45,40	37,70
7 - Porcentagem do território nacional (5)	13,05	9,31		21,03	56,59
8 - Grau relativo de ruralização(4)	51,35	36,98	30,58	54,60	62,30

TABELA 2

INDICADORES DE DISTRIBUIÇÃO ESPACIAL, SEGUNDO ZONAS DE OCUPAÇÃO
1960-1980

(conclusão)

INDICADORES E ANO	ZONAS DE OCUPAÇÃO				
	Total	Parcial		Subocupada	Vazia
		Urbana	Rural		
1970					
1 - Porcentagem da população total.	65,88	-	-	31,10	3,02
2 - Porcentagem da população urbana	73,16	-	-	24,48	2,36
3 - Porcentagem da população urbana (excluindo as Regiões Metropolitanas)	49,40	-	-	75,03	-
4 - Porcentagem da população rural.	56,60	-	-	39,51	3,89
5 - Grau relativo de concentração urbana (1)	1,4197	-	-	0,0934	0,0063
6 - Grau relativo de concentração rural (2)	0,9647	-	-	0,1292	0,0069
7 - Grau de urbanização (3)	62,11	-	-	44,01	43,62
8 - Grau relativo de ruralização (4)	37,89	-	-	55,99	56,38
9 - Porcentagem do território nacional (5)	15,47	-	-	44,26	40,27
1980					
1 - Porcentagem da população total	57,84	18,75	-	21,04	2,47
2 - Porcentagem da população urbana	63,70	19,12	-	15,12	2,06
3 - Porcentagem da população urbana (excluindo as Regiões Metropolitanas)	49,33	70,01	-	79,18	-
4 - Porcentagem da população rural	45,58	18,04	-	33,36	3,02
5 - Grau relativo de concentração urbana (1)	1,8771	0,7089	-	0,0834	0,0075
6 - Grau relativo de concentração rural (2)	0,9364	0,1005	-	0,1184	0,0067
7 - Grau de urbanização (3)	80,75	68,82	-	48,60	58,93
8 - Grau relativo de ruralização (4)	19,75	31,18	-	51,40	41,07
9 - Porcentagem do território nacional (5)	12,24	10,12	-	44,40	33,24

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.

NOTAS - As porcentagens de população total, urbana e rural, referentes a cada zona de ocupação, foram calculadas com relação às respectivas populações do País.

(1), (2) Os graus relativos de concentração urbana e rural correspondem à soma dos indicadores de ocupação física; (3) Grau de urbanização $U = \frac{Gu}{1+Gu}$ onde $Gu = \frac{Pop.urbana}{Pop.rural}$; (4) Grau de ruralização $R = \frac{1}{1+Gu}$; (5) A porcentagem do território corresponde à fração do território nacional correspondente a cada zona.

TABELA 3

INDICES DE OCUPAÇÃO DEMOGRAFICA DAS AREAS URBANA E RURAL, SEGUNDO ZONAS DE OCUPAÇÃO - 1960-1980

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	INDICES DE OCUPAÇÃO DEMOGRAFICA (%)					
	1960		1970		1980	
	Urbana	Rural	Urbana	Rural	Urbana	Rural
TOTAL	119,67	99,66	141,97	96,47	187,71	93,64
Ceará	4,44	7,14	4,81	6,83	6,31	7,21
Rio Grande do Norte	4,79	6,45	5,53	5,95	6,92	6,31
Paraíba	7,34	11,03	7,06	9,50	8,46	10,00
Pernambuco	11,01	11,02	11,36	9,28	12,67	10,25
Alagoas	9,01	11,11	9,08	13,42	11,64	15,56
Sergipe	7,82	10,02	7,50	8,57	9,25	10,15
Espírito Santo	4,93	9,62	6,29	7,47	9,33	6,83
Rio de Janeiro (1) (2) ..	41,63	11,50	34,86	8,78	78,80	9,09
São Paulo	19,23	9,24	22,94	5,49	29,52	4,90
Santa Catarina	4,19	7,21	5,18	6,73	7,42	6,58
Rio Grande do Sul (e) ...	5,28	5,32	5,28	4,52	6,46	4,03
PARCIAL	16,88	12,53	-	-	76,49	10,05
Minas Gerais (c)	3,91	4,96	4,13	3,62	5,08	3,22
Paraná (a, d)	3,89	7,10	5,00	8,63	7,39	6,76
Distrito Federal (b, e) .	9,08	0,47	17,08	1,30	64,95	2,80
SUBOCUPADA	4,16	9,11	9,34	12,92	8,34	12,44
Maranhão	0,80	2,99	0,92	2,68	1,27	3,61
Piauí	0,68	1,82	0,85	1,77	1,18	2,11
Bahia	2,15	3,31	2,19	3,06	2,74	3,66
Goiás	0,53	0,99	0,77	1,03	1,23	0,97
VAZIA	0,93	1,32	0,63	0,69	0,75	0,67
Rondônia (f)	0,07	0,08	0,10	0,08	0,32	0,46
Acre (c, f)	0,12	0,39	0,15	0,40	0,29	0,47
Amazonas	0,09	0,15	0,10	0,14	0,18	0,16
Roraima	0,03	0,03	0,03	0,04	0,07	0,06
Pará (c, f)	0,30	0,35	0,33	0,36	0,45	0,61
Amapá	0,15	0,11	0,18	0,14	0,25	0,22
Mato Grosso	0,17	0,21	0,22	0,29	0,25	0,23
Mato Grosso do Sul (f) ..	-	-	-	-	0,86	0,55

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.

NOTAS - (a), (b) Passaram em 1970 para as zonas de ocupação total; (c) passaram em 1970 para as zonas de ocupação subocupada; (d) passaram em 1980 para as zonas de ocupação total; (e) passaram em 1980 para as zonas de ocupação parcial; (f) passaram em 1980 para as zonas de ocupação subocupada.

(1) Em 1960 o Distrito Federal era o atual Município do Rio de Janeiro. (2) Inclui em 1970 e 1980 o antigo Estado da Guanabara e o Estado do Rio de Janeiro.

TABELA 4

INDICES DE HETEROGENEIDADE DEMOGRAFICA DAS AREAS URBANA E RURAL, SEGUNDO ZONAS DE OCUPAÇÃO - 1960-1980

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	INDICES DA HETEROGENEIDADE DEMOGRAFICA					
	1960		1970		1980	
	Urbana I n u	Rural I n r	Urbana I n u	Rural I n r	Urbana I n u	Rural I n r
TOTAL	29,9175	24,9150	35,4925	23,6675	48,8130	24,3464
Ceará	1,1100	1,7850	1,2025	1,7075	1,6406	1,8746
Rio Grande do Norte	1,1975	1,6125	1,3825	1,4875	1,7300	1,6406
Paraíba	1,8350	2,7575	1,7650	2,3750	2,1996	2,6000
Pernambuco	2,7525	2,7550	2,8400	2,3200	3,2942	2,6650
Alagoas	2,2525	2,7775	2,2700	3,3550	3,0264	4,0456
Sergipe	1,9550	2,5050	1,8750	2,1425	2,4050	2,6390
Espírito Santo	1,2325	2,4050	1,5725	1,8675	2,4258	1,7758
Rio de Janeiro (1) (2) ..	10,4075	2,8750	8,7150	2,1950	20,4888	2,3634
São Paulo	4,8075	2,3100	5,7350	1,3725	7,752	1,2740
Santa Catarina	1,0475	1,8025	1,2950	1,6825	1,9292	1,7108
Rio Grande do Sul (e ...	1,3200	1,3300	1,3200	1,1300	1,6796	1,0478
PARCIAL	4,2200	3,1325	-	-	19,8874	2,6130
Minas Gerais (c)	0,9775	1,2400	1,0325	0,9050	1,3208	0,8372
Paraná (a, d)	0,9725	1,7750	1,2500	1,7075	1,9214	1,7576
Distrito Federal (b, e) .	2,2700	0,1175	4,2700	0,3250	16,8870	0,7280
SUBOCUPADA	1,0100	2,2775	2,3350	3,2300	2,1684	3,2344
Maranhão	0,2000	0,7475	0,2300	0,6700	0,3302	0,9386
Piauí	0,1700	0,4550	0,2125	0,4425	0,3068	0,5486
Bahia	0,5075	0,8275	0,5475	0,7650	0,7124	0,9516
Goiás	0,1325	0,2475	0,1875	0,2575	0,3198	0,2522
VAZIA	0,2325	0,3300	0,1700	0,1725	0,1950	0,5486
Rondônia (f)	0,0175	0,0200	0,0250	0,0200	0,0832	0,1196
Acre (c, f)	0,0300	0,0975	0,0375	0,1000	0,0754	0,1222
Amazonas	0,0225	0,0375	0,0375	0,0350	0,0468	0,4160
Roraima	0,0075	0,0075	0,0075	0,0100	0,0182	0,0156
Pará (c, f)	0,0750	0,0875	0,0825	0,0900	0,1170	0,1586
Amapá	0,0375	0,0275	0,0450	0,0350	0,0650	0,0572
Mato Grosso	0,0425	0,0525	0,0550	0,0725	0,0650	0,0598
Mato Grosso do Sul (f) ..	-	-	-	-	0,2236	0,1430

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.

NOTAS - (a), (b) Passaram em 1970 para as zonas de ocupação total; (c) passaram em 1970 para as zonas de ocupação subocupada; (d) passaram em 1980 para as zonas de ocupação total; (e) passaram em 1980 para as zonas de ocupação parcial; (f) passaram em 1980 para as zonas de subocupação.

(1) Em 1960 o Distrito Federal era o atual Município do Rio de Janeiro. (2) Inclui em 1970 e 1980 o antigo Estado da Guanabara e o Estado do Rio de Janeiro.

TABELA 5

TAXAS DE CRESCIMENTO GEOMÉTRICO ANUAL DA POPULAÇÃO RESIDENTE, SEGUNDO AS ZONAS DE OCUPAÇÃO E REGIÕES METROPOLITANAS ENTRE 1970-1980

ZONAS DE OCUPAÇÃO E REGIÕES METROPOLITANAS	TAXA DE CRESCIMENTO MÉDIO GEOMÉTRICO ENTRE 1970-1980	
	Total (%)	Urbano (%)
BRASIL	2,49	4,45
REGIÕES METROPOLITANAS	3,79	4,27
Zona de ocupação total	1,25	3,15
Região Metropolitana	2,84	3,03
Zona de subocupação	5,05	8,01
Região Metropolitana	8,27	8,42

FONTE - IBGE. Censos Demográficos - 1970 e 1980.

TABELA 6

INDICES DE CONCENTRAÇÃO DE GINI (G) E DE SIMILARIDADES ($) DA POPULAÇÃO RESIDENTE EM CENTROS URBANOS DE 10 MIL E MAIS HABITANTES, SEGUNDO ZONAS DE OCUPAÇÃO DEMOGRAFICA - 1960-1980

ZONAS DE OCUPAÇÃO DEMOGRAFICA E INDICES	INDICES DE CONCENTRAÇÃO E DE SIMILARIDADE		
	1960	1970	1980
BRASIL: G	0,6500	0,6691	0,6659
Δ	0,5289	0,5471	0,5355
Zona de ocupação total: G	0,6811	0,6905	0,6930
Δ.....	0,5462	0,5665	0,5993
Zonas subocupadas - Parcial: G	0,5952	0,5892	0,6111
Δ....	0,4797	0,4825	0,5058
- Vazia: G	0,5883	0,5139	0,6485
Δ.....	0,5072	0,4157	0,5247

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.

TABELA 7

TAMANHO DOS CENTROS URBANOS, POR ZONA DE OCUPAÇÃO
1960-1980

(continua)

TAMANHO DOS CENTROS URBANOS	BRASIL					
	1960		1970		1980	
	% de Cidades	% da População	% de Cidades	% da População	% de Cidades	% da População
TOTAL ABSOLUTO	406	22 948 845	576	37 845 720	899	59 770 501
TOTAL RELATIVO	100,00	100,00	100,00	100,00	100,00	100,00
De 10 000 a 19 999 habitantes	52,96	13,06	48,26	10,56	46,38	9,72
20 000 a 49 999 habitantes	29,06	16,07	31,94	14,95	31,59	14,82
50 000 a 99 999 habitantes	10,34	12,87	9,72	10,22	11,57	11,37
100 000 a 499 999 habitantes	6,16	18,48	8,68	25,87	9,01	27,03
500 000 a 999 999 habitantes	0,98	11,68	0,70	7,77	0,78	7,88
1 000 000 e mais habitantes	0,50	27,84	0,69	30,63	0,67	29,18

TAMANHO DOS CENTROS URBANOS	ZONA DE OCUPAÇÃO TOTAL					
	1960		1970		1980	
	% de Cidades	% da População	% de Cidades	% da População	% de Cidades	% da População
TOTAL ABSOLUTO	266	17 284 551	379	28 967 849	486	38 078 371
TOTAL RELATIVO	100,00	100,00	100,00	100,00	100,00	100,00
De 10 000 a 19 999 habitantes	51,50	11,12	44,07	8,42	42,80	7,58
20 000 a 49 999 habitantes	27,82	13,58	33,51	13,38	32,92	12,73
50 000 a 99 999 habitantes	12,41	13,42	10,55	9,68	11,73	10,14
100 000 a 499 999 habitantes	6,77	16,78	10,55	27,53	11,11	28,81
500 000 a 999 999 habitantes	0,75	8,14	0,53	4,79	0,82	6,67
1 000 000 e mais habitantes	0,75	36,96	0,79	36,20	0,62	34,07

TABELA 7

TAMANHO DOS CENTROS URBANOS, POR ZONA DE OCUPAÇÃO
1960-1980

(continua)

TAMANHO DOS CENTROS URBANOS	ZONA DE OCUPAÇÃO PARCIAL					
	1960		1970		1980	
	% de Cidades	% da População	% de Cidades	% da População	% de Cidades	% da População
TOTAL ABSOLUTO	84	3 075 423	-	-	211	10 776 391
TOTAL RELATIVO	100,00	100,00	-	-	100,00	100,00
De 10 000 a 19 999 habitantes	50,00	18,91	-	-	45,97	11,87
20 000 a 49 999 habitantes	39,29	30,12	-	-	30,81	19,05
50 000 a 99 999 habitantes	7,14	14,79	-	-	14,69	18,55
100 000 a 499 999 habitantes	2,38	15,27	-	-	7,58	26,85
500 000 a 999 999 habitantes	1,19	20,91	-	-	-	-
1 000 000 e mais habitantes	-	-	-	-	0,95	23,68

TAMANHO DOS CENTROS URBANOS	ZONA SUBOCUPADA					
	1960		1970		1980	
	% de Cidades	% da População	% de Cidades	% da População	% de Cidades	% da População
TOTAL ABSOLUTO	39	1 742 573	180	8 032 738	185	8 699 817
TOTAL RELATIVO	100,00	100,00	100,00	100,00	100,00	100,00
De 10 000 a 19 999 habitantes	66,67	20,53	57,20	17,92	56,75	16,57
20 000 a 49 999 habitantes	17,95	16,12	28,81	19,95	28,65	18,23
50 000 a 99 999 habitantes	5,13	6,65	7,78	11,57	7,57	10,96
100 000 a 499 999 habitantes	7,69	20,50	4,48	17,44	5,41	20,23
500 000 a 999 999 habitantes	2,56	36,20	1,11	19,33	1,08	16,81
1 000 000 e mais habitantes	-	-	0,56	13,79	0,54	17,20

TABELA 7

TAMANHO DOS CENTROS URBANOS, POR ZONA DE OCUPAÇÃO
1960-1980

(conclusão)

TAMANHO DOS CENTROS URBANOS	ZONAS VAZIAS					
	1960		1970		1980	
	% de Cidades	% da População	% de Cidades	% da População	% de Cidades	% da População
TOTAL ABSOLUTO	17	846 298	17	845 133	17	1 215 892
TOTAL RELATIVO	100,00	100,00	100,00	100,00	100,00	100,00
De 10 000 a 19 999 habitantes	58,83	16,00	44,07	8,42	41,19	7,59
20 000 a 49 999 habitantes	23,53	15,64	33,51	13,38	35,29	16,48
50 000 a 99 999 habitantes	5,88	7,62	10,55	9,68	11,76	11,70
100 000 a 499 999 habitantes	11,76	60,74	10,55	27,53	5,88	13,81
500 000 a 999 999 habitantes	-	-	0,53	4,79	5,88	50,42
1 000 000 e mais habitantes	-	-	0,79	36,20	-	-

FONTE - IBGE. Sinopse dos Censos Demográficos de 1960, 1970 e 1980.

TABELA 8

SUBDIVISÃO DAS ZONAS DE OCUPAÇÃO POR GRAU DE DISPERSÃO
POPULAÇÃO URBANA QUE APRESENTA CADA UNIDADE DA FEDERAÇÃO - 1960-1980

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	GRAU DE DISPERSÃO DA POPULAÇÃO URBANA (1) E (2)								
	1960			1970			1980		
	Alta	Média	Baixa	Alta	Média	Baixa	Alta	Média	Baixa
TOTAL									
Ceará		X		X			X		
Rio Grande do Norte		X			X			X	
Paraíba		X			X			X	
Pernambuco			X		X			X	
Alagoas		X			X			X	
Sergipe		X			X			X	
Espírito Santo		X			X			X	
Rio de Janeiro		X	X			X			X
São Paulo			X			X			X
Santa Catarina		X			X			X	
Rio Grande do Sul (e)		X				X			X
PARCIAL									
Minas Gerais (c)		X			X			X	
Paraná (a, d)		X			X			X	
Distrito Federal (b, e)			X			X	X		
SUBOCUPADA									
Maranhão	X				X			X	
Piauí	X	X			X			X	
Bahia		X			X			X	
Goiás	X				X			X	
VAZIA									
Rondônia (f)		X				X			X
Acre (c, f)		X			X			X	
Amazonas		X				X			X
Roraima			X			X			X
Pará (c, f)		X			X				X
Amapá			X			X			X
Mato Grosso (f)		X			X			X	X(3)

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.
NOTAS - (a, b) Passaram em 1970 para as zonas de ocupação total; (c) passaram em 1970 para as zonas de ocupação subocupada; (d) passaram em 1980 para as zonas de ocupação total; (e) passaram em 1980 para as zonas de ocupação parcial; (f) passaram em 1980 para as zonas de subocupação.
(1) Grau de dispersão: 0 - 49% alta; 50 - 74% média e 75 - 100% baixa.
(2) Em função da porcentagem de população que vive nos centros urbanos de 10 000 e mais habitantes da população urbana; (3) Mato Grosso do Sul.

TABELA 9

REPARTIÇÃO DO CRESCIMENTO RELATIVO DA PEA, POR ZONAS DE OCUPAÇÃO, SEGUNDO SETORES DE ATIVIDADES ENTRE 1960-1970

SETORES DE ATIVIDADES	ZONAS DE OCUPAÇÃO			TOTAL
	Total	Subocupada (1)	Vazia	
TOTAL ABSOLUTO	-	-	-	6 920 369
TOTAL RELATIVO (%)	93,7	9,6	-3,3	100,0
Primário	19,6	-5,2	-3,1	11,3
Secundário	29,2	6,3	0,5	36,0
Terciário	44,9	8,5	-0,7	52,7

FONTE - IBGE. Censos Demográficos de 1960 e 1970.
(1) Subocupada inclui as zonas de ocupação parcial e subocupação parcial.

TABELA 10

DISTRIBUIÇÃO DO CRESCIMENTO RELATIVO DA PEA, POR ZONAS DE OCUPAÇÃO DEMOGRAFICAS, SEGUNDO SETORES DE ATIVIDADES ENTRE 1970-1980

SETORES DE ATIVIDADES	ZONAS DE OCUPAÇÃO DEMOGRAFICAS			TOTAL
	Total	Subocupada (1)	Vazia	
TOTAL ABSOLUTO	-	-	-	13 678 835
TOTAL RELATIVO (%)	42,9	56,6	0,5	100,0
Primário	-11,3	9,2	-1,0	-3,1
Secundário	23,7	15,8	0,5	40,0
Terciário	30,6	31,6	0,9	63,1

FONTE - IBGE. Censos Demográficos de 1970 e 1980.
(1) Inclui as zonas de subocupação total e parcial.

TABELA 11

INDICADORES SÓCIO-ECONÔMICOS,
SEGUNDO ZONAS DE OCUPAÇÃO - 1960-1980

(continua)

ZONAS DE OCUPAÇÃO E ANOS	TAXA DE ALFABETIZAÇÃO (5 ANOS E MAIS)	% DA PEA NA POPULAÇÃO DE 10 ANOS E MAIS (TAXA BRUTA DE ATIVIDADE)	TAXA DE ATIVIDADE (1)
OCUPAÇÃO TOTAL			
1960	58,8	46,2	48,6
1970	65,6	45,3	47,6
1980	71,3	50,0	53,0
OCUPAÇÃO PARCIAL			
1960	52,3	46,6	48,3
1970	-	-	-
1980	75,0	49,8	52,7
SUBOCUPAÇÃO			
1960	35,7	47,3	49,2
1970	49,9	44,2	46,0
1980	52,9	46,9	49,4
VAZIA			
1960	50,0	46,0	47,6
1970	55,6	45,4	46,8
1980	60,6	47,7	49,5
TOTAL			
1960	53,3	46,4	48,6
1970	60,5	45,0	47,1
1980	68,0	49,3	52,1

TABELA 11

INDICADORES SÓCIO-ECONÔMICOS,
SEGUNDO ZONAS DE OCUPAÇÃO - 1960-1980

(conclusão)

ZONAS DE OCUPAÇÃO E ANOS	PORCENTAGEM DA POPULAÇÃO ECONOMICAMENTE ATIVA (2)			URBANIZAÇÃO
	Setor Primário	Setor Secundário	Setor Terciário	
OCUPAÇÃO TOTAL				
1960	45,2	15,9	38,9	48,7
1970	37,2	21,0	41,8	62,1
1980	23,2	29,4	47,4	80,3
OCUPAÇÃO PARCIAL				
1960	62,9	9,4	27,7	36,6
1970	-	-	-	-
1980	29,5	24,4	46,1	68,8
SUBOCUPAÇÃO				
1960	73,7	5,3	21,0	45,4
1970	59,1	11,5	29,5	44,0
1980	50,7	14,8	34,5	48,6
VAZIA				
1960	66,2	5,8	27,9	37,7
1970	58,7	11,6	29,7	43,6
1980	40,7	19,4	39,9	58,9
TOTAL				
1960	54,4	12,4	33,2	44,7
1970	44,3	17,9	37,8	55,9
1980	30,0	25,5	44,5	67,6

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.

(1)Taxa de Atividade = $\frac{\text{PEA}}{\text{População 10 a 64 anos}} \times 100$

(2) Exclusive os que estavam procurando trabalho em 1980.

TABELA 12

CRESCIMENTO DA POPULAÇÃO ECONOMICAMENTE ATIVA E TAXAS DE CRESCIMENTO MÉDIO GEOMÉTRICO ANUAL, SEGUNDO ZONAS DE OCUPAÇÃO E SETORES DE ATIVIDADE - 1960-1980

ZONAS DE OCUPAÇÃO E SETORES DE ATIVIDADE	CRESCIMENTO DA POPULAÇÃO ECONOMICAMENTE ATIVA				TAXAS DE CRESCIMENTO MÉDIO GEOMÉTRICO ANUAL DA PEA	
	Entre 1960-1970		Entre 1970-1980			
	Dados Absolutos	Dados Relativos (%)	Dados Absolutos (1)	Dados Relativos (%)	1960-1970	1970-1980
TOTAL	6 920 269	30,6	13 678 835	46,3	2,70	3,88
Setor primário .	780 348	6,3	- 429 322	- 3,3	0,62	- 0,33
Setor secundário	2 488 605	88,7	5 477 065	103,4	6,55	7,36
Setor terciário.	3 651 316	48,6	7 666 906	68,6	4,04	5,36
OCUPAÇÃO TOTAL ...	6 553 865	48,9	5 868 084	29,4	4,06	2,61
Setor primário .	1 359 298	22,4	- 1 550 319	- 20,9	2,04	- 2,32
Setor secundário	2 062 680	96,9	3 235 732	77,2	7,01	5,89
Setor terciário.	3 131 887	60,2	3 619 455	43,4	4,82	3,67
SUBOCUPADA (2) ...	594 358	7,3	7 743 206	88,3	0,70	6,53
Setor primário .	- 364 624	- 6,6	1 253 445	24,2	- 0,68	2,19
Setor secundário	390 753	63,4	2 167 921	215,2	5,03	12,17
Setor terciário.	568 229	28,2	3 943 817	152,6	2,51	9,71
VAZIA	- 227 954	- 21,4	67,545	8,1	- 2,38	0,78
Setor primário .	- 214 326	- 30,4	- 132 448	- 26,9	- 3,55	- 3,09
Setor secundário	35 172	56,5	73 412	75,4	4,58	5,78
Setor terciário.	- 48 800	- 16,4	103 634	41,6	- 1,77	3,54

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.
(1) Inclusive os que estavam procurando trabalho em 1980. (2) Inclui as zonas subocupadas e parcial em 1960 e 1980.

TABELA 13

CRESCIMENTO DA POPULAÇÃO DE 5 ANOS E MAIS, SEGUNDO ZONAS DE OCUPAÇÃO E CONDIÇÃO DE ALFABETIZAÇÃO - 1960-1980

ZONAS DE OCUPAÇÃO E SETORES DE ATIVIDADE	CRESCIMENTO DA POPULAÇÃO DE 5 ANOS E MAIS				TAXA DE CRESCIMENTO MÉDIO ANUAL DA POPULAÇÃO DE 5 ANOS E MAIS (%)	
	Entre 1960-1970		Entre 1970-1980			
	Dados Absolutos	Dados Relativos (%)	Dados Absolutos	Dados Relativos (%)	1960-1970	1970-1980
TOTAL	20 453 881	34,9	23 307 712	29,5	3,04	2,61
Alfabetizados ..	16 593 344	53,1	21 839 462	45,6	4,35	3,83
Analfabetos	3 860 537	14,1	1 468 250	4,7	1,33	0,46
OCUPAÇÃO TOTAL ...	18 098 756	52,5	7 201 468	13,7	4,31	1,29
Alfabetizados ..	14 211 852	70,1	8 159 629	23,7	5,46	2,15
Analfabetos	3 886 904	27,4	- 958 161	- 5,3	2,45	- 0,54
SUBOCUPADA	2 919 339	13,7	16 077 631	66,2	1,29	4,00
Alfabetizados ..	2 536 992	26,5	13 547 903	111,9	2,38	7,80
Analfabetos	382 347	3,2	2 529 728	20,8	0,32	1,91
VAZIA	- 564 214	- 19,7	28 613	1,2	- 2,18	0,12
Alfabetizados ..	- 155 500	- 10,9	131 930	10,4	- 1,14	0,99
Analfabetos	- 408 714	- 28,6	- 103 317	- 10,1	- 3,32	- 1,06

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.
NOTA - A zona subocupada inclui a zona de subocupação total e parcial.

TABELA 14

SUPERFICIE APROVEITADA E INDICES DE APROVEITAMENTO COM AGROPECUARIA, SEGUNDO ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1960-1980

(continua)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	SUPERFICIE (EM Km²) (A)	SUPERFICIE APROVEITADA COM AGROPECUARIA (Km²) (B)		
		1960	1970	1980
BRASIL	8 456 508	1 509 299	2 215 314	2 317 281
OCUPAÇÃO TOTAL	1 103 364	544 282	838 034	591 927
Ceará	146 817	49 472	81 361	72 048
Rio Grande do Norte	53 015	24 627	33 220	22 800
Paraíba	56 372	28 912	37 644	34 038
Pernambuco	98 281	33 474	51 749	43 349
Alagoas	27 652	9 754	17 766	18 146
Sergipe	21 994	9 151	14 725	14 944
Espírito Santo	45 597	15 815	29 291	28 609
Rio de Janeiro (1) (2)	43 305	20 898	26 234	24 003
São Paulo	247 320	146 959	173 240	164 072
Santa Catarina	95 483	30 068	49 096	46 842
Rio Grande do Sul (e)	267 528	175 152	207 982	204 490
OCUPAÇÃO PARCIAL	787 417	358 187	-	556 266
Minas Gerais (c)	582 586	295 908	356 694	349 882
Paraná (a, d)	199 060	61 393	114 320	123 026
Distrito Federal (b, e)	5 771	886	1 406	1 894
SUBOCUPADA	1 777 537	350 566	1 000 637	975 349
Maranhão	324 616	33 697	80 602	73 319
Piauí	250 934	30 788	65 452	54 659
Bahia	559 951	84 503	152 650	184 055
Goiás	642 036	201 578	292 912	353 449
OCUPAÇÃO VAZIA	4 782 871	256 264	376 643	193 739
Rondônia (f)	243 044	165	5 182	11 656
Acre (c, f)	152 589	409	1 820	3 921
Amazonas	1 558 987	2 184	8 607	8 517
Roraima	230 104	7 100	11 800	16 635
Pará (c, f)	1 227 530	12 871	53 507	63 120
Amapá	139 068	3 813	3 688	2 279
Mato Grosso	881 001	229 722	347 366	166 308
Mato Grosso do Sul (f)	350 548	-	-	231 170

TABELA 14

SUPERFICIE APROVEITADA E INDICES DE APROVEITAMENTO COM AGROPECUARIA, SEGUNDO ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1960-1980

(continua)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	INDICE DE APROVEITAMENTO COM AGROPECUARIA B/A.100			CRESCIMENTO RELATIVO DA SUPERFICIE APROVEITADA $= \left(\frac{B-A}{A}\right) \cdot 100$	
	1960	1970	1980	1960-1970	1970-1980
BRASIL	17,85	26,20	27,40	46,78	- 4,60
OCUPAÇÃO TOTAL	49,33	64,06	57,20	53,97	- 29,37
Ceará	33,70	55,42	49,07	64,46	- 11,45
Rio Grande do Norte	46,45	62,66	43,01	34,89	- 31,37
Paraíba	51,29	66,78	60,38	30,20	- 9,58
Pernambuco	34,06	52,65	44,11	54,59	- 16,23
Alagoas	35,27	64,25	65,62	82,14	2,14
Sergipe	41,61	66,95	68,17	60,91	1,83
Espírito Santo	34,68	64,24	62,74	85,21	- 2,33
Rio de Janeiro (1) (2)	48,26	60,58	55,43	25,53	- 8,50
São Paulo	59,42	70,05	66,34	17,88	- 5,29
Santa Catarina	31,49	51,42	49,06	63,28	- 4,59
Rio Grande do Sul (e)	65,47	77,74	76,44	18,74	- 1,68
OCUPAÇÃO PARCIAL	45,49	-	64,99	-	-
Minas Gerais (c)	50,79	60,06	60,06	19,53	- 1,08
Paraná (a, d)	30,84	61,80	61,80	86,21	7,62
Distrito Federal (b, e)	15,35	32,82	32,82	58,69	34,71
SUBOCUPADA	19,72	26,75	26,00	185,43	- 2,53
Maranhão	10,38	24,83	22,59	139,20	- 9,04
Piauí	12,27	26,08	21,78	112,59	- 16,49
Bahia	15,09	27,26	32,87	80,64	20,57
Goiás	31,40	45,62	55,05	45,31	20,67
OCUPAÇÃO VAZIA	5,36	11,07	6,90	46,97	- 48,56
Rondônia (f)	0,07	2,13	4,80	3 040,61	124,93
Acre (c, f)	0,27	1,19	2,57	344,99	115,44
Amazonas	0,14	0,55	0,55	294,09	- 1,05
Roraima	3,09	5,13	7,23	66,20	40,97
Pará (c, f)	1,05	4,36	5,14	315,72	17,97
Amapá	2,74	2,65	1,64	- 3,28	- 38,20

TABELA 14

SUPERFICIE APROVEITADA E INDICES DE APROVEITAMENTO COM AGROPECUARIA, SEGUNDO ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1960-1980

(conclusão)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	INDICE DE APROVEITAMENTO COM AGROPECUARIA B/A.100			CRESCIMENTO RELATIVO DA SUPERFICIE APROVEITADA $=\left(\frac{B-A}{A}\right)\cdot 100$	
	1960	1970	1980	1960-1970	1970-1980
Mato Grosso	18,65	28,21	18,88	51,21	- 52,12
Mato Grosso do Sul (f)	-	-	65,95	-	-

FONTE - IBGE. Censo Agropecúario de 1960, 1970 e 1980.

NOTAS - (1) Indice de aproveitamento com agropecúaria corresponde ao percentual de áreas utilizadas em lavouras, pastagens e terras em descanso sobre a superfície total de cada zona demográfica de cada respectiva Unidade da Federação. (2) a , b - passaram em 1970 para as zonas de ocupação total. c - passaram em 1970 para as zonas de ocupação subocupada. d - passaram em 1980 para as zonas de ocupação total. e - passaram em 1980 para as zonas de ocupação parcial. f - passaram em 1980 para as zonas de ocupação subocupada. (1) Em 1960 o Distrito Federal era o atual município do Rio de Janeiro. (2) Inclui em 1970 e 1980 o antigo Estado da Guanabara e o Estado do Rio de Janeiro.

TABELA 15

TAXAS DE INDUSTRIALIZAÇÃO E DE TERCIARIZAÇÃO, SEGUNDO ZONAS DE OCUPAÇÃO - 1960-1980

ZONAS DE OCUPAÇÃO	TAXA DE INDUSTRIALIZAÇÃO (%) (1)		TAXA DE TERCIARIZAÇÃO (%) (1)	
	Entre 1960-1970	Entre 1970-1980	Entre 1960-1970	Entre 1970-1980
BRASIL	6,14	6,82	3,91	5,11
OCUPAÇÃO TOTAL	6,53	4,89	4,63	2,61
SUBOCUPADA (2)	4,81	10,37	2,47	8,66
VAZIA	4,41	4,88	- 1,79	2,71

FONTE - IBGE. Censos Demográficos de 1960, 1970 e 1980.

(1) Taxa de industrialização de terciarização:

$$= \frac{2}{n} \times [(S_1 - S_0)/(S_1 + S_0)] \times 100$$

S_1 = PEA no setor respectivo (secundário ou terciário) no ano $t + n$

S_0 = PEA no setor respectivo (secundário ou terciário) no ano t

n = Número de anos transcorridos entre t , $t + n$

(2) As zonas subocupadas incluíram em 1960 e 1980 as de ocupação parcial.

TABELA 16

DENSIDADE POPULACIONAL E INDICES DE OCUPAÇÃO DEMOGRAFICA DO TERRITÓRIO URBANO E RURAL, SEGUNDO AS UNIDADES DA FEDERAÇÃO - BRASIL - 1960

UNIDADES DA FEDERAÇÃO	DENSIDADES POPULACIONAIS			INDICES DE OCUPAÇÃO DEMOGRAFICA		
	t D_i	u D_i	r D_i	t I_i	u I_i	r I_i
BRASIL	8,29	3,70	4,59	0,0199	0,0219	0,0169
Rondônia	0,29	0,12	0,16	0,0007	0,0007	0,0008
Acre	1,04	0,21	0,82	0,0025	0,0012	0,0039
Amazonas	0,45	0,15	0,31	0,0011	0,0009	0,0015
Roraima	0,12	0,05	0,07	0,0003	0,0003	0,0003
Pará	1,25	0,50	0,74	0,0030	0,0030	0,0035
Amapá	0,49	0,25	0,24	0,0012	0,0015	0,0011
Maranhão	7,61	1,35	6,26	0,0183	0,0080	0,0299
Piauí	4,95	1,14	3,81	0,0119	0,0068	0,0182
Ceará	22,45	7,48	14,97	0,0569	0,0444	0,0714
Rio Grande do Norte	21,61	8,08	13,53	0,0519	0,0479	0,0645
Paraíba	35,49	12,37	23,12	0,0852	0,0734	0,1103
Pernambuco	41,67	18,57	23,10	0,1000	0,1101	0,1102
Alagoas	45,50	15,20	(30,23)	0,1092	0,0901	(0,1111)
Sergipe	34,20	13,18	21,02	0,0821	0,0782	0,1002
Bahia	10,57	3,63	6,94	0,0254	0,0215	0,0331
Minas Gerais	17,01	6,60	10,41	0,0408	0,0391	0,0496
Espírito Santo	28,47	8,31	20,17	0,0633	0,0493	0,0962
Rio de Janeiro (1)	152,66	120,42	(32,24)	(0,3664)	(0,7141)	(0,1150)
São Paulo	51,79	32,43	19,37	0,1243	0,1923	0,0924
Paraná	21,44	6,56	14,88	0,0515	0,0389	0,0710
Santa Catarina	22,18	7,06	15,12	0,0532	0,0419	0,0721
Rio Grande do Sul	20,06	8,90	11,16	0,0481	0,0528	0,0532
Mato Grosso	0,72	0,28	0,44	0,0017	0,0017	0,0021
Goiás	2,98	0,90	2,08	0,0072	0,0053	0,0099
Distrito Federal	24,29	15,31	0,98	0,0583	0,0908	0,0047
$\sum_{i=1}^{n}$ (2)	416,63	168,63	247,99	1,0000	1,0000	1,0000
$\bar{D}_T$	22,77	11,56	11,21	-	-	-

FONTE - IBGE. Censo Demográfico de 1960.

(1) Inclui o Estado da Guanabara e o Rio de Janeiro. (2) O Somatório dos índices é igual a 1, mas para efeito de cálculo do grau relativo das UFs, calculou-se o índice de ocupação demográfica para o Rio de Janeiro e Pernambuco, na área rural, em separado. Assim, encontram-se excluídas do somatório, por apresentarem valor de I_i superior a 10,0% do $\sum D_i$

TABELA 17

DENSIDADE POPULACIONAL E INDICES DE OCUPAÇÃO DEMOGRAFICA DO TERRITÓRIO, TOTAL, URBANO E RURAL, SEGUNDO AS UNIDADES DA FEDERAÇÃO - BRASIL - 1970

UNIDADES DA FEDERAÇÃO	DENSIDADES POPULACIONAIS			INDICES DE OCUPAÇÃO DEMOGRAFICA		
	t	u	r	t	u	r
	D/i	D/i	D/i	I/i	I/i	I/i
BRASIL	11,02	6,16	4,86	0,0216	0,0245	0,0189
Rondônia	0,46	0,25	0,21	0,0009	0,0010	0,0008
Acre	1,41	0,39	1,02	0,0028	0,0015	0,0040
Amazonas	0,61	0,26	0,35	0,0012	0,0010	0,0014
Roraima	0,18	0,08	0,10	0,0004	0,0003	0,0004
Pará	1,77	0,83	0,93	0,0035	0,0033	0,0036
Amapá	0,82	0,45	0,37	0,0016	0,0018	0,0014
Maranhão	9,22	2,32	6,90	0,0181	0,0092	0,0268
Piauí	6,70	2,14	4,56	0,0132	0,0085	0,0177
Ceará	29,71	12,12	17,58	0,0583	0,0481	0,0683
Rio Grande do Norte	29,24	13,91	15,33	0,0574	0,0553	0,0595
Paraíba	42,27	17,78	24,49	0,0830	0,0706	0,0950
Pernambuco	52,51	28,60	23,91	0,1031	0,1136	0,0928
Alagoas	57,43	22,85	34,59	0,1127	0,0908	0,1342
Sergipe	40,95	18,89	22,07	0,0804	0,0750	0,0857
Bahia	13,38	5,51	7,87	0,0263	0,0219	0,0306
Minas Gerais	19,72	10,40	9,32	0,0387	0,0413	0,0362
Espírito Santo	35,08	15,83	19,24	0,0689	0,0629	0,0747
Rio de Janeiro	(207,71)	(182,57)	25,14	(0,2563)	(0,3486)	(0,0878)
São Paulo	71,86	57,72	14,13	0,1410	0,2294	0,0549
Paraná	34,81	12,58	22,23	0,0683	0,0500	0,0863
Santa Catarina	30,39	13,05	17,34	0,0597	0,0518	0,0673
Rio Grande do Sul	24,91	13,28	11,63	0,0489	0,0528	0,0452
Mato Grosso	1,30	0,56	0,74	0,0026	0,0022	0,0029
Goiás	4,58	1,93	2,65	0,0090	0,0077	0,0103
Distrito Federal	*(93,14)*	*(89,43)*	3,71	*(0,1150)*	*(0,1708)*	*(0,0130)*
$\sum_{i=1}^{n}$ (1)	509,31	251,73	257,56	1,0000	1,0000	1,0000
$\bar{D}_T$	32,41	20,95	11,46	-	-	-

FONTE - IBGE. Censo Demográfico de 1970.

(1) O somatório do índice é igual a 1, mas para efeito do grau relativo das UFs, calcula-se o índice de ocupação demográfica para o Rio de Janeiro e Distrito Federal, que se encontram excluídos do somatório, por apresentarem valor de I_i superior a 10,0% do $\sum D_i$

TABELA 18

DENSIDADE POPULACIONAL E INDICES DE OCUPAÇÃO DEMOGRAFICA DO TERRITÓRIO, TOTAL, URBANO E RURAL, SEGUNDO AS UNIDADES DA FEDERAÇÃO - BRASIL - 1980

UNIDADES DA FEDERAÇÃO	DENSIDADES POPULACIONAIS			INDICES DE OCUPAÇÃO DEMOGRAFICA		
	D_i^t	D_i^u	D_i^r	I_i^t	I_i^u	I_i^r
BRASIL	14,09	9,52	4,57	0,0262	0,0313	0,0195
Rondônia	2,03	0,96	1,07	0,0038	0,0032	0,0046
Acre	1,98	0,87	1,11	0,0037	0,0029	0,0047
Amazonas	0,92	0,55	0,37	0,0017	0,0018	0,0016
Roraima	0,34	0,21	0,13	0,0006	0,0007	0,0006
Pará	2,78	1,36	1,42	0,0052	0,0045	0,0061
Amapá	1,26	0,75	0,52	0,0023	0,0025	0,0022
Maranhão	12,33	3,87	8,46	0,0229	0,0127	0,0361
Piauí	8,53	3,58	4,95	0,0158	0,0118	0,0211
Ceará	36,06	19,17	16,90	0,0670	0,0631	0,0721
Rio Grande do Norte	35,83	21,04	14,79	0,0665	0,0692	0,0631
Paraíba	49,18	25,73	23,46	0,0913	0,0846	0,1000
Pernambuco	62,55	38,52	24,03	0,1162	0,1267	0,1025
Alagoas	71,88	35,39	36,49	0,1335	0,1164	0,1556
Sergipe	51,92	28,11	23,80	0,0964	0,0925	0,1015
Bahia	16,92	8,33	8,59	0,0314	0,0274	0,0366
Minas Gerais	22,99	15,43	7,56	0,0427	0,0508	0,0322
Espírito Santo	44,38	28,36	16,02	0,0824	0,0933	0,0683
Rio de Janeiro	260,88	239,55	21,32	(0,4844)	(0,7880)	(0,0909)
São Paulo	101,25	89,74	11,50	(0,1880)	(0,2952)	(0,0490)
Paraná	38,33	22,47	15,86	0,0712	0,0739	0,0676
Santa Catarina	38,00	22,56	15,44	0,0706	0,0742	0,0658
Rio Grande do Sul	29,07	19,63	9,44	0,0540	0,0646	0,0403
Mato Grosso do Sul	3,91	2,62	1,29	0,0073	0,0086	0,0055
Mato Grosso	1,30	0,75	0,55	0,0024	0,0025	0,0023
Goiás	6,02	3,74	2,20	0,0112	0,0123	0,0097
Distrito Federal	204,02	197,45	6,57	(0,3789)	(0,6495)	(0,0280)
$\sum_{i=1}^{n}$ (1)	538,51	304,00	234,54	1,0000	1,0000	1,0000
$\bar{D}_T$	23,41	13,22	10,20	-	-	-

FONTE - IBGE. Censo Demográfico de 1980.

(1) O somatório do índice é igual a 1, mas para efeito de cálculo do grau relativo das UFs calcula-se o índice de ocupação demográfica para o Rio de Janeiro e Distrito Federal, que se encontram excluídos do somatório, por apresentarem valor de I_i superior a 10,0% do $\sum D_i$

TABELA 19

INDICADORES SÓCIO-ECONÔMICOS, SEGUNDO UNIDADES DA FEDERAÇÃO - 1960

(continua)

UNIDADES DA FEDERAÇÃO	POPULAÇÃO RESIDENTE DE 5 ANOS E MAIS (1)		POPULAÇÃO RESIDENTE DE 10 ANOS E MAIS	
	Total	Alfabetizada	Total	10 - 64 anos
TOTAL (2)	58 689 372	31 271 187	48 748 141	46 574 586
Rondônia	56 979	26 528	46 278	45 278
Acre	128 795	40 375	102 278	98 695
Amazonas	583 076	250 246	470 755	455 422
Roraima	22 975	10 387	17 888	17 469
Pará	271 124	662 495	1 035 767	995 792
Amapá	54 873	28 043	43 711	42 462
Maranhão	2 060 567	621 325	1 666 200	1 612 493
Piauí	1 029 828	284 494	831 853	800 608
Ceará	2 743 910	912 797	2 256 975	2 144 588
Rio Grande do Norte ..	951 726	364 976	792 678	749 988
Paraíba	1 677 167	556 189	1 383 463	1 317 354
Pernambuco	3 428 181	1 276 906	2 835 406	2 713 333
Alagoas	1 050 468	287 547	1 010 740	818 876
Sergipe	624 763	218 468	508 694	481 644
Bahia	4 934 589	1 810 726	4 048 458	3 863 657
Minas Gerais	8 083 235	4 071 211	6 616 201	6 368 153
Espírito Santo	960 185	484 375	772 962	746 005
Rio de Janeiro	2 832 963	1 779 211	2 336 748	2 241 504
Guanabara	2 897 689	2 373 403	2 544 327	2 409 708
São Paulo	10 987 015	7 647 399	9 323 041	8 909 274
Paraná	3 533 198	1 989 053	2 881 145	2 795 674
Santa Catarina	1 749 935	1 165 073	1 409 591	1 356 824
Rio Grande do Sul	4 575 755	3 207 042	3 816 200	3 646 576
Mato Grosso	740 039	411 655	603 125	584 486
Goiás	1 589 734	711 146	1 287 097	1 253 343
Distrito Federal	120 603	80 117	106 493	105 410

TABELA 19

INDICADORES SÓCIO-ECONÔMICOS, SEGUNDO UNIDADES DA FEDERAÇÃO - 1960

(conclusão)

UNIDADES DA FEDERAÇÃO	POPULAÇÃO ECONOMICAMENTE ATIVA			
	Total	Setor Primário	Setor Secundário	Setor Terciário
TOTAL	22 636 608	12 309 991	2 806 793	7 519 824
Rondônia	22 552	12 263	1 512	7 777
Acre	48 056	38 103	1 470	8 483
Amazonas	212 949	151 041	10 874	51 034
Roraima	7 855	5 055	266	2 534
Pará	475 514	301 466	31 731	142 317
Amapá	18 934	9 097	2 211	7 626
Maranhão	790 228	651 378	23 271	115 579
Piauí	378 893	287 081	19 029	72 783
Ceará	1 042 871	690 935	127 055	224 881
Rio Grande do Norte ..	337 216	231 202	18 672	87 342
Paraíba	595 354	437 615	34 196	123 543
Pernambuco	1 304 573	806 596	120 487	377 490
Alagoas	426 009	314 036	29 633	82 340
Sergipe	258 598	178 057	21 646	58 695
Bahia	1 942 181	1 355 576	122 369	464 236
Minas Gerais	2 989 798	1 827 334	285 415	877 049
Espírito Santo	352 474	240 223	19 961	92 290
Rio de Janeiro	981 494	297 106	207 001	477 387
Guanabara	1 176 228	20 927	237 209	918 092
São Paulo	4 517 594	1 475 732	1 053 310	1 988 552
Paraná	1 413 460	983 404	93 323	336 733
Santa Catarina	641 195	402 149	70 300	168 746
Rio Grande do Sul	1 759 420	964 460	188 486	606 474
Mato Grosso	280 403	188 226	14 164	78 013
Goiás	591 476	436 382	30 673	124 421
Distrito Federal	71 283	3 547	42 529	25 207

FONTE - IBGE. Censo Demográfico de 1960.

(1) Inclusive os de idade ignorada; (2) Excluído Fernando de Noronha.

TABELA 20

INDICADORES SÓCIO-ECONÔMICOS, POR ZONAS DE OCUPAÇÃO
E UNIDADES DA FEDERAÇÃO - 1960

(continua)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	TAXA DE ALFABETIZAÇÃO DA POPULAÇÃO DE 5 ANOS E MAIS	PORCENTAGEM DA POPULAÇÃO DE 10 A 64 ANOS NA POPULAÇÃO DE 10 ANOS E MAIS	PORCENTAGEM DA PEA NA POPULAÇÃO DE 10 ANOS E MAIS
TOTAL	53,3	95,5	46,4
OCUPAÇÃO TOTAL	58,8	95,0	46,2
Ceará	33,3	95,0	46,2
Rio Grande do Norte	38,3	94,6	42,5
Paraíba	33,2	95,2	43,0
Pernambuco	37,2	95,7	46,0
Alagoas	27,4	81,0	42,1
Sergipe	35,0	94,7	50,8
Espírito Santo	50,4	96,5	45,6
São Paulo	69,6	95,6	48,5
Santa Catarina	66,6	96,3	45,5
Rio Grande do Sul	70,1	95,6	46,1
Rio de Janeiro	72,5	95,3	44,2
OCUPAÇÃO PARCIAL	52,3	96,5	46,6
Minas Gerais	50,4	96,3	45,2
Paraná	56,3	97,0	49,1
Distrito Federal	66,4	99,0	66,9
SUBOCUPAÇÃO	35,7	96,1	47,3
Maranhão	30,2	96,8	47,4
Piauí	27,6	96,2	45,5
Bahia	36,7	95,4	48,0
Goiás	44,7	97,4	46,0
VAZIA	50,0	96,5	46,0
Rondônia	46,6	97,7	48,7
Acre	31,3	96,5	47,0
Amazonas	42,9	96,7	45,2
Roraima	45,2	97,7	43,9
Pará	52,1	96,1	45,9
Amapá	51,1	97,1	43,3
Mato Grosso	55,6	96,9	46,5

TABELA 20

INDICADORES SÓCIO-ECONÔMICOS, POR ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1960

(conclusão)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	TAXA DE ATIVIDADE (1)	PORCENTAGEM DA POPULAÇÃO ECONOMICAMENTE ATIVA (PEA)		
		Setor Primário	Setor Secundário	Setor Terciário
TOTAL	48,6	54,4	12,4	33,2
OCUPAÇÃO TOTAL	48,6	45,2	15,9	38,9
Ceará	48,6	66,2	12,2	21,6
Rio Grande do Norte	45,0	68,6	5,5	25,9
Paraíba	45,2	73,5	5,7	20,8
Pernambuco	48,1	61,8	9,3	28,9
Alagoas	52,0	73,7	7,0	19,3
Sergipe	53,7	68,8	8,4	22,8
Espírito Santo	47,2	68,1	5,7	26,2
São Paulo	50,7	32,7	23,3	44,0
Santa Catarina	47,3	62,7	11,0	26,3
Rio Grande do Sul	48,2	54,8	10,7	34,5
Rio de Janeiro	46,4	14,7	20,6	64,7
OCUPAÇÃO PARCIAL	48,3	62,9	9,4	27,7
Minas Gerais	46,9	61,1	9,6	29,3
Paraná	50,6	69,6	6,6	23,8
Distrito Federal	67,6	5.0	59,6	35,4
SUBOCUPAÇÃO	49,2	73,7	5,3	21,0
Maranhão	49,0	82,4	3,0	14,6
Piauí	47,3	75,8	5,0	19,2
Bahia	50,3	69,8	6,3	23,9
Goiás	47,2	73,8	5,2	21,0
VAZIA	47,6	66,2	5,9	27,9
Rondônia	49,8	58,8	6,7	34,5
Acre	48,7	79,3	3,1	17,7
Amazonas	46,8	70,9	5,1	24,0
Roraima	45,0	64,4	3,4	32,3
Pará	47,8	63,4	6,7	29,9
Amapá	44,6	48,0	11,7	40,3
Mato Grosso	48,0	67,1	5,1	27,8

FONTE - IBGE. Censo Demográfico de 1960.

(1) Taxa de Atividade = $\frac{\text{PEA}}{\text{População 10 a 64 anos}} \times 100$

TABELA 21

INDICADORES SÓCIO-ECONÔMICOS, SEGUNDO UNIDADES DA FEDERAÇÃO - 1970

(continua)

UNIDADES DA FEDERAÇÃO	POPULAÇÃO DE 5 ANOS E MAIS		POPULAÇÃO DE 10 ANOS E MAIS	
	Total	Alfabetizada	Total	10 - 64 anos
TOTAL	79 143 253	47 864 531	65 683 745	62 758 664
Rondônia	90 638	50 236	73 792	72 109
Acre	174 088	61 507	137 880	134 043
Amazonas	745 956	394 451	590 693	572 167
Roraima	33 514	18 471	26 709	26 709
Pará	1 784 328	1 032 492	1 455 061	1 399 345
Amapá	92 127	53 550	73 291	71 178
Maranhão	2 487 409	862 438	2 018 581	1 949 625
Piauí	1 379 716	439 899	1 107 136	1 061 446
Ceará	3 610 262	1 369 494	2 922 767	2 774 002
Rio Grande do Norte ..	1 285 743	516 348	1 044 934	983 483
Paraíba	1 982 703	761 639	1 620 326	1 528 466
Pernambuco (1)	4 353 781	1 863 571	3 583 766	3 417 203
Alagoas	1 314 803	436 716	1 069 425	1 022 808
Sergipe	745 687	303 836	603 997	571 476
Bahia	6 221 968	2 620 715	5 069 462	4 839 575
Minas Gerais	9 775 948	5 769 124	8 066 745	7 732 858
Espírito Santo	1 351 623	809 415	1 106 951	1 062 102
Rio de Janeiro (2) ...	7 917 267	6 148 825	6 805 633	6 448 384
São Paulo	15 557 319	12 093 640	13 294 432	12 658 742
Paraná	5 765 804	3 602 663	4 683 582	4 530 406
Santa Catarina	2 444 831	1 823 423	1 990 306	1 907 956
Rio Grande do Sul	5 802 377	4 418 347	4 898 215	4 653 638
Mato Grosso	1 331 412	757 521	1 081 348	1 048 628
Goiás	2 447 616	1 318 885	1 991 503	1 932 106
Distrito Federal	446 313	337 321	367 210	360 876

TABELA 21

INDICADORES SÓCIO-ECONÔMICOS, SEGUNDO UNIDADES DA FEDERAÇÃO - 1970

(conclusão)

UNIDADES DA FEDERAÇÃO	POPULAÇÃO ECONOMICAMENTE ATIVA			
	Total	Setor Primário	Setor Secundário	Setor Terciário
TOTAL	29 556 877	13 090 339	5 295 398	11 171 140
Rondônia	33 903	15 915	4 214	13 774
Acre	64 540	44 328	3 985	16 227
Amazonas	169 333	161 648	29 679	78 006
Roraima	11 466	5 536	1 134	4 796
Pará	620 314	347 161	72 772	200 381
Amapá	29 104	11 287	5 659	12 158
Maranhão	973 160	762 900	49 184	161 076
Piauí	484 666	346 875	38 024	99 767
Ceará	1 255 440	749 090	163 754	342 596
Rio Grande do Norte ..	410 111	240 955	45 283	123 873
Paraíba	675 409	437 937	58 576	178 896
Pernambuco (1)	1 504 803	764 719	201 594	538 490
Alagoas	482 574	323 155	46 960	112 459
Sergipe	265 582	161 815	30 340	73 427
Bahia	2 301 697	1 437 364	254 230	110 103
Minas Gerais	3 460 615	1 717 333	512 060	1 231 222
Espírito Santo	457 787	240 383	62 264	155 140
Rio de Janeiro (2) ...	2 916 114	256 161	749 058	1 910 895
São Paulo	6 372 842	1 301 830	2 003 684	3 067 328
Paraná	2 276 754	1 438 838	232 575	605 341
Santa Catarina	882 229	451 697	174 020	256 512
Rio Grande do Sul	2 268 935	1 044 760	378 127	846 048
Mato Grosso	494 503	297 539	56 714	140 250
Goiás	866 685	524 117	77 107	265 461
Distrito Federal	178 311	6 996	44 401	126 914

FONTE - IBGE. Censo Demográfico de 1970.

(1) Inclusive Fernando de Noronha; (2) Inclusive o Estado da Guanabara.

TABELA 22

INDICADORES SÓCIO-ECONÔMICOS, POR ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1970

(continua)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	TAXA DE ALFABETIZAÇÃO DA POPULAÇÃO DE 5 ANOS E MAIS	PORCENTAGEM DA POPULAÇÃO DE 10 A 64 ANOS NA POPULAÇÃO DE 10 ANOS E MAIS	PORCENTAGEM DA PEA NA POPULAÇÃO DE 10 ANOS E MAIS
TOTAL	60,5	95,5	45,0
OCUPAÇÃO TOTAL	65,6	95,3	45,3
Ceará	37,9	94,9	43,0
Rio Grande do Norte	40,2	94,1	39,2
Paraíba	38,4	94,3	41,7
Pernambuco	42,8	95,4	42,0
Alagoas	33,2	95,6	45,1
Sergipe	40,7	94,6	44,0
Espírito Santo	59,9	95,9	41,4
Rio de Janeiro	77,7	94,8	42,8
São Paulo	77,7	95,2	47,9
Paraná	62,5	96,7	48,6
Santa Catarina	74,6	95,9	44,3
Rio Grande do Sul	76,1	95,0	46,3
Distrito Federal	75,6	98,3	48,6
SUBOCUPAÇÃO	49,9	96,0	44,2
Acre	35,3	97,2	46,8
Pará	57,9	96,2	42,6
Maranhão	34,7	96,6	48,2
Piauí	31,9	95,9	43,8
Bahia	42,1	95,5	45,4
Minas Gerais	59,0	95,9	42,9
Goiás	53,9	97,0	43,5
VAZIA	55,6	97,0	45,4
Rondônia	55,4	97,7	45,9
Amazonas	52,9	96,9	45,6
Roraima	55,1	97,5	42,9
Amapá	58,1	97,1	39,7
Mato Grosso	56,9	97,0	45,7

TABELA 22

INDICADORES SÓCIO-ECONÔMICOS, POR ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1970

(conclusão)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	TAXA DE ATIVIDADE (1)	PORCENTAGEM DA POPULAÇÃO ECONOMICAMENTE ATIVA (PEA)		
		Setor Primário	Setor Secundário	Setor Terciário
TOTAL	47,1	44,3	17,9	37,8
OCUPAÇÃO TOTAL	47,6	37,2	21,0	41,8
Ceará	45,3	59,7	13,0	27,3
Rio Grande do Norte	41,7	58,8	11,0	30,2
Paraíba	44,2	64,8	8,7	26,5
Pernambuco	44,0	50,8	13,4	35,8
Alagoas	47,2	67,0	9,7	23,3
Sergipe	46,5	60,9	11,5	27,6
Espírito Santo	43,1	52,5	13,6	33,9
Rio de Janeiro	45,2	8,8	25,7	65,5
São Paulo	50,3	20,5	31,4	48,1
Paraná	50,3	63,2	10,2	26,6
Santa Catarina	46,2	51,2	19,7	29,1
Rio Grande do Sul	48,8	46,0	16,7	37,3
Distrito Federal	49,4	3,9	24,9	71,2
SUBOCUPAÇÃO	46,0	59,0	11,5	29,5
Acre	48,1	68,7	6,2	25,1
Pará	44,3	56,0	11,7	32,3
Maranhão	49,9	78,3	5,1	16,6
Piauí	45,7	71,6	7,8	20,6
Bahia	47,6	62,4	11,0	26,5
Minas Gerais	44,8	49,6	14,8	35,6
Goiás	44,9	60,5	8,9	30,6
VAZIA	46,8	58,7	11,6	29,7
Rondônia	47,0	46,9	12,5	40,6
Amazonas	47,1	60,0	11,0	29,0
Roraima	44,0	48,3	9,9	41,8
Amapá	40,9	38,8	19,4	41,8
Mato Grosso	47,2	60,1	11,5	28,4

FONTE - IBGE. Censo Demográfico de 1970.

(1) Taxa de Atividade = $\frac{\text{PEA}}{\text{População 10 a 64 anos}} \times 100$

TABELA 23

INDICADORES SÓCIO-ECONÔMICOS - 1980

(continua)

UNIDADES DA FEDERAÇÃO	POPULAÇÃO RESIDENTE DE 5 ANOS E MAIS		POPULAÇÃO RESIDENTE DE 10 ANOS E MAIS	
	Total	Alfabetizada	Total	10 - 64 anos
TOTAL	102 450 965	69 703 993	87 677 224	82 906 792
Rondônia	404 350	239 649	330 678	322 975
Acre	245 553	109 339	198 609	191 020
Amazonas	1 161 793	683 171	941 520	906 347
Roraima	64 708	42 950	53 367	51 679
Pará	2 808 043	1 725 260	2 301 191	2 195 427
Amapá	141 497	93 233	113 338	108 238
Maranhão	3 285 046	1 417 608	2 702 963	2 559 707
Piauí	1 776 059	771 475	1 461 133	1 381 312
Ceará	4 485 731	2 138 214	3 760 077	3 525 511
Rio Grande do Norte ..	1 616 809	809 687	1 363 896	1 266 707
Paraíba	2 342 388	1 034 959	1 955 699	1 806 977
Pernambuco (2)	5 226 095	2 605 328	4 377 108	4 099 014
Alagoas	1 650 106	656 546	1 359 930	1 275 780
Sergipe	955 533	450 849	789 462	734 279
Bahia	7 896 001	3 877 554	6 541 367	6 149 277
Minas Gerais	11 581 903	8 072 663	9 931 096	9 393 183
Espírito Santo	1 733 187	1 217 880	1 482 765	1 406 290
Rio de Janeiro	10 034 661	8 228 279	8 889 192	8 356 809
São Paulo	21 977 985	18 034 452	19 311 632	18 291 350
Paraná	6 600 492	4 895 474	5 607 805	5 365 965
Santa Catarina	3 156 994	2 573 199	2 715 819	2 589 839
Rio Grande do Sul	6 893 549	5 709 011	6 088 091	5 726 237
Mato Grosso do Sul ...	1 170 911	807 270	985 964	946 394
Mato Grosso	954 262	586 805	791 717	764 488
Goiás	3 285 674	2 096 417	2 759 092	2 647 548
Distrito Federal	1 001 635	826 721	863 713	844 439

TABELA 23

INDICADORES SÓCIO-ECONÔMICOS - 1980

(conclusão)

UNIDADES DA FEDERAÇÃO	POPULAÇÃO ECONOMICAMENTE ATIVA			
	Total (1)	Setor Primário	Setor Secundário	Setor Terciário
TOTAL	43 235 712	12 661 017	10 772 463	18 838 046
Rondônia	171 448	89 167	23 017	56 747
Acre	93 065	48 134	8 036	35 418
Amazonas	445 174	176 680	93 190	165 190
Roraima	26 727	9 592	3 636	12 854
Pará	1 026 863	440 668	182 455	381 006
Amapá	49 127	10 887	10 492	25 657
Maranhão	1 308 102	884 472	110 110	285 293
Piauí	659 830	395 774	69 037	184 698
Ceará	1 715 066	741 215	320 869	608 156
Rio Grande do Norte ..	595 171	239 160	111 384	224 525
Paraíba	843 166	412 609	126 124	282 682
Pernambuco (2)	2 036 771	788 402	379 948	805 977
Alagoas	612 145	323 683	85 627	183 217
Sergipe	353 723	149 794	61 325	131 417
Bahia	3 034 291	1 464 985	477 862	984 108
Minas Gerais	4 736 190	1 518 442	1 115 624	1 991 950
Espírito Santo	710 605	242 241	151 578	302 327
Rio de Janeiro	4 317 373	195 580	1 241 157	2 755 915
São Paulo	10 411 726	1 175 002	3 998 442	5 062 567
Paraná	2 863 043	1 182 082	521 522	1 116 430
Santa Catarina	1 356 186	418 249	428 392	484 161
Rio Grande do Sul	3 204 117	903 641	812 608	1 437 897
Mato Grosso do Sul ..	502 921	176 126	87 289	232 352
Mato Grosso	384 826	162 318	63 494	148 917
Goiás	1 304 874	501 216	211 558	565 093
Distrito Federal	473 182	10 898	77 687	373 492

FONTE - IBGE. Censo Demográfico de 1980, dados gerais e mão-de-obra.
(1) Inclusive os que estavam procurando trabalho. (2) Inclusive Fernando de Noronha.

TABELA 24

INDICADORES SÓCIO-ECONÔMICOS, POR ZONAS DE OCUPAÇÃO
E UNIDADES DA FEDERAÇÃO - 1980

(continua)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	TAXA DE ALFABETIZAÇÃO DA POPULAÇÃO DE 5 ANOS E MAIS	PORCENTAGEM DA POPULAÇÃO DE 10 A 64 ANOS NA POPULAÇÃO DE 10 ANOS E MAIS	PORCENTAGEM DA PEA NA POPULAÇÃO DE 10 ANOS E MAIS
TOTAL	68,0	94,6	49,3
OCUPAÇÃO TOTAL	71,3	94,4	50,0
Ceará	47,7	93,8	45,6
Rio Grande do Norte	50,1	92,9	43,6
Paraíba	44,2	92,4	43,1
Pernambuco	49,9	93,6	46,5
Alagoas	39,8	93,8	45,0
Sergipe	47,2	93,0	44,8
Espírito Santo	70,3	94,8	47,9
Rio de Janeiro	82,0	94,0	48,6
São Paulo	82,1	94,7	53,9
Paraná	74,2	95,7	51,1
Santa Catarina	81,5	95,4	49,9
OCUPAÇÃO PARCIAL	75,0	94,6	49,8
Minas Gerais	69,7	94,6	47,7
Rio Grande do Sul	82,8	94,1	52,6
Distrito Federal	82,5	97,8	54,8
SUBOCUPADA	52,9	94,9	46,9
Rondônia	59,3	97,7	51,8
Acre	44,5	96,2	46,9
Pará	61,4	95,4	44,6
Maranhão	43,2	94,7	48,4
Piauí	43,4	94,5	45,2
Bahia	49,1	94,0	46,4
Mato Grosso do Sul	68,9	96,0	51,0
Goiás	63,8	96,0	47,3
VAZIA	60,6	96,4	47,7
Amazonas	58,8	96,3	47,3
Roraima	66,4	96,8	50,1
Amapá	65,9	95,5	43,3
Mato Grosso	61,5	96,6	48,6

TABELA 24

INDICADORES SÓCIO-ECONÔMICOS, POR ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO - 1980

(conclusão)

ZONAS DE OCUPAÇÃO E UNIDADES DA FEDERAÇÃO	TAXA DE ATIVIDADE (1)	PORCENTAGEM DA POPULAÇÃO ECONOMICAMENTE ATIVA (PEA) (2)		
		Setor Primário	Setor Secundário	Setor Terciário
TOTAL	52,1	30,0	25,4	44,6
OCUPAÇÃO TOTAL	53,0	23,2	29,4	47,4
Ceará	48,6	44,4	19,2	36,4
Rio Grande do Norte	47,0	41,6	19,4	39,0
Paraíba	46,7	50,2	15,4	34,4
Pernambuco	49,7	39,9	19,2	40,8
Alagoas	48,0	54,6	14,5	30,9
Sergipe	48,2	43,7	17,9	38,4
Espírito Santo	50,5	34,8	21,8	43,4
Rio de Janeiro	51,7	4,7	29,6	65,7
São Paulo	56,9	11,5	39,0	49,5
Paraná	53,4	41,9	18,5	39,6
Santa Catarina	52,4	31,4	32,2	36,4
OCUPAÇÃO PARCIAL	52,7	29,5	24,4	46,1
Minas Gerais	50,4	32,8	24,1	43,1
Rio Grande do Sul	56,0	28,6	25,8	45,6
Distrito Federal	56,0	2,4	16,8	80,8
SUBOCUPADA	49,4	50,7	14,8	34,5
Rondônia	53,1	52,8	13,6	33,6
Acre	48,7	52,5	8,8	38,7
Pará	46,8	43,9	18,2	37,9
Maranhão	51,1	69,1	8,6	22,3
Piauí	47,8	60,9	10,6	28,5
Bahia	49,3	50,1	16,3	33,6
Mato Grosso do Sul	53,1	35,5	17,6	46,9
Goiás	49,3	39,2	16,6	44,2
VAZIA	49,5	40,7	19,4	39,9
Amazonas	49,1	40,6	21,4	38,0
Roraima	51,7	36,8	13,9	49,3
Amapá	45,4	23,1	22,3	54,6
Mato Grosso	-	43,3	16,9	39,8

FONTE - IBGE. Censo Demográfico de 1980.

(1) Taxa de Atividade = $\frac{\text{PEA}}{\text{População 10 a 64 anos}} \times 100$ (2) Exclui os que estavam procurando trabalho.

RESUMO

Descreve e regionaliza a ocupação demográfica do Território Nacional entre 1960 e 1980 em quatro zonas de ocupação demográficas: total, subocupada, parcial e vazia. Para cada zona demográfica são descritos os principais aspectos demoeconômicos e da concentração populacional no país nesse período.

ABSTRACT

It describes the demographic occupation of the national territory, between 1960 and 1980, taking into account four areas according to their population and area: totally occupied, suboccupied, partially occupied and unoccupied. For each occupied area it describes main economic and demographic features and observes the populational redistribution occurred in Brazil during that period. Federação.

NOTAS

1 - Este trabalho foi elaborado e redigido em 1982 e atualizado com dados definitivos do Censo Demográfico de 1980 ém 1986.

2 - Cabe mencionar a contribuição dada pela Socióloga Leda Maria N. de Oliveira que participou da realização de algumas fases deste projeto. Agradece também as sugestões de Joop Alberts, Manoel Augusto Costa e Luiz Antonio Pinto de Oliveira. Ressalta-se que os possíveis erros estampados neste trabalho como as idéias desenvolvidas são de inteira responsabilidade do autor.

3 - MEÓT, Henry - El concepto de región, segunda parte, aspectos metodológicos - Santiago; CEPAL-1974 (Curso de Planificación del desarrolho, doc.c/25-B)

4 - Uma forma que melhor se ajustaria à realidade brasileira seria aplicar a técnica para microrregiões ou municípios por grandes regiões, pois permitiria descrever e analisar com mais detalhes as condições de ocupação demográfica, tanto para o país como para o interior deste.

5 - Chama-se atenção de que a variabilidade e homogeneidade aqui tratadas não se referem a cada unidade de análise e, sim, ao número total de unidade de observação (N_T).

6 - Recordando que a $\sum_{i=1}^{n} I_i$ é igual a 1 ou100%.

7 - MEÓT, Henry; DOMICELJ, Sergio - Cambio estrategico en la ocupación del território algunas cuestiones derivadas de lo experiencia peruana. In: PLANIFICACIÒN REGIONAL Y URBANA EN AMERICA LATINA, México, Editores ILPES/ILDIS - Siglo XXI, 1978.

8 - Exclui-se o território de Fernando de Noronha por falta de representatividade territorial e populacional no contexto nacional para 1960 e 1970. Em 1980, encontra-se agregado ao Estado de Pernambuco

9 - MEÓT; COMICELJ. Op. cit., p. 135 - acrescentam mais dois aspectos que não analisaremos neste trabalho que são o subaproveitamento dos recursos minerais e os custos crescentes das grandes cidades, que requerem alocações da maioria dos escassos recursos disponíveis

10 - Cabe lembrar que os graus relativos de concentração urbana correspondem à soma dos indicadores de ocupação demográfica.

11 - Fórmula do índice de Gini para dados agrupados:

$$G = \frac{\sum_{i=1}(X_{i-1} \cdot Y_i) - \sum_{i=1}(X_i \cdot Y_{i-1})}{10000}$$

X_{i-1} = representa o percentual acumulado da população residente urbanos de 10 000 e mais habitantes por tamanho dos centros urbanos.

$X_{i=1}$ = representa o percentual acumulado da população residente em centros urbanos de 10 000 e mais habitantes na classe anterior.

Y_i = representa o percentual acumulado do número de centros urbanos de 10 000 e mais habitantes.

$Y_{i=1}$ = representa o percentual acumulado do número de centros urbanos de 10 000 e mais habitantes na classe anterior.

Fórmula do índice Similaridade:

$$\delta = \sum_{i=1} \left| \frac{x_i - y_i}{2} \right| \rightarrow \quad \text{onde:}$$

x_i = percentual da população residente em centros urbanos de 10 000 e mais habitantes.

y_i = percentual do número de centros urbanos de 10 000 e mais habitantes.

12 - Neste trabalho somente analisaremos a dispersão da população urbana, dado que o objetivo é ver como está se comportando o processo de urbanização ao longo dos últimos 20 anos, principalmente em relação aos centros urbanos de 10 000 e mais habitantes. Por outro lado, além das zonas de ocupação considerou-se importante apresentar as medidas segundo as Unidades da Federação.

13 - Ver Tabelas 11, 12, 13, 14 e 15 no final do trabalho.

14 - O conceito de Força de Trabalho, aqui utilizado, refere-se à taxa de atividade (% da PEA por setores de atividade na PEA de 10 anos e mais).

15 - População urbana residente na sede municipal

16 - Idem, idem.

RESENHAS BIBLIOGRÁFICAS

- GENERALIZED LINEAR MODELS por P. McCullagh e J. A. Nelder. 1a. Edição. Chapman and Hall. 1983. 261p.

O livro de McCullagh e Nelder, publicado em 1983, propiciou um grande avanço nas pesquisas e parte aplicada dos modelos lineares generalizados. O livro apresenta um resumo bastante consistente da teoria dos modelos lineares generalizados (MLGs) com ênfase na construção de modelos para a análise de dados univariados. Neste texto são descritas quase todas as áreas de trabalho com os MLGs e citadas as principais pesquisas publicadas até o ano de 1982.

O texto é dividido em 12 capítulos. O 1º é uma introdução a vários modelos estatísticos, como o modelo clássico de regressão, o modelo log-linear, o modelo logístico para estudo de proporções, polinômios inversos e modelos para análise de dados de sobrevivência. Os MLGs são introduzidos no Capítulo 2, onde se destacam as medidas de adequação e análise dos resíduos. No Capítulo 3, são estudados os modelos para análise de dados contínuos com variância constante, dando ênfase à formação da estrutura do modelo, a estimação dos parâmetros e a seleção de covariáveis.

No Capítulo 4, discute-se a análise de dados binários com alguns tópicos de teoria assintótica

e verossimilhanças condicionadas, enquanto no capítulo 5, os modelos multinomiais são descritos com a apresentação da função de quase-verossimilhança de vários modelos de chances proporcionais e, mais fundamentalmente, definindo modelos em termos da escala de medição.

Para mim, os Capítulos 6 e 7 são os melhores do texto. No capítulo 6, os autores apresentam os modelos log-lineares de uma forma bastante simples, com uma notação elegante, o que não é comum nos livros que tratam desses modelos, sendo feitas duas análises de dados reais que ilustram muito bem as potencialidades dos modelos para análise de dados na forma de contagens. No capítulo 7, os modelos para análise de dados com coeficiente de variação constante são revistos com três bons exemplos de análise de dados.

O Capítulo 8 é um resumo do artigo do McCullagh "Quasi-likelihood functions" no *Annals of Statistics* em 1983, sendo o mais difícil de ser entendido, além de apresentar alguns erros, principalmente, em relação à função de quase-verossimilhança para dados correlacionados. A meu ver, este capítulo deveria ser o penúltimo do texto.

Os Capítulos 9 e 10 tratam, respectivamente, de modelos para a análise de dados de sobrevivência e de modelos com parâmetros não-lineares adicionais. O Capítulo 11, o 3° melhor do livro, apresenta, de uma forma bastante clara e com uma visão prática, as principais técnicas de diagnóstico nos MLGs. As formas apresentadas para verificar a adequação da estrutura linear da função de ligação e da função de variância são guias bastante úteis para o analista na prática.

O último Capítulo, o 12, tenta apresentar alguns tópicos adicionais para pesquisas futuras.

O texto apresenta um pequeno número de erros que são facilmente localizados pelo leitor mais atento. Acreditamos que ele deva ser recomendado para alunos de pós-graduação que tenham noções de teoria da verossimilhança. Admite-se ainda que o leitor já esteja familiarizado com o pacote GLIM (Generalized Linear Interactive Modelling).

O livro do McCullagh e Nelder representou o 1° passo na análise de dados segundo a teoria dos MLGs. Depois desse livro, vários outros já foram publicados, mas certamente este texto continua a ser o mais importante nessa área. Após 1983, as pesquisas na área dos MLGs cresceram a passos agigantados e o texto foi ficando um pouco desatualizado. Tanto assim que seus autores lançarão uma 2a. edição, provavelmente no início de 1990, com muito mais capítulos tentando manter o livro a par das pesquisas correntes na área.

Finalmente, pelo papel que teve no desenvolvimento das pesquisas na área dos MLGs, e por ter sido o 1° livro publicado nessa área, o texto de McCullagh e Nelder tornou-se um best-seller na literatura Estatística, sendo essencial sua consulta a pessoas interessadas na análise de dados.

Gauss M. Cordeiro (Depto. de Economia, CCSA/UFPE, 5070 - Pe)

- BAYESIAN ANALYSIS OF LINEAR MODELS, Lyle D. Broemeling, Marcel Dekker Inc., N. York and Basel. 1985. 454p.

Este é um livro de enfoque Bayesiano, como deixa claro o próprio título. Logo no prefácio o autor coloca a pergunta/provocativa: "Why the Bayesian approach?", e responde de forma simples e indiscutível: "Because one may solve the inferential problems in the analysis of linear models with one tool, namely Bayes theorem."

Uma resposta mais completa a essa questão acabaria por envolver tecnicismos que não cabem nessa resenha. Entretanto, além da simplicidade (um único teorema), destacamos a crescente aplicação do método, como pode ser constatado nas revistas e jornais técnicos, bem como pela quantidade de novos textos publicados na área. A título de ilustração citamos a aposta feita por Good em 1947 com Bartlett a cerca de qual corrente da estatística predominaria ao fim de um século. No artigo mencionado ele conclui que passado um terço de século a "tendência" é plenamente favorável a corrente Bayesiana, a considerar o número de novos textos e publicações técnicas.

É com base nesse pano de fundo que nos propusemos a essa resenha. Embora sendo um texto da série "Statistic Text Books and Monographs " tem algumas falhas tipográficas grosseiras, devendo ser examinado muito mais como um livro de consulta do que como um "livro - texto"

Seu principal mérito é reunir num mesmo texto os resultados mais recentes em modelos lineares Bayesianos.

O livro está organizado em nove capítulos e um apêndice contendo as principais distribuições uni e multivariadas (por exemplo: Wishart, t-multivariada, poli-t e matriz t).

Os capítulos são denominados: Bayesian Inference for the General Linear Model, Linear Statistical Models and Bayesian Inference, The Traditional Linear Models, The Mixed Models, Time Series Models, Linear Dynamic Systems, Structural Change in Linear Models, Multivariate Linear Models, Looking Ahead e Appendix.

Descrevemos a seguir o conteúdo de cada um desses capítulos destacando os pontos relevantes.

No capítulo 1 os conceitos fundamentais de inferência Bayesiana são apresentados no contexto de modelos lineares, o que pressupõe conhecimentos prévios do modelo de regressão e de métodos Bayesianos, ao nível, por exemplo, de um curso de mestrado, aliás, como é destacado no prefácio do texto à página (iv). Outro ponto de destaque neste capítulo é uma seção com comentários sobre referências bibliográficas recentes.

Na primeira parte do capítulo 2 são apresentados diversos modelos especiais de: regressão, planejamento de experimento, mudanças estruturais e outros; enquanto que na segunda parte,

I. J.Good (1980), "Some History of the Hierarchical Bayesian Methodology", in J.M. Bernardo et alii, Ed., Bayesian Statistics 1, Spain, pg 484

o autor reapresenta de forma mais formal vários conceitos próprios da Inferência Bayesiana incluindo seções sobre: probabilidade subjetiva, prioris vagas, distribuição preditiva etc. À página 40 encontramos uma nota histórica sobre os processos Bayesianos. Ao fim deste capítulo é apresentada uma lista de referências bibliográficas atualizada até meados da década de 80.

Os modelos lineares tradicionais são discutidos no capítulo 3. Inicialmente apresentam-se os modelos normais para uma e duas populações, e ilustra-se o uso do método Bayesiano com análise das distribuições a priori, a posteriori e preditiva. A seguir, os modelos de regressão linear simples e múltipla são analisados, além de serem introduzidos modelos não-lineares de regressão. Alguns modelos de planejamento de experimento são também discutidos, incluindo a análise de modelos de dois fatores de classificação, blocos aleatorizados e experimentos com interação. A última seção deste capítulo discorre sobre análise de covariância.

O capítulo 4 abrange os modelos mistos, mais gerais que aqueles de efeito fixo, pois incluem fatores aleatórios. Estas são variáveis aleatórias não observáveis e suas variâncias – componentes de variância – são os parâmetros elementares de interesse.

Modelos de séries temporais são analisados do ponto vista Bayesiano, no capítulo 5. Destacam-se os modelos auto-regressivos de média móvel e de defasagem polinomial. Este capítulo estende os trabalhos de Zellner, A – "Bayesian Inference in Econometrics". Alguns aspectos de identificação são apresentados superficialmente. Por último, destacamos um estudo numérico de um processo auto-regressivo quase estacionário. Este capítulo inclui ainda o modelo de regressão linear com estrutura auto-regressiva.

O objeto do capítulo 6 são os sistemas lineares dinâmicos, os quais se prestam ao monitoramento e ao controle de estados de um sistema. Os principais desenvolvimentos teóricos e aplicados nesta área ocorreram em engenharia de controle e comunicações. Uma exceção na literatura apresentada é o artigo de Harrison & Stevens Os principais aspectos da análise de sistemas lineares são: filtragem, predição e suavização. A natureza seqüencial do método Bayesiano adapta-se perfeitamente aos modelos lineares dinâmicos, tendo papel relevante na solução de problemas de estimação, identificação e controle de sistemas. Este capítulo introduz as idéias básicas de sistemas lineares, revisa o Filtro de Kalman e desenvolve a análise Bayesiana para controle, estimação adaptativa e filtragem não-linear.

No capítulo 7, modelos lineares com mudanças estruturais são amplamente discutidos. Suas aplicações em Economia, Ciências Sociais, Física etc. são bastantes conhecidas. O capítulo descreve modelos de deslocamento em seqüências normais, mudanças estruturais em modelos lineares, estabilidade estrutural em modelos de regressão com erros auto-regressivos.

O conteúdo do capítulo 8 inclui o estudo de alguns modelos lineares multivariados de regressão e de planejamento de experimento. Modelos vetoriais auto-regressivos são também discutidos. Sob o título de "outros modelos multivariados" encontram-se tópicos tais como: regressão multivariada com erro auto-regressivo, função de transferência e modelos multivari-

P.J.Harrison and C.F.Stevens (1976) - "Bayesian Forecasting" - Journal Royal Statistical Soc., series B,38, 205-247.

ados para mudanças estruturais.

Finalmente, no capítulo 9, é apresentada uma "visão futura da inferência Bayesiana em modelos lineares". É feita uma revisão da análise Bayesiana em modelos lineares e são apresentados tópicos para pesquisas futuras incluindo estudos de aproximações para a distribuição poli-t em sistemas lineares dinâmicos e em modelagem de mudanças estruturais e suas relações com outliers.

Helio S. Migon (ENCE/IBGE & IM-UFRJ)

CORREÇÃO

No artigo intitulado "Crítica de Razões no Censo Econômico", publicado no n.191 (1988/1º semestre), foi omitido, por um lapso, o nome da co-autora, Rosana de Freitas Castro.

PUBLICAÇÕES RECEBIDAS

A Gerência de Editoração do IBGE recebeu as seguintes publicações:

Consul, P.C. *Generalized Poisson distributions, properties and applications.* Marcel Dekker, Inc. New York, Marcel Dekker, 1989, 302p.

Dielman, Terry E. *Pooled Cross-sectional and time series data analysis.* Marcel Dekker, Inc. New York,1989, 249p

Vallin, Jacques; Meslé, Franco. *Les causes de décès en France de 1925 a 1978.* Paris, Institut National d' Études Démographiques, Presses Universitaries de France, 1988, 607p.

Population & Societes. Paris, Institut Nacional d' Études Demographiques, nº 230, dec. 1988

Population & Societes. Paris, Institut Nacional d' Études Demographiques, nº 231, 232, 234, 236, jan., fev/mar., abr./maio, jun. 1989

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBE's

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

Djalma G.C.Pessoa
Editor Responsável - RBEs
ENCE
Rua André Cavalcanti, l06
Bairro de Fátima
20231 - Rio de Janeiro - RJ

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

– Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

Instruções para preparo de originais:

1. 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 sequencialmente,

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) nomes(s) completo(s) do(s) autor(es), indicando-se para 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, 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 $\xrightarrow{D}$ 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. Refências 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 página do artigo, entidade que publicou, e localidade onde foi publicado. Na lista final das referências, todos os auotres 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, nº 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 apresentadas em folhas separadas, precedidas de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas sequencialmente (Tabela 1, Figura 3, etc...) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. O apêndices devem ter título 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 parentênses 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 $\hat{\hat{x}}$, ou $\boldsymbol{x}$ ou $(x+y)$ 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 (dabliu) e ω (omega minúsculo)

v (vê) e ν (ni minúsculo)

o (o minúsculo), O (o maiúsculo) e 0 (zero)

1 (um) e l (ele minúsculo);

h) procure usar a notação indicada nos exemplos a seguir

Use	Não Use
$\mathrm{var}(x)$	var$x$ ou VAR(x)
cov	Cov ou COV
pr para probabilidade	P ou Pr
tr para traço de matriz	Traço ou traço
$E(X)$para esperança de$X$	EX ou $\mathcal{E}(X)$
$\log(x)$	$\log_e(x)$ ou $ln(x)$
$a/(bc)$ ou $a(bc)^{-1}$	a/bc
$0,2$	0.2 ou .2 ou ,2
1.000.00	1000 ou 1000,00
$x_1, \cdots, x_n$	$x_1, x_2, \cdots, x_n$
$\Gamma\left(\frac{1}{2}n+\frac{1}{2}\right)$	$\Gamma[(n+1)/2]$ ou $\Gamma\binom{n+1}{2}$
ab	$a \cdots b$ ou $a \times b$

9. Equações devem se 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 $\Sigma_{i=1}^{n} x_i$; não deve ser deixada no texto, ou então deve ser escrita Σx_i (se os limites de somação forem óbvios). A forma Σ_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 qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho quer nas legendas ou títulos.

Impresso no Departamento de Produção Gráfica – IBGE, em maio de 1990 – O.S. 281-10

ENTRE EM CONTATO COM O IBGE
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

SEDE — Presidência
Av. Franklin Roosevelt, 166
20021 — RIO DE JANEIRO — RJ — Telefone: 220-9442

CDDI — Centro de Documentação e Disseminação de Informações
Av. Beira Mar, 436
20021 — RIO DE JANEIRO — RJ — Telefone: 220-0511

BRASÍLIA
SCS — Quadra 06 — Bloco "A"
70300 — BRASÍLIA — DF — Telefone: 224-6998

UNIDADES REGIONAIS NAS CAPITAIS
ENDEREÇO E TELEFONE

AC/RIO BRANCO — Rua Benjamin Constant, 506, Tel.: 224-1540
AL/MACEIÓ — Rua Tibúrcio Valeriano, 125, Tel.: 223-5088
AM/MANAUS — *Rua Lobo D'Almada, 272, Tel.: 232-0152*
AP/MACAPÁ — Av. Antonio Coelho de Carvalho, 301, Tel.: 222-2796
BA/SALVADOR — Av. Estados Unidos, 50, Tel.: 243-9277
CE/FORTALEZA — Rua Major Facundo, 733, Tel.: 231-5352
ES/VITÓRIA — Rua Duque de Caxias, 267, Tel.: 222-5004
GO/GOIÂNIA — Av. Tocantins, 675, Tel.: 223-3307
MA/SÃO LUÍS — Rua Joaquim Távora, 49, Tel.: 222-0350
MT/CUIABÁ — Av. XV de Novembro, 235, Tel.: 322-2121
MS/CAMPO GRANDE — Rua Barão do Rio Branco, 1431, Tel.: 721-1902
MG/BELO HORIZONTE — Rua Oliveira, 523, Tel.: 223-0554
PA/BELÉM — Av. Gentil Bittencourt, 418, Tel.: 222-7195
PE/RECIFE — Rua do Hospício, 387, Tel.: 231-0811
PB/JOÃO PESSOA — Rua Irineu Pinto, 94, Tel.: 241-1560
PI/TERESINA — Rua Simplício Mendes, 436, Tel.: 222-4161
PR/CURITIBA — Rua Carlos de Carvalho, 552, Tel.: 234-9122
RJ/RIO DE JANEIRO — Rua Humaitá, 85, Tel.: 286-2672
RN/NATAL — Praça Pedro Velho, 435, Tel.: 222-3695
RO/PORTO VELHO — Av. Duque de Caxias, 1223, Tel.: 221-5143
RR/BOA VISTA — Av. Getúlio Vargas, 76-E, Tel.: 224-4425
RS/PORTO ALEGRE — Av. Augusto de Carvalho, 1205, Tel.: 228-6444
SC/FLORIANÓPOLIS — Rua João Pinto, 12, Tel.: 222-0733
SE/ARACAJU — Rua Riachuelo, 1017, Tel.: 222-8197
SP/SÃO PAULO — Rua Urussuí, 93, Tel.: 883-0077