Imputação de valores faltantes referentes às variáveis de rendimento do trabalho na Pesquisa Mensal de Emprego

Instituto Brasileiro de
Geografia
e Estatística - IBGE

APRESENTAÇÃO

A Fundação Instituto Brasileiro de Geografia e Estatística – IBGE apresenta, nesta publicação, a descrição e os resultados do método de imputação de valores faltantes, referentes às variáveis de rendimento de trabalho, aplicado na Pesquisa Mensal de Emprego – PME.

Com esta publicação documenta-se a aplicação deste método que aprimora os resultados dos indicadores de rendimento da Pesquisa Mensal de Emprego.

Esses Indicadores, com a aplicação do procedimento de imputação, passaram a ser divulgados a partir dos resultados da PME de março de 2007.

IBGE coloca à disposição do usuário, além do presente documento, um texto resumo descrevendo o impacto da aplicação da metodologia nos principais indicadores da pesquisa, a série histórica completa recalculada e os microdados.

Wasmália Socorro Barata Bivar
Diretora de Pesquisas

26 de abril de 2007

1. INTRODUÇÃO:

Entre os erros não-amostrais a que uma pesquisa está sujeita, a não-resposta é um dos que sempre se verificam, em maior ou menor grau, seja por dificuldade de estabelecer contato com o informante, seja por dificuldade do respondente em oferecer as informações solicitadas.

Tais dificuldades são mais percebidas em pesquisas repetidas mensalmente, como é caso da Pesquisa Mensal de Emprego, cujo cronograma de coleta requer entrevistas num calendário apertado. Ademais, o rendimento , que é visto como uma informação confidencial sensível, é um dos quesitos mais vulneráveis à não-resposta, mesmo quando o restante da pesquisa é realizado com sucesso. Este comportamento também é observado em pesquisas censitárias e por amostragem, tanto no Brasil como internacionalmente.

Além disso, a não-resposta de rendimentos tende a ocorrer com mais freqüência em todas as investigações quando esses são mais altos, o que, também, se verifica na Pesquisa Mensal de Empregos. Esse comportamento é identificado como não-resposta diferencial e um tratamento estatístico é requerido para correção desse vício das estimativas produzidas. Optou-se, dessa forma, por adotar o procedimento de imputação.

Este documento apresenta a descrição da metodologia utilizada na imputação[1] dos valores faltantes referentes às variáveis de rendimentos de trabalho captados na Pesquisa Mensal de Emprego - PME.

Este procedimento é mais uma etapa do processo de crítica e imputação das variáveis investigadas pela pesquisa, que se inicia com a detecção e

[1] **Definição** : Substituição do dado faltante, ou inconsistente (coletado) por valor definido em escritório (não coletado);

Justificativas :
- facilita análise e processamento de dados ;
- maior consistência nas estimativas;
- pesos amostrais: uma pessoa pode representar outras 200 ou 300.

correção automática das inconsistências verificadas na parte 3 do questionário - Características de Educação dos Moradores de 10 anos ou mais e na parte 4 do questionário - Características de Trabalho dos Moradores de 10 anos, utilizando o sistema DIA[2].

Com base nas informações previamente criticadas pelo DIA, os seis quesitos que investigam o rendimento mensal dos trabalhadores são analisados. São eles:

1. Rendimento bruto mensal habitual do trabalho principal dos empregados e trabalhadores domésticos;

2. Rendimento bruto mensal efetivo do trabalho principal dos empregados e trabalhadores domésticos, recebido no mês de referência;

3. Retirada mensal habitual do trabalho principal dos trabalhadores por conta própria e empregadores;

4. Retirada mensal efetiva do trabalho principal dos trabalhadores por conta própria e empregadores, no mês de referência;

5. Rendimento mensal habitual dos trabalhadores no(s) outro(s) trabalho(s) que tinha(m) na semana de referência; e

6. Rendimento mensal efetivo dos trabalhadores no(s) outro(s) trabalho(s) que tinha(m) na semana de referência, recebido no mês de referência.

Aqueles que não são declarados pelo informante são submetidos a um processo de imputação, utilizando uma metodologia baseada em Árvores de Regressão[3] .

[2] DIA - Aplicativo de informática desenvolvido pelo Instituto Nacional de Estatística -INE- da Espanha, para crítica e imputação de dados qualitativos. Baseia-se na metodologia de Fellegi Holt, com certas modificações para tratar erros sistemáticos.

[3] A mesma metodologia foi utilizada para imputação dos quesitos de renda do Censo Demográfico 2000, porém com algumas diferenças de ordem operacional.

Para facilitar a implementação do processo de imputação, foram criadas novas variáveis de rendimento a partir do conjunto de quesitos do questionário. São elas:

a) Rendimento habitual do trabalho principal (composto pelos itens 1 e 3 anteriores);

b) Rendimento efetivo do trabalho principal (composto pelos itens 2 e 4);

c) Rendimento habitual do(s) outro(s) trabalho(s) (item 5); e

d) Rendimento efetivo do(s) outro(s) trabalho(s) (item 6).

Cabe ressaltar a necessidade de que o processo de imputação das variáveis de rendimento seja extremamente ágil ou o mais automatizado possível, assim como todas as outras etapas de apuração da pesquisa. Isto visa garantir que o tempo que separa o término da operação de entrevista e a divulgação dos resultados da pesquisa (cerca de sete dias úteis) seja preenchido prioritariamente com a análise dos resultados obtidos e elaboração dos relatórios de divulgação.

2. ASPECTOS GERAIS:

Atualmente, a PME investiga cerca de 40.000 domicílios nas 6 Regiões Metropolitanas abrangidas pela pesquisa, dos quais aproximadamente 80% resultam em entrevistas realizadas, que resultam em cerca de 43.000 trabalhadores remunerados, que, conseqüentemente, deveriam ter seus rendimentos informados.

Entretanto, é conhecida a dificuldade de se obter esses dados do informante. Cabe ressaltar que, os esforços que têm sido direcionados para melhorar a coleta estão se refletindo na sensível redução das taxas de não-resposta nos quesitos de rendimento, desde a implantação da PME reformulada, em março de 2002, como pode ser verificado na Tabela 1.

Tabela 1: Taxa de não-resposta dos quesitos de rendimento[1] do trabalho principal da PME.

Região Metropolitana	Média histórica [2]	Média de 2002	Média de 2003	Média de 2004	Média de 2005	Média de 2006
Total	5,6%	8,0%	9,0%	6,2%	2,9%	2,3%
Recife	7,7%	10,1%	12,9%	11,9%	2,4%	1,5%
Salvador	6,2%	7,6%	10,5%	8,2%	3,1%	1,8%
Belo Horizonte	3,9%	8,0%	7,1%	2,6%	1,2%	1,2%
Rio de Janeiro	8,3%	12,9%	10,6%	8,3%	5,7%	4,9%
São Paulo	3,7%	4,7%	6,1%	4,0%	2,1%	1,7%
Porto Alegre	4,7%	5,6%	8,8%	4,7%	2,4%	2,3%

(1) Incluindo habitual e efetivo.

(2) Média calculada de março de 2002 a dezembro de 2006, sobre o total de remunerados.

Entre as possibilidades de não-resposta estão incluídas a não-resposta total dos quesitos de rendimento, ou seja, aquela em que não foram informados nem o rendimento habitual nem o efetivo, e a não-resposta parcial, em que um dos rendimentos foi declarado.

O comportamento histórico destas taxas de não-resposta é a justificativa para o processo de imputação de rendimentos. A taxa de não-resposta de rendimento é diferencial, ou seja, não é ao acaso, fato comprovado por uma participação maior das pessoas ocupadas com rendimento ignorado entre as mais escolarizadas e as que foram classificadas como empregadoras. As Tabelas 2 e 3, apresentam uma média histórica, de março de 2002 a dezembro de 2006, da distribuição por anos de estudo e por posição na ocupação, respectivamente, da população ocupada remunerada com rendimentos ignorados no trabalho principal e da população ocupada como um todo, nas seis regiões metropolitanas investigadas pela pesquisa. É importante ressaltar que embora seja uma média do período em questão, o mesmo comportamento é observado para cada mês individualmente.

Tabela 2: Distribuição da população ocupada total e ocupada com rendimento ignorado segundo as faixas de anos de estudo.

Anos de estudo	População Ocupada com rendimentos ignorados (%)	População Ocupada (%)
Sem instrução e menos de 1 ano	2,2	2,8
1 a 3 anos	4,2	6,2
4 a 7 anos	16,8	24,9
8 a 10 anos	14,6	19,0
11 anos ou mais	62,2	47,1
Total	100,0	100,0

Tabela 3: Distribuição da população ocupada total e ocupada com rendimento ignorado segundo a posição na ocupação.

Posição na Ocupação	População Ocupada com rendimentos ignorados (%)	População Ocupada (%)
Empregado Doméstico	3,3	8,4
Militar e Funcionário Público	14,4	7,6
Empregado com carteira	28,1	42,0
Empregado sem carteira	19,1	16,6
Conta Própria	26,4	20,4
Empregador	8,7	5,0
Total	100,0	100,0

Nota-se uma maior participação de pessoas com 11 anos ou mais de estudo entre os que não possuem rendimentos declarados, em confronto com a população ocupada total. O mesmo comportamento é observado entre os empregadores. Já observando os empregados domésticos e os trabalhadores com menores níveis de escolaridade, o que se nota é uma menor participação entre os que não possuem rendimentos declarados do que na população ocupada total. Este é um comportamento observado em todas as regiões metropolitanas.

3. METODOLOGIA:

Como dito anteriormente, a metodologia utilizada combina Árvores de Regressão com seleção probabilística de doadores em cada estrato construído através da árvore (Breiman *et al*, 1984). Em linhas gerais, a técnica de árvore de regressão consiste em um método de estratificação que utiliza um conjunto

de características das pessoas respondentes da PME para classificar os registros em grupos homogêneos, a partir de um grupo de variáveis explicativas. Para tal procedimento foi utilizado a função RPART do software R.[4]

A formação da árvore se dá através de partições binárias, sempre distribuindo os indivíduos em dois grupos mutuamente exclusivos, que são chamados de nós. O grupo inicial que contém todos os indivíduos é chamado de nó raiz e os estratos finais, de nós terminais. Estes formam as classes de imputação.

A cada mês, é construída uma árvore para cada uma das seis regiões metropolitanas investigadas pela PME (Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo e Porto Alegre). A variável utilizada como variável dependente do modelo é o rendimento habitual do trabalho principal.

As variáveis explicativas selecionadas para a construção da árvore foram: sexo, idade, condição do morador no domicílio, anos de estudo, posição na ocupação no trabalho principal, tamanho do empreendimento do trabalho principal e horas habitualmente trabalhadas no trabalho principal. As variáveis sexo, condição do morador no domicílio, posição na ocupação e tamanho do empreendimento, foram divididas nas seguintes categorias:

- Sexo(sexo):

a = homem
b = mulher

- Condição do morador no domicílio (cond):

a = principal responsável
b = outros

- Posição na ocupação no trabalho principal (Pos_ocup):

a = empregado doméstico
b = militar e funcionário público
c = empregado com carteira
d = empregado sem carteira
e = conta própria

[4] RPART - Recursive Partitioning. É um função do software R que trabalha tanto com árvores de regressão quanto de classificação . O R é um software livre e pode ser obtido através do endereço http://www.R-project.org.

f = empregador

- Tamanho do empreendimento do trabalho principal (Tam_empr):

a = 2 a 5 pessoas
b = 6 a 10 pessoas
c = 11 ou mais

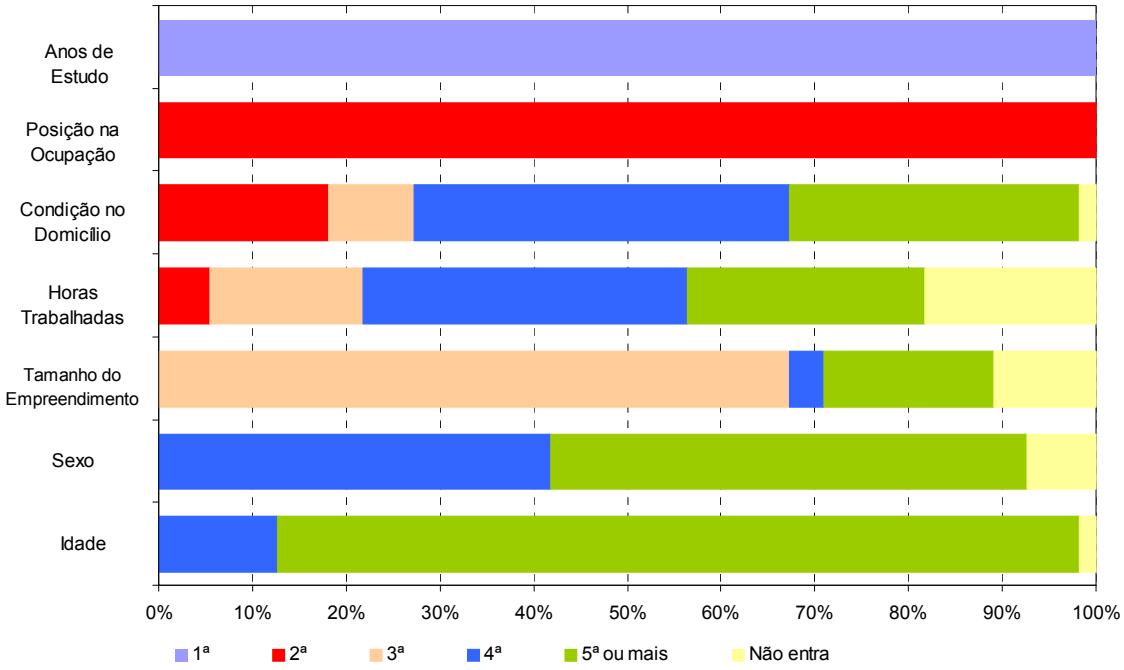
As demais variáveis são quantitativas e foram usadas diretamente.

As melhores partições são estabelecidas pela metodologia de geração das árvores, utilizando ou não todas as variáveis apontadas como preditoras, inclusive permitindo a reutilização das variáveis, de forma que, ao final, se obtenha a menor variância possível dentro de cada nó.

Para a construção das árvores adotou-se o critério de que um nó terminal não poderia ter menos do que 1% do total de registros utilizados para sua construção (afim de se obter um número razoável de doadores), e foi definido um valor para o parâmetro de complexidade que fornece árvores com 20 nós terminais em média (número utilizado também no Censo Demográfico).

A Figura 1 apresenta a participação da posição de entrada das variáveis explicativas (levando em conta apenas a primeira aparição) na construção da árvore de regressão, considerando as árvores construídas para a Região Metropolitana de São Paulo, no período de março de 2002 a dezembro de 2006 (58 árvores).

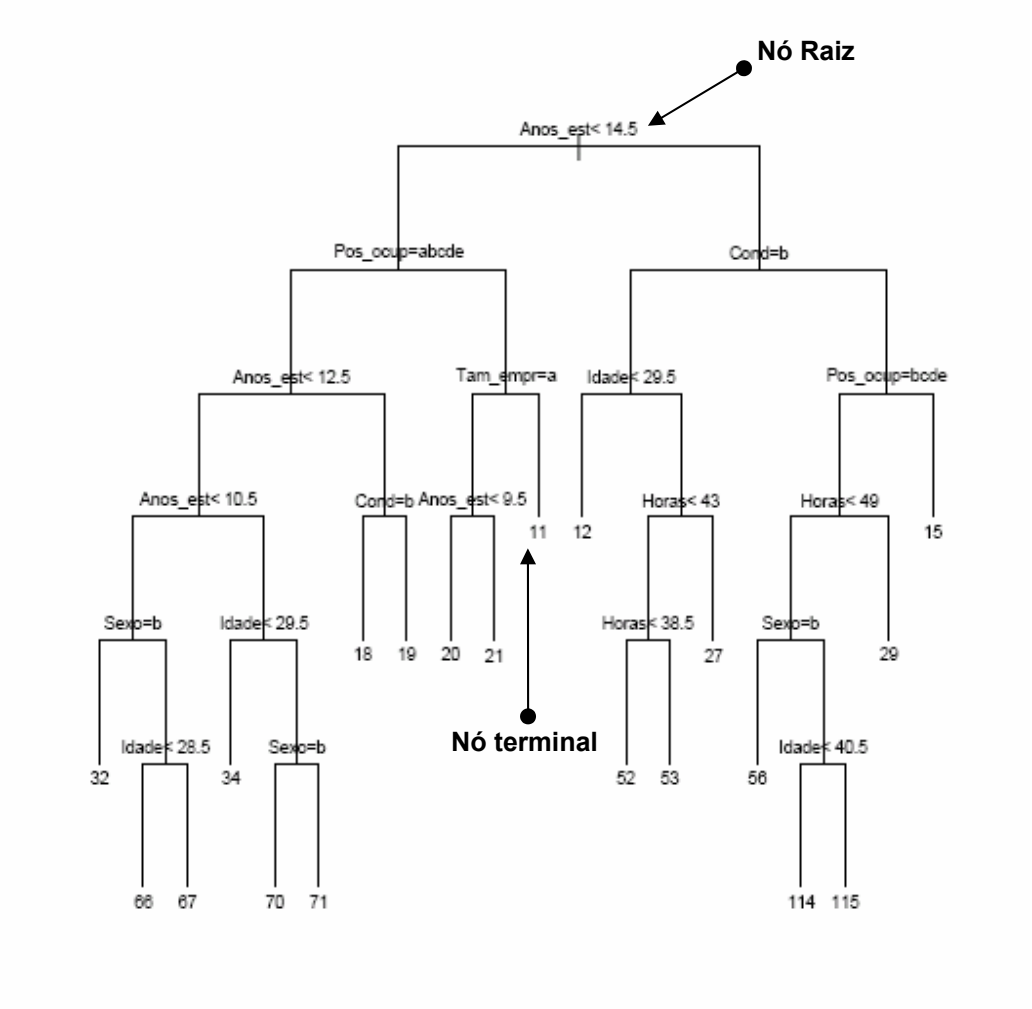
Figura 1: Posições de entrada das variáveis explicativas na construção da árvores.



Desta maneira nota-se que, em todas as árvores construídas da árvore, a variável ‘Anos de estudo’ foi a primeira a ser utilizada para a partição do nó raiz e que também a variável ‘Posição na ocupação’ apareceu na segunda partição.

A Figura 2 apresenta um exemplo de árvore obtida. Trata-se de uma árvore binária onde, a cada passo, o registro é classificado como tendo ou não a característica. Primeiro é verificado se a variável 'Anos de estudo' é menor ou igual a 14,5 anos. Em caso afirmativo, segue-se o caminho da esquerda. Se a variável 'Anos de estudo' for maior que 14,5 anos, segue-se o da direita e assim sucessivamente até os nós terminais.

Figura 2: Árvore construída para a Região Metropolitana de São Paulo em dezembro de 2005.



Neste exemplo o nó de número 11 é um nó terminal e nele estão todas as pessoas que têm menos de 14,5 anos de estudo, são empregadores e estão em empreendimentos com mais de 6 funcionários.

4. PROCEDIMENTO DE IMPUTAÇÃO:

O procedimento de imputação é diferenciado de acordo com o tipo de não-resposta observado. No caso de rendimentos habituais e efetivos ignorados a imputação é feita pela seleção de doadores nos nós terminais da árvore construída para a imputação. É um procedimento do tipo *hot-deck*[5] com seleção aleatória dentro de classes, ou seja, em cada nó seleciona-se aleatoriamente, com probabilidade proporcional ao peso de cada indivíduo, um doador para os rendimentos ignorados. Estes doadores são aqueles que possuem rendimentos habituais e efetivos não ignorados, excluindo em cada nó aqueles que apresentam valores extremos (superiores e inferiores). A obtenção destes limites é função do intervalo interquartílico, calculado na escala logarítmica, ou seja:

$LI = Q1 - 2{,}0(Q3-Q1)$ e

$LS = Q1 + 2{,}0(Q3-Q1)$,

onde LI e LS são os limites inferiores e superiores, respectivamente e Q1 e Q3 são o primeiro e terceiro quartis. A quantidade Q3-Q1 é a distância interquartílica.

Desta forma, se uma pessoa possui um vetor de 2 rendimentos não respondidos, rendimento habitual do trabalho principal e rendimento efetivo do trabalho principal, o doador irá ceder seus dois rendimentos a este receptor. O mesmo procedimento é adotado para aqueles que têm apenas os rendimentos do(s) outro(s) trabalho(s) ignorados e para aqueles que possuem um vetor de 4 rendimentos ignorados (dois do trabalho principal e dois do(s) outro(s) trabalho(s)).

Caso não haja doadores para algum destes casos dentro de seu nó correspondente, a seleção é feita no nó raiz tentando-se preservar como doadores aqueles que possuem a mesma posição na ocupação no trabalho principal do receptor. Se ainda assim não forem encontrados doadores, a

[5] Procedimento em que os dados ignorados são substituídos por valores de outros informantes da mesma pesquisa

seleção é feita, sem restrições, no nó raiz, ou seja, considerando todos os informantes respondentes da mesma região metropolitana.

A imputação dos poucos casos de rendimentos ignorados do(s) outro(s) trabalho(s) dos não remunerados no trabalho principal é feita exclusivamente no nó raiz, ou seja, considerando todos as pessoas que responderam o rendimento na região metropolitana.

Para a não-resposta parcial, ou seja, para o caso em que um dos dois rendimentos foi informado (rendimento habitual do trabalho principal), o procedimento adotado foi outro. Neste caso, optou-se por aproveitar a informação de um dos rendimentos respondidos, visto que em geral cerca de 90% dos indivíduos possuem o mesmo valor para ambos os rendimentos pesquisados (habitual e efetivo). Portanto aquele que possuir rendimento habitual respondido e efetivo ignorado, irá doar o seu valor de rendimento habitual para o efetivo e vice-versa. Isto acontece de maneira análoga com o(s) rendimentos do(s) outro(s) trabalho(s), ou seja, variáveis: rendimentos efetivo do trabalho principal e rendimento habitual do(s) outro(s) trabalho(s).

A exceção para este procedimento acontece no mês de janeiro, que tem como mês de referência o mês de dezembro. Nos meses de janeiro, a equivalência entre os rendimentos habitual e efetivo é menor, devido, em grande parte, ao 13º salário recebido pelos trabalhadores. Neste caso, os rendimentos efetivos são obtidos multiplicando os rendimentos habituais por uma razão média entre os rendimentos habituais e os efetivos, calculada dentro de cada nó para três grupos distintos. Da mesma maneira, os rendimentos efetivos serão divididos por esta razão, para se obter os rendimentos habituais. O primeiro grupo é formado pelos trabalhadores domésticos, o segundo, por militares ou funcionários públicos e empregados com carteira e o terceiro, por empregados sem carteira, conta própria e empregadores.

A imputação dos rendimentos habituais dos que não possuem rendimento efetivo (rendimento efetivo igual a zero), no trabalho principal ou

no(s) outro(s) trabalho(s), também é feita selecionando-se um doador aleatoriamente nos nós da árvore.

Após o procedimento de imputação, em cada nó terminal da árvore efetua-se o teste de Kolmogorov-Smirnov (LEHMANN,1997), que é freqüentemente utilizado para avaliar se duas amostras têm distribuições semelhantes, ou melhor, se foram extraídas de uma mesma população. O teste é aplicado usando os valores das variáveis de rendimento antes e depois da imputação. Para os meses de março de 2002 até março de 2006, não foram encontrados p-valores abaixo do nível de significância (5%), desta forma concluímos que as distribuições não sofreram alterações após o procedimento de imputação.

5. REFERÊNCIAS:

BREIMAN, L., FRIEDMAN, J.H., OLSHEN R.H. and STONE, C.J. *Classification and Regression Trees*, 1984, Monterrey:Wadsworth and Brooks/Cole.

PESSOA, D.G.C. e SANTOS, A.R. *Imputação de renda dos responsáveis por domicílios - conjunto universo do Censo Demográfico 2000*, 2003, Relatório Técnico, DEMET/DPE/IBGE.

PESSOA,D.G.C., MOREIRA, G.G. e SANTOS, A.R. *it Imputação de rendimentos no questionário da amostra do Censo Demográfico 2000*, 2003, Relatório Técnico, DEMET/DPE/IBGE.

PESSOA, D.G.C., SILVA, P.L.N. e SANTOS, A.R. *Imputação para não-resposta parcial de renda na Pesquisa Mensal de Emprego*, Relatório Técnico, 2000, DEMET/DPE/IBGE.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

LEHMAN, E. L. *Testing Statistical Hypotheses*. New York: Springer Verlag, 1997.