ESTUDO PRELIMINAR DE DADOS PARA ANÁLISE DE DETERMINANTES DE SAÚDE NO BRASIL

C.A.C. PIRES[1], R.M.V.R. ALMEIDA[2], F.F. NOBRE[3], R.B.P. PANERAI[4]

RESUMO -- A disponibilidade de dados confiáveis sobre o Sistema de Saúde é indispensável para o desenvolvimento de trabalhos na área de planejamento em saúde. O presente trabalho descreve a implementação de um banco de dados relativos aos setores Saúde, Social, Econômico e Ambiental. As dificuldades de seleção dos dados são apresentadas bem como um método exclusivamente matemático para interpolação de populações. Um estudo do erro quando se utiliza interpolação linear para as variáveis não demograficas, indica um erro médio de 12% para interpolação de uma amostra em variáveis do setor saúde, mostrando a viabilidade deste procedimento quando necessário. A caracterização da função de probalidade das principais variáveis sugere uma distribuição gaussiana na maioria dos casos.

INTRODUÇÃO

A melhoria das condições de saúde de uma população depende de inúmeros fatores além da incorporação de novas tecnologias no atendimento médico (Dubos,1978). Entende-se a saúde como o bem estar físico, mental e social do indivíduo. A saude de uma população depende de fatores (ou determinantes) tais como: saneamento, moradia, educação, condições econômicas, nutrição e condições ambientais. Diferentes intervenções e/ou linhas de ação apresentam diferentes relações custo/benefício, sendo portanto fundamental que se desenvolvam estudos objetivos e quantitativos no sentido de se estabelecer as soluções ótimas, sob o ponto de vista da saúde condicionada pelo econômico e social(Evans,1983).

No Brasil, a importância qualitativa dos vários determinantes de saúde acima mencionados, tem sido constatada, porém de forma limitada e para um número reduzido de fatores e/ou populações (Paim,1988). Existe portanto a necessidade de estudos mais abrangentes utilizando um maior número de variáveis em populações mais representativas, obtendo-se assim subsídios para um planejamento integrado, possibilitando a tomada de decisões quanto a priorização de ações e alocação de recursos em saúde e a análise do impacto de diferentes intervenções (Banta,1982).

O modelamento dos determinantes de saúde de uma população ampla, utilizando um grande número de variáveis (indicadores) e com a preocupação de refletir a dinâmica do processo, representa não só uma urgente necessidade para otimizar-se a utilização de recursos, mas também o desenvolvimento de metodologias de análise multivariável aplicada a sistemas de saúde. Além de

---

1-Pesquisador do Programa de Engenharia Biomédica,COPPE/UFRJ
2-Aluno de mestrado do Programa de Engenharia Biomédica, COPPE/UFRJ
3-Professor Adjunto do Programa de Engenharia Biomédica, COPPE/UFRJ
4-Professor Adjunto do Programa de Engenharia Biomédica, COPPE/UFRJ

//Trabalho recebido em 30/5/87 e aceito em 17/7/87//

apontar intervenções setoriais ótimas no sentido de promoção da saúde, o modelo poderá ser extremamente útil para o planejamento e racionalização de serviços médicos. Pelas diferenças regionais observadas nos indicadores, seria possível identificar as cidades nas quais existam não só carência assistencial, mas também mortalidade excessiva por diferentes causas, indicando assim a necessidade de incorporação de tecnologias específicas a nível de prevenção, diagnóstico e terapia.

Como primeira etapa para o posterior desenvolvimento de modelos mais complexos, o Programa de Engenharia Biomédica da COPPE/UFRJ vem implementando um banco de dados relativos a variáveis dos setores ambiental, saúde e sócio-econômico. A dificuldade deste tipo de dados, tem sido até agora um dos motivos que dificultam o desenvolvimento de pesquisas nesta linha no nosso país. O objetivo do presente trabalho é descrever a implementação deste banco de dados, bem como avaliar alguns parâmetros úteis para a utilização correta destes dados. Isto permitirá a posterior definição de critérios para a interpolação ou coleta dos dados ausentes.

MATERIAIS

Idealmente os dados devem ser disponíveis a nível de município, visando uma comparação inter-regional, e com a maior duração possível (séries longas) a fim de permitir estudo do comportamento dinâmico destas variáveis (relação causa-efeito). O trabalho de coleta de dados é ilimitado pela própria natureza do projeto , pois exige a atualização anual da base de dados.

Ao início do projeto, foram escolhidos como pontos de coleta de dados, 59 municípios do Brasil, entre os 138 municípios que em 1980 apresentavam uma população maior que 100 000 habitantes. Estes municípios foram escolhidos segundo três critérios:

1-Todas as capitais (com mais de 100 000 habitantes).

2-Municípios com alta taxa de crescimento populacional no decênio 70-80, segundo ordem decrescente de crescimento.

3-Distribuição geográfica, visando melhor representatividade das diversas regiões na amostra.

Segundo estes critérios, foram selecionadas quase todas capitais de estado, cidades pertencentes as grandes regiões metropolitanas, e alguns municípios no interior dos estados com elevadas taxas de crescimento.

Os municípios selecionados representavam em 1980 32.9% da população total do País, e apresentam predominância acentuada de população urbana. A cada município atribuiu-se um número de código, em ordem decrescente conforme a população em 1980. Os municípios selecionados encontram-se listados junto com seu número de código na Tabela 1.

As variáveis selecionadas visam medir as alterações ocorridas nos diversos setores ao longo do tempo. Estas variáveis foram escolhidas não apenas objetivando indicadores ótimos quanto a sensibilidade e especificidade (WHO,1981), mas principalmente a partir da disponibilidade da informação a nível de município para todo o país. A maior parte dos dados utilizados foram obtidos junto ao Instituto Brasileiro de Geografia e Estatística (IBGE) ou outros orgãos públicos, procura-se assim evitar os erros que podem ocorrer devido a utilização de diferentes técnicas de coleta ou amostragem.

As variáveis inicialmente escolhidas foram 41, representativas dos três setores já citados, e ao longo do trabalho de coleta sofreram modificação em número, sendo algumas aglutinadas, outras suprimidas e novas variáveis inseridas. A seleção de diversas variáveis para cada setor permitirá a composição de indicadores mais específicos e sensiveis para o acompanhamento das alterações do setor. Deste modo são atualmente utilizadas 62 variáveis, que estão relacionadas na Tabela 2.

O objetivo inicial na coleta de dados era obter dados a partir de 1950, porém com o desenvolvimento do trabalho, observou-se que poucos dados anteriores a 1960 são disponiveis. Isto forçou a utilização de séries mais curtas, abrangendo apenas o período 1960-1985.

A coleta de dados foi iniciada em abril de 1983 e tem-se atualmente cerca de 33 000 dados já coletados, em sua maioria relativos ao período 1960-1984 e com predominância de dados do setor saúde. Estes dados são inicialmente digitados formando arquivos onde apresentam-se em ordem aleatória. Cada arquivo contém 1000 dados, em formato ASCII , sendo cada dado representado por um conjunto de 12 dígitos, onde dois dígitos são referentes ao código da cidade, dois relativos ao ano, três para o código da variável e os últimos 5 dígitos representam o valor do dado. As vantagens da utilização deste formato para o arquivo básico, são:

-possibilidade de transferência dos dados para qualquer máquina e em qualquer sistema operacional.
-facilidade de correção com o uso de qualquer editor de texto disponivel.
-permite o acesso dos dados por qualquer linguagem de programação.

Como desvantagem tem-se que os arquivos ocupam muito volume, não sendo portanto adequados para pesquisa direta (processamento lento), porém isto é facilmente resolvido logo ao início do processamento pela separação das séries históricas que são guardadas em arquivos não formatados. Os arquivos básicos são todos disponíveis em sistemas CP/M , RT11, e MS-DOS.

Todo o processamento dos dados está sendo realizado com programas desenvolvidos em liguagem FORTRAN, e ASSEMBLY em micro-computador PC compatível, o que facilitará bastante a divulgação dos resultados obtidos devido ao seu baixo custo e grande difusão.

PROCESSAMENTO DOS DADOS

A primeira etapa no processamento dos dados é a separação das séries históricas a partir dos arquivos de dados brutos. Obtém-se assim arquivos que contém cada um 59 séries, uma para cada cidade. A seguir é feita a normalização das variáveis, a fim de permitir um estudo comparativo entre cidades de diferentes tamanhos. As variáveis são normalizadas pela população (variáveis relativas à produção agrícola são normalizadas também pela área). Nesta etapa também os valores de produção agrícola e rebanho são convertidos para dólar, utilizando o câmbio médio anual, deflacionado para o ano de 1970.

O primeiro problema surgido no tratamento dos dados é a interpolação das variáveis demográficas. Os dados de população são disponíveis apenas para os anos em que são realizados os censos demográficos pelo IBGE, ou seja, os anos de 1960, 1970, e 1980, sendo que os dados de população rural não são disponíveis para 1960, porque neste ano estes dados não foram discriminados por faixa etária. Uma vez que o período do estudo é 1960-1985 e que os

demais dados tem periodicidade anual, torna-se então necessário não só interpolar no período intercensitário, mas também estimar a população a partir de 1980.

Como a extrapolação dos pontos a partir de 1980 pela utilização de uma função baseada na taxa de crescimento 70-80 levou em alguns casos a valores absurdos para 1985, optou-se por uma estimativa da população de 1990, baseado na média ponderada da variação populacional nos períodos de 60-70 e 70-80. Assim estima-se:

$$T_{1990} = (1/3)T_{1970} + (2/3)T_{1980}$$

Onde T_N = taxa de incremento no período N.

Deste modo tem-se 4 pontos e o problema pode ser então considerado como sendo apenas de interpolação.

Para interpolação utiliza-se uma função exponencial que representa uma população crescente com taxa de crescimento variável. Esta função foi escolhida após diversas tentativas, visando obter o ajustamento dos 4 pontos de população disponíveis. A função logística modificada de melhor resultado em termos de abrangência pode ser representada por:

$$Y_1 = Y_0.\exp(a.t)$$

Onde:

$$a = A.\exp(Bt+Ct^2)$$

$$t = (ano-60)/10$$

A, B e C podem ser obtidos por:

$$A = a_1/\exp(B+C)$$

$$B = -4\log(a_1/a_2) + (2/3)\log(a_1/a_3)$$

$$C = (2.\log(a_1/a_2)-\log(a_1/a_3))/2$$

Sendo:

a_1 = taxa de incremento populacional no período 60-70

a_2 = idem para o período 60-80

a_3 = idem para o período 60-90

Utiliza-se uma função com taxa de crescimento exponencial e do segundo grau, porque desta forma toda informação disponível (4 pontos) é utilizada. O resultado obtido pelo uso desta função para interpolar populações é apresentado nas figuras 1 e 2.

Uma das limitações desta função é que ela não pode ser aplicada para populações com taxas de crescimento nula ou negativa, neste caso é utilizada

uma estimativa baseada simplesmente na taxa de decremento da população. Deve-se observar que isto ocorre nas cidades em estudo em um número muito pequeno de casos e apenas para a população rural. Esta função foi utilizada apenas para as variáveis do setor demográfico (populações), e o erros em relação aos métodos clássicos de estimativa de população ainda não foram avaliados.

No atual estado de coleta, as séries ainda apresentam "lacunas" (pontos ausentes consecutivos) de vários tamanhos. Para verificar-se a possibilidade dos dados ausentes serem interpolados, foi realizado um teste, utilizando interpolação linear, visando verificar o erro de interpolação de algumas das variáveis disponíveis. Para isso, foram tomadas as séries mais completas (N > 15), e nelas era retirada uma de suas amostras, a qual foi substituída pela interpolação linear de seus dois pontos adjacentes. O resultado assim obtido foi comparado ao valor do ponto verdadeiro, e o erro percentual armazenado. O processo foi repetido para cada um dos pontos das séries obtendo-se tabelas do erro de interpolação de uma amostra por variável. A seguir foi realizada a mesma análise com a retirada de duas e três amostras em lugar de uma. Para o setor saúde, o erro médio de interpolação foi de aproximadamente 12%, excetuando-se as variáveis de código 160 e 170 que apresentaram erro em torno de 35%, o que pode ser explicado devido a natureza casuística destas variáveis. Estes erros de interpolação são apresentados na Tabela 3.

Também é interessante pesquisar-se as características de distribuição das variáveis diponíveis, previamente à sua utilização. Para isso, todos os dados não-interpolados de cada uma das variáveis, normalizadas por número de habitantes, foram reunidos, sem discriminação de cidades ou regiões, e seu histograma de distribuição de frequência traçado. Pode-se observar que a maior parte das variáveis do setor saúde apresentam uma distribuição normal, como por exemplo as variáveis 100, 110 e 111. Outras como a 101 e 141 possuem certo grau de curtose, o que pode ser reflexo de problemas de coleta ou diferenças regionais. Nas variáveis do setor sócio-econômico, via de regra não se observa distribuição normal, sendo que algumas apresentam distribuição bimodal, o que também sugere a existência de diferenças regionais. As figuras 3 e 4 ilustram estas distribuições.

DISCUSSÃO E CONCLUSÕES

Os principais problemas para o desenvolvimento deste banco de dados foram a indisponibilidade dos dados mais antigos e a descontinuidade das séries históricas devido a mudanças de conceituação da variável, como por exemplo, a mudança do registro de óbitos que era feito segundo o local de ocorrência e passou na decada de 70 a ser segundo local de residência. A impossibilidade de acesso a informações reservadas também é um problema a ser resolvido.

Os métodos clássicos de estimativa de população baseiam-se em técnicas de estimativa por faixa etária (estratificação) utilizando também o número de nascimentos, óbitos por faixa etária, migração, etc, para estimar a população anual (IBGE,1978). Como o objetivo do trabalho é verificar a interrelação entre muitas destas variáveis, a utilização desta técnica fica prejudicada, dado que ela fatalmente iria induzir falsas correlações entre as variáveis. Isto porém não exclui a possibilidade de utilização de métodos clássicos de correção de dados demográficos para algum tipo de análise, desde que o erro introduzido pelo 'uso cíclico' da informação possa ser desprezado. A utilização de métodos puramente matemáticos para a estimativa de população, embora não sendo o mais correto, apresenta a vantagem de não induzir falsas correlações além de ser de fácil aplicação. Uma característica fundamental

para a interpolação é a ausência de pontos de inflexão nas séries interpoladas. Isto exige a aplicação de uma função que tenha derivada contínua e expresse o crescimento da população a uma taxa não constante.

Com relação ao erro de interpolação linear, o baixo valor encontrado permite afirmar que a interpolação linear de amostras ausentes não interfere na estrutura interna da série, pelo menos para lacunas de tamanho 1.

Esse é um resultado interessante, pois permite a extensão do tamanho das séries epidemiológicas possíveis de serem usadas, sem que trabalho de modelamento posteriores, por exemplo com métodos de análise multivariada, tenham seus resultados afetados. Isso será particularmente útil para países do Terceiro Mundo, onde sabemos que por problemas de coleta, erros de sub-registro ou de desorganização do sistema de estatísticas de saúde, muitas vezes as séries apresentam-se incompletas.

Em outro trabalho (Almeida,1987), esses resultados foram usados para a análise de alguns parâmetros úteis ao modelamento de determinantes de saúde no Brasil, com ênfase em séries de pequena extensão temporal.

A normalidade encontrada na distribuição de frequência de variáveis do setor saúde é importante, pois este é um pré requisito para a aplicação de vários testes estatísticos clássicos.

Para variáveis de outros setores, como por exemplo a variável 621 (consumo de energia elétrica, Figura 3), nem sempre caracterizou-se uma distribuição normal, o que exigirá cuidados adequados em seu tratamento, como por exemplo o uso de estatísticas não-paramétricas ou técnicas de simulação de Monte-Carlo.

Com o objetivo de demonstrar a potencialidade da análise multivariada dos determinantes de saúde, como ferramenta de apoio em planejamento, encontra-se em desenvolvimento um sistema demonstrativo, implementado em micro-computador, que utiliza recursos gráficos e técnicas de processamento permitindo a visualização de séries históricas (normalizadas ou não), diagramas de espalhamento e seleção de grupos de cidades por parâmetro. O sistema interage com o operador em diversos níveis, permitindo a seleção de cidades, variáveis, etc. Estas características tornam o sistema demonstrativo ponto de partida para o desenvolvimento de um sistema especialista voltado ao apoio à decisão na área de planejamento de saúde coletiva.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro concedido pela CAPES, FINEP e CNPq e as seguintes entidades pelo fornecimento de dados: Ministério da Saúde, Serviço Estadual de Análise de Dados Estatísticos de São Paulo (SEADE), Fundação Estadual de Estatística do RGS (FEE), Instituto Brasileiro de Geografia Estatística (IBGE), e as Secretarias Estatuais de Saúde dos estados do Rio de Janeiro, Minas Gerais, Bahia, Paraíba, e Rio Grande do Sul.

Agradecimentos também são apresentados aos Profs. B. McA. Sayers do Imperial College e E. O. Attinger da Universidade de Virginia pela colaboração na fase inicial deste projeto.

REFERÊNCIAS

ALMEIDA, R. M. V. R.; NOBRE, F. F.; PANERAI, R. B.;(1987) - "Caracterização Dinâmica de Séries Históricas de Determinantes de Saúde. Submetido ao X Congresso Brasileiro de Engenharia Biomédica, Rio de Janeiro.

BANTA, D.;(1982) - "Resources for Health", Progen, New York.

DUBOS, R.;(1978) - "Man Adapting", Yale University Press, New Haven.

EVANS, R. W.;(1983) - "Health Care Technology and the Inevitability Rationing Decisions" Journal of American Medical Association , volume 15 pagina 249.

IBGE; (1978) - "Métodos Para Preparar Projeções de População por Sexo e Idade", Rio de Janeiro.

MARTINI, C. S.;ALLAN, B.;DEVIDSO, J.;BACKETT, R. M.;(1980) - "Health Care Indexes Sensitive to Medical Care Variations", International Journal of Health Services, volume 2 paginas 293-309.

PAIM, J. S.;DIAS, C. N.;ARAUJO, J, D,;(1980) - "Influência de Fatores Sociais e Ambientais na Mortalidade Infantil", Boletim da Oficina Sanitária Panamericana, volume 88 paginas 327-391.

SIEGEL, S.;(1985) - "Estatística Não-Paramétrica", McGraw-Hill.

WORLD HEALTH ORGANIZATION (1981) - "Development of Indicators For Monitoring Progress Towards Health For All By The Year 2000", World Health Organization, Suiça.

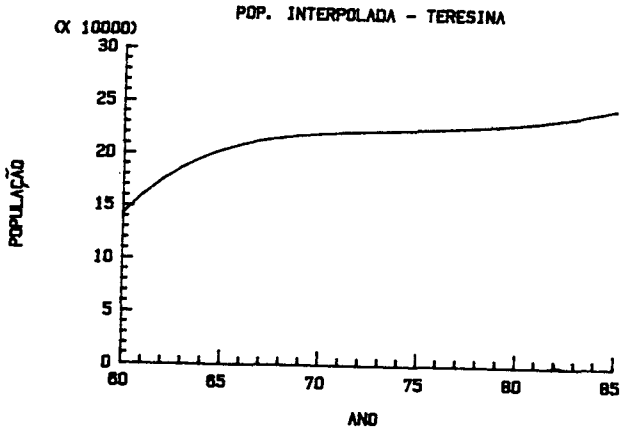

Figura 1 - População interpolada para o município de Teresina.

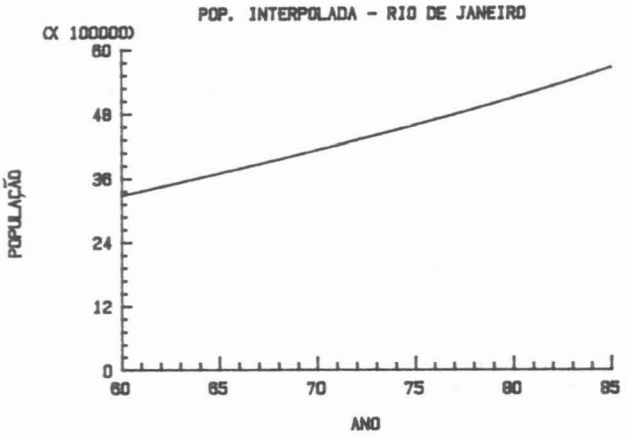

Figura 2 - População interpolada para o município do Rio de Janeiro.

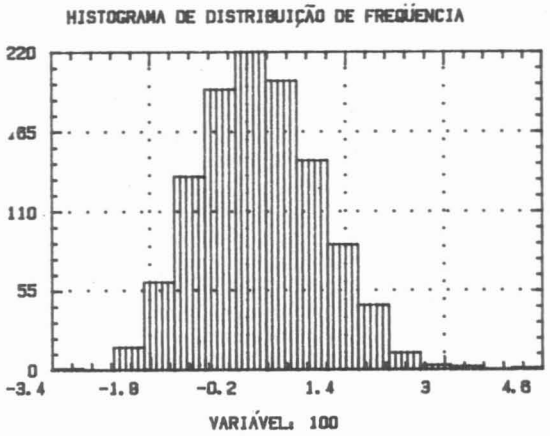

Figura 3 - Distribuição de freqüência para variável 100 (mortalidade total).

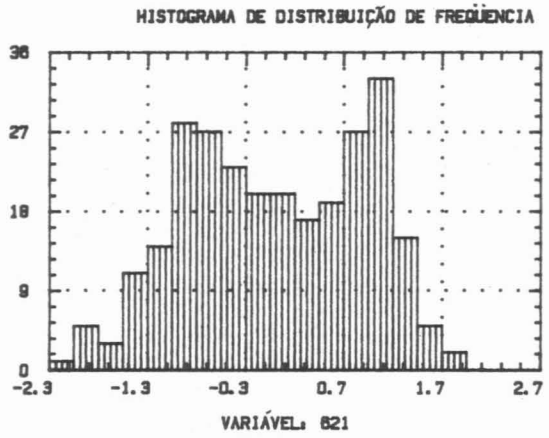

Figura 4 - Ditribuição de freqüência para variável 621 (consumo de energia elétrica)

Porto Alegre	Ribeirão Preto	Goiânia
Pelotas	S. J. dos Campos	Campo Grande
Canoas	Diadema	Cuiabá
Rio Grande	Mauá	Salvador
Novo Hamburgo	Carapicuíba	Feira de Santana
Viamão	Sumaré	Fortaleza
Gravataí	Rio de Janeiro	Recife
Curitiba	Nova Iguaçu	Jaboatão
Londrina	São Gonçalo	Olinda
Cascavel	Duque de Caxias	Paulista
Foz do Iguaçu	Niterói	São Luís
Joinville	S. J. do Meriti	Imperatriz
Florianópolis	Vitória	Natal
São Paulo	Vila Velha	Maceió
Campinas	Belo Horizonte	Teresina
Santo André	Contagem	João Pessoa
Guarulhos	Uberlândia	Aracaju
Osasco	Uberaba	Belém
S. B. do Campo	Ipatinga	Manaus
Santos	Brasília	

Tabela 1 - Relação dos municípios selecionados

Tabela 2 - Relação das variáveis selecionadas com seus códigos.

010- Nascidos vivos
060- Número de auxilio-doença concedios
070- Número de aposentados por invalidez
100- Número de óbitos todas faixas etárias, todas as causas.
101- Número de óbitos de 1-4 anos
102- Número de óbitos de 5-19 anos
103- Número de óbitos de 20-49 anos
104- Número de óbitos de 50-64 anos
105- Número de óbitos de maiores de 65 anos
110- Mortalidade infantil .
111- Número de natimortos.
120- Mortalidade por doenças infecciosas e parasitárias
130- Mortalidade por doenças cardiovasculares.
140- Mortalidade por doenças respiratórias.
141- Mortalidade por doenças respiratórias, 0 a 1 ano.
142- Mortalidade por doenças respiratórias, 1 a 4 anos.
150- Mortalidade por doenças neoplásicas.
160- Mortalidade por acidentes de trânsito.
170- Mortalidade por eventos violentos.
180- Mortalidade por doenças diarreicas.
181- Mortalidade por doenças diarreicas, 0 a 1 ano.
200- Número de leitos.
210- Número de baixas hospitalares.
220- Número de consultas ambulatoriais.
300- Precipitação pluviométrica anual.
310- Temperatura mínima média mensal no ano.
500- População do município de menores de 1 ano.
501- População do município de 1 a 4 anos.
502- População do município de 5 a 19 anos.
503- População do município de 20 a 49 anos.
504- População do município de 50 a 65 anos.
505- População do município de maiores de 65 anos.
510- População rural do município de menores de 1 ano.
511- População rural do município de 1 a 4 anos.
512- População rural do município de 5 a 19 anos.
513- População rural do município de 20 a 49 anos.
514- População rural do município de 50 a 65 anos.
515- População rural do município de maiores de 65 anos.
530- Quantidade de água produzida.
535- Extensão da rêde de esgotos.
537- Número de moradias c/algum sistema de esgoto.
560- Número total de moradias.
561- Número de moradias c/agua encanada.
562- Número de moradias c/poço ou nascente.
570- Número de veiculos automotores
580- Número de veiculos de transporte coletivo.
590- Número de ferimentos de acidentes de trânsito.
600- Número de aparelhos telefônicos.
610- Area do município.
620- Consumo total de energia elétrica.
621- Consumo de energia elétrica industrial.
700- Número de pessoas com alfabetizadas maiores de 5 a.
701- Número de pessoas com 8 ou mais anos de estudo.
702- Número de pessoas com 11 ou mais anos de estudo.
730- Total de terra arável.
735- Total de área cultivada.
740- Valor da produção agrícola do município.
741- Valor do rebanho.
742- Número de culturas para atingir 80% do valor da produção.
743- Número de estabelecimentos agrícolas.
750- Número de empregos regulares.
780- População rural total.

	número de interpolações		
variável	1	2	3
10	.044	.076	.103
100	.065	.100	.129
101	.214	.303	.370
102	.144	.227	.288
103	.093	.142	.180
104	.097	.150	.189
105	.088	.128	.160
110	.116	.178	.232
111	.099	.164	.214
120	.224	.346	.443
130	.108	.177	.233
140	.177	.275	.349
141	.188	.271	.326
142	.338	.514	.654
150	.112	.177	.224
160	.353	.633	.915
170	.314	.558	.809
180	.221	.357	.485
181	.201	.306	.400
200	.185	.455	.735
210	.350	.649	.951
220	.390	.602	.858
300	.179	.273	.329
310	.061	.096	.120
530	.078	.126	.177
535	.014	.027	.043
561	.057	.097	.126
570	.078	.125	.165
580	.076	.121	.158
600	.125	.248	.363
735	.258	.492	.576
740	.480	.684	.777
742	.516	.773	1.032

Tabela 3 - Erro relativo de interpolação linear por variável, para lacunas de tamanho 1, 2 e 3.

PRELIMINARY ANALYSIS OF DETERMINANTS OF HEALTH IN BRAZIL

ABSTRACT -- The availability of reliable data about the Health System is the first step towards the development of research work in health planning. The present work describes the implementation of a data base incorporating health, social, economic and environmental variables. The difficulties with the identification of suitable data are discussed and a mathematical method for use in the interpolation of demographic data is introduced. An evaluation of the error using linear interpolation for estimating missing values of non-demographic variables resulted in a mean error of 12% when interpolating one sample of health variables. This result showed the possibility of using this approach when necessary. The characterization of the probability distribution of some of the available variables suggests a Gaussian distribution in most cases.