THE RECONSTRUCTION OF AGE-SPECIFIC DISEASE DEATH RATES FROM INCOMPLETE DATA

Joseph E. Quinn, U. S. Bureau of the Census

This paper describes a methodology that is being used to reconstruct and estimate some social indicators for the Brazilian health sector in the twentieth century. First, the kinds of partial data available in Brazil are described. Second, a method for constructing a matrix of age-specific death rates using marginal data on Brazilian death rates by age and disease and R.A.S. techniques is outlined. Third, an illustration of the method is given using Portuguese data. Fourth, some tests useful for comparing any projections made with actual data, when available, are outlined. Fifth, an example of the usefulness of constructing age-disease death rates is given. Sixth, a transition matrix for a Brazilian state is shown and its implications noted. MORTE, the computer program used, is part of the LRPM (Long-Range Planning Model) system, and will be available from the author.

This paper is part of a larger project to build a structural model of Brazilian society. Structuralism as described by Celso Furtado 3/ is an attempt: "Starting from a global historical perspective to identify those elements which will permit for a given time period the simplification of social reality to a reduced system on which we can apply the tools of economic analysis. Focusing on a short time period is necessary so that certain elements will be stable enough for awhile that we can consider them unchangeable parts of the structure. The global historical perspective for its part will permit us to observe these at one point structural elements as variables whose importance will change when we move from one time period to another".

First, Brazilian data on the structure of deaths by age is extensive and fairly robust estimates of the life table can be obtained for (in order of quality) 1970, 1950, 1940, and 1920 by region and by some definitions of social class (occupation, education, and income). However the structure of deaths by cause is harder to obtain. Deaths by disease are usually available only in the country registries where they have been entered in longhand because these statistics are seldom compiled outside of the state capitals and some large cities. Data on deaths by cause are rarely available for rural areas (although there is a good recent survey of ten districts of the state of São Paulo and another of a northeastern state). Death rate statistics by both age and cause have rarely been compiled. It is outside this paper but a picture of the marginal rates for all of Brazil by disease can be gradually built by interrelating the available fragments of data.

Second, once reasonably robust estimates of deaths by both age and cause are available, it is possible to contemplate the construction of a matrix whose cells would be age-specific death rates by cause, for example: deaths from cancer at age 20. The matrix can take many forms. The cells can be: (a) actual numbers of deaths, (b) percent of age group who die, (c) percent of people who die from this cause in these age groups (columns add to one), (d) percent of people who die in this age group from these causes (rows add to one), and (e) percent of all deaths that occur in this cell (matrix adds to one), etc. An analogous and parallel set of matrices can be built around the life tables or around a stable population or around standardized population data. All of the matrices have different possible uses (see Boudon 2/).

If we have a matrix of age-disease specific death rates, A_{ij}, rates by age $A_{.j}$, and by disease $A_{i.}$ -- and we have just the marginal rates for a second matrix, $B_{.j}$ and $B_{i.}$; we can by an iterative process estimate the B_{ij} using the A_{ij} and the marginals known for B. This is often done in projecting input-output tables (see Bacharach 1/).

Third, using data from both the World Health Organization and that gathered by Preston, et.al., 5/, I have tested the methodology on actual data from a large number of countries over periods as long as a century. Tables 1 and 3 show two matrices for Portugal in 1920 -- one from the actual rates and the other from a life table that conforms to the data. Tables 2 and 4 show predicted deaths in 1930. Tables 2 and 4 are practically identical and can be compared with the actual results in Table 5. Despite the dramatic reduction in unknown causes from 50% of deaths in 1920 to 20% in 1930 and the rise in life expectancy from 34 to 45 years at birth, the predictions are quite good.

The marginal disease rates for several countries were estimated from time series regressions -- incidentally the regressions show no significant correlation between diseases (except infant mortality and all other diseases) once time trends were removed -- and again the results (using the methodology and time series estimates to use one known year to predict another known year) were compared to actual data and indicated that this methodology was a very good prediction device. Brazilian data when available was tested against matrices from other countries to select the best prediction vectors.

Fourth, the results of a prediction can be tested against actual results when available using either information theory or the inequality coefficient (see Theil 6/, 7/, and 8/).

Three measures from information theory are the most useful:

(a) information entropy,
(b) information inaccuracy,
(c) information improvement.

(a) Information Entropy

$$H = \sum_{i=1}^{n} q_i \cdot \log_2 \frac{1}{q_i}$$

(b) Information Inaccuracy

$$I(q:p) = \sum_{i=1}^{n} q_i \log_2 \frac{q_i}{p_i}$$

(c) Information Improvement

$$I(q:p) - I(q:p') = \sum_{i=1}^{n} q_i \log_2 \frac{p_i'}{p_i}$$

where q_i = actual results,

p_i = first prediction,

p_i' = second prediction.

For example, Table 6 gives five estimates of deaths in Portugal for 1930 with a given life expectancy from the West, North, East, and South Model Life Tables and a fifth estimate which resulted from only one iteration of the 1920 matrix (by causes of death). The Information Inaccuracy of the predictions compared with the actual results are (expressed in 10^{-5} bits):

North	4509
West	4002
South	3529
East	2264
R.A.S.	1794

Thus even with only one iteration and no supplementary changes of cells known to be weak, the R.A.S. technique was superior to all the model life tables in this example (the lower the score, the lower the inaccuracy). Actually with a time series and the use of the inequality coefficients and its decompositions, it is possible to build predictions that have only 5% to 10% of the inaccuracy of the model life tables compared with actual data for many countries.

The inequality coefficient, U, takes the form:

$$U^2 = \Sigma \frac{(p_i - q_i)^2}{\Sigma q_i^2}$$

The numerator of the inequality coefficients is the root mean square prediction error and can be decomposed into a number of terms, each of which will reflect a particular type of prediction error. Consider the following decomposition:

$$\frac{1}{n} \Sigma(p_i - q_i)^2 = (p-q)^2 + (s_p - rs_q)^2 + (1-r^2)s_a^2$$

This can be represented as three terms which add up to one:

(i) The <u>Bias Proportion</u> U^m

which shows whether the average predicted change deviates substantially from the average realized change.

$$U^m = \frac{(\bar{p}-\bar{q})^2}{\frac{1}{n} \Sigma (p_i - q_i)^2}$$

(ii) The <u>Regression Proportion</u> U^r

which deals with the deviation of the regression scope from 1.

$$U^k = \frac{(s_p - rs_q)^2}{\frac{1}{n} \Sigma (p_i - q_i)^2}$$

(iii) The <u>Disturbance Proportion</u> U^d

which deals with the variance of the regression disturbance.

$$U^d = \frac{(1-r)^2 s_q^2}{\frac{1}{n} \Sigma (p_i - q_i)^2}$$

$$U^p + U^m + U^r = 1$$

(iv) Since U^m and U^r are sustematic errors, an <u>optimal linear correction</u> of the forecasts would be: $a + bp_i$,

where $a = \bar{q} - b\bar{p}$

and $b = \frac{rs_q}{s_p}$

Fifth, to illustrate the usefulness of knowing age-specific and cause-specific death rates, Table 7 shows some results of what it would mean to eliminate Chargas disease in areas of Brazil where it is epidemic (actually some areas have rates three times as high as this example which is for males only). Chargas disease is spread by an insect vector (the triatoma) and usually kills its victims several years after the initial infection. Some die suddenly from heart attacks or cerebral hemorrhages but many die by inches growing weaker and weaker over a period of months or years. No cure is known. In the example, 1.4 years of working life per person would be gained. For 100,000 people in risk with an average life expectancy of 60 years, the estimated cost of repeatedly spraying their homes throughout their lifetimes to kill the insect carrier of the disease would be $240,000, the gains just in wages even at the minimum levels prevailing in rural São Paulo and rural Minas Gerais would be over $150,000,000 -- which leaves plenty of room for fighting the disease more vigorously than presently is the case even if the costs are much higher than estimated here. A more complete simulation was also run with an age-sex distribution similar to rural Brazil with the disease being eliminated over a 20-year period, the cost benefit ratio was still over 100 to 1.

Sixth, Table 8 shows the distribution of every 10,000 deaths in Rio de Janeiro in 1940 and 1970. Most notable is the small change in the large share of total mortality that is childhood mortality between the two years, and the big improvement in the numbers not dying during the working years but living to after age 60. While changing age structure and falling death rates have had their influence, this matrix reflects the rapid improvements in conditions for some social classes but not for others. Where class differences were around five to six years in life expectancy in 1940, they are now approaching 30 years with some groups having only around 35-40 years of life expectancy and others 65-70.

In conclusion, this methodology offers some promise as a tool for less developed countries to prepare health statistics and to move away from reliance for predictions on the recognized faulty model life tables (see Lederman 4/). It will also produce some statistics that should influence policy and could aid in cost-benefit studies.

BIBLIOGRAPHY

1/ Bacharach, Michael -- Biproportional Matrices and Input-Output Change, Cambridge, 1970.

2/ Boudon, Raymond -- Mathematical Structures of Social Mobility, Amsterdam, 1973.

3/ Furtado, Celso -- Analise Do "Modelo" Brasileiro, Rio de Janeiro, 1972.

4/ Lederman, Sully -- Nouvelles Tables - Types de Mortalite, Paris, 1969.

5/ Preston, Samuel, N. Keyfitz, and R. Schoen -- Causes of Death, Life Tables for National Populations, New York, 1972.

6/ Theil, Henri -- Applied Economic Forecasting, Amsterdam, 1966.

7/ ________ -- Economics and Information Theory, Amsterdam, 1967.

8/ ________ -- Statistical Decomposition Analysis, Amsterdam, 1972.

TABLE 1 -- DEATHS BY DISEASE AND AGE FOR PORTUGAL IN 1920 (Males Only)
(Percent of Age Group Who Die From Specific Cause Times 10,000)

Age at Start of Interval	TB	P/I	C	H	FL	D	DG	M	IF	MV	V	O
0	37	1168	2	68	1743	7333	55	0	4730	0	109	13868
1	43	559	2	18	423	1137	48	0	0	0	69	2457
5	25	128	1	10	53	132	21	0	0	0	27	446
10	21	63	1	11	31	41	10	0	0	0	33	246
15	108	75	1	17	46	21	9	0	0	0	54	268
20	273	101	3	23	102	16	15	0	0	0	75	374
25	298	78	6	23	67	19	21	0	0	0	78	376
30	244	63	10	41	103	22	25	0	0	0	66	382
35	257	62	14	66	112	19	38	0	0	0	64	446
40	228	81	20	110	110	29	35	0	0	0	58	607
45	226	66	47	166	154	39	67	0	0	0	57	701
50	192	64	67	274	162	51	93	0	0	0	60	951
55	198	76	107	454	218	75	104	0	0	0	67	1282
60	127	72	125	744	291	100	127	0	0	0	72	1943
65	128	76	165	1321	448	155	155	0	0	0	84	3046
70	68	99	170	2303	655	267	212	0	0	0	124	5441
75	104	92	180	3765	978	481	276	0	0	0	163	9427
80	52	118	140	3468	789	325	229	0	0	0	74	13288
85	31	173	157	5228	1413	581	314	0	0	0	110	27869

CODES:

TB -- Respiratory T.B.
P/I -- Other Infections and Parasites
C -- Neoplasms
H -- Cardiovascular
FL -- Infl., Pneu., Bronch.
D -- Diarrheal
DG -- Certain Degenerative
M -- Maternal
IF -- Certain Diseases of Infancy
MV -- Motor Vehicle
V -- Other Violence
O -- Other and Unknown

TABLE 2 -- PREDICTED DEATHS BY DISEASE AND AGE FOR PORTUGAL IN 1930 (Males Only)
(Prediction Based on Total Deaths in 1920)

Age at Start of Interval	TB	P/I	C	H	FL	D	DG	M	IF	MV	V	O	SUM	DEA
0	45	903	2	60	773	5559	55	0	3238	0	121	5224	15977	15980
1	151	1249	7	46	542	2490	139	0	0	0	211	2674	7518	7519
5	67	220	3	20	52	223	47	0	0	0	67	374	1073	1073
10	75	143	4	29	40	92	30	0	0	0	108	273	793	793
15	406	180	4	47	63	49	28	0	0	0	185	313	1274	1274
20	839	198	9	52	115	31	38	0	0	0	211	357	1850	1850
25	852	142	18	48	70	34	50	0	0	0	204	334	1751	1751
30	684	113	29	84	106	39	58	0	0	0	169	333	1614	1614
35	670	103	37	125	107	31	82	0	0	0	152	361	1669	1669
40	520	118	47	183	92	41	66	0	0	0	121	430	1617	1617
45	482	90	103	258	120	52	118	0	0	0	111	464	1798	1798
50	384	82	138	399	118	64	154	0	0	0	110	591	2040	2040
55	348	85	193	581	140	82	152	0	0	0	108	700	2389	2389
60	215	78	218	920	181	106	179	0	0	0	112	1024	3032	3032
65	153	58	203	1152	196	116	154	0	0	0	92	1132	3254	3255
70	60	56	154	1479	211	147	155	0	0	0	100	1490	3850	3851
75	50	28	89	1327	173	145	111	0	0	0	72	1416	3411	3412
80	16	23	44	773	88	62	58	0	0	0	21	1262	2347	2347
85	5	18	26	610	83	58	42	0	0	0	16	1387	2244	2244

NOTE: SUM -- Is sum of rows,
DEA -- Is marginal age death rate.

ALSO SEE CODE ON TABLE 1.

TABLE 3 -- LIFE TABLE DEATHS BY DISEASE AND AGE FOR PORTUGAL IN 1920 (Males Only)
(Out of An Original Cohort of 100,000)

Age at Start of Interval	TB	P/I	C	H	FL	D	DG	M	IF	MV	V	O
0	33	1037	2	60	1547	6509	49	0	4198	0	97	12310
1	114	1480	5	48	1120	3011	127	0	0	0	183	6506
5	75	385	3	30	160	397	63	0	0	0	81	1343
10	61	183	3	32	90	119	29	0	0	0	96	716
15	310	214	3	48	132	59	26	0	0	0	154	763
20	750	276	8	63	279	44	41	0	0	0	205	1024
25	774	202	16	60	174	49	55	0	0	0	203	977
30	604	156	25	102	256	54	62	0	0	0	163	946
35	605	146	33	156	264	45	90	0	0	0	151	1051
40	506	180	44	244	244	64	78	0	0	0	129	1348
45	468	137	97	344	319	81	139	0	0	0	118	1452
50	365	122	128	523	308	97	177	0	0	0	114	1813
55	337	129	182	774	371	128	177	0	0	0	114	2184
60	185	105	183	1089	426	146	186	0	0	0	105	2843
65	149	89	193	1546	524	181	181	0	0	0	98	3563
70	55	80	138	1868	531	217	172	0	0	0	101	4413
75	46	41	79	1655	430	211	121	0	0	0	72	4138
80	9	24	26	638	146	61	42	0	0	0	14	2465
85	2	11	10	319	86	35	19	0	0	0	7	1700

SEE CODE ON TABLE 1.

TABLE 4 -- PREDICTED DEATHS BY DISEASE AND AGE FOR PORTUGAL IN 1930 (Males Only)
(Prediction Based on Life Table for 1920)

Age at Start of Interval	TB	P/I	C	H	FL	D	DC	M	IF	MV	V	O	SUM	DEA
0	45	903	2	60	773	5559	55	0	3238	0	121	5224	15977	15980
1	151	1249	7	46	542	2490	139	0	0	0	221	2674	7518	7519
5	67	220	3	20	52	223	47	0	0	0	67	374	1073	1073
10	75	143	4	29	40	92	30	0	0	0	108	273	793	793
15	408	179	4	46	63	49	28	0	0	0	185	312	1274	1274
20	841	197	10	51	115	31	38	0	0	0	210	357	1850	1850
25	852	142	18	48	70	34	50	0	0	0	204	334	1751	1751
30	684	113	29	84	106	39	58	0	0	0	169	333	1614	1614
35	670	103	37	126	107	31	82	0	0	0	152	361	1669	1669
40	520	118	47	183	92	41	66	0	0	0	121	430	1617	1617
45	482	90	103	258	120	52	118	0	0	0	111	464	1798	1798
50	383	82	138	400	118	64	154	0	0	0	110	591	2040	2040
55	348	85	193	582	140	82	152	0	0	0	108	700	2389	2389
60	215	78	218	921	181	106	179	0	0	0	112	1024	3032	3032
65	153	58	202	1152	196	116	154	0	0	0	92	1132	3255	3255
70	60	56	154	1479	211	147	155	0	0	0	100	1490	3850	3851
75	50	28	90	1347	173	145	111	0	0	0	72	1415	3411	3412
80	15	25	44	768	88	63	58	0	0	0	22	1264	2347	2347
85	5	18	26	610	83	58	42	0	0	0	16	1387	2244	2244

NOTE: SUM -- Is sum of rows,
DEA -- Is marginal age death rate.

ALSO SEE CODE ON TABLE 1.

TABLE 5 -- ACTUAL DISTRIBUTION OF DEATHS BY DISEASE AND AGE FOR PORTUGAL IN 1930 (Males Only)

Age at Start of Interval	TB	P/I	C	H	FL	D	DG	M	IF	MV	V	O	SUM	DEA
0	64	1044	1	74	726	6057	61	0	3238	0	178	4537	15976	15980
1	129	1288	9	57	388	2686	172	0	0	0	315	2475	7517	7519
5	65	223	4	23	35	123	34	0	0	0	123	442	1073	1073
10	75	148	3	53	36	33	33	0	0	0	109	303	793	793
15	434	189	7	47	57	17	23	0	0	0	152	348	1274	1274
20	915	206	12	67	67	18	33	0	0	0	158	375	1850	1850
25	853	123	10	106	62	20	37	0	0	0	158	384	1751	1751
30	698	107	26	125	67	12	52	0	0	0	122	405	1614	1614
35	642	70	36	157	95	15	66	0	0	0	125	464	1669	1669
40	511	76	60	174	84	17	89	0	0	0	105	502	1617	1617
45	474	66	92	270	97	23	103	0	0	0	127	547	1798	1798
50	361	83	122	359	148	29	141	0	0	0	124	672	2040	2040
55	304	58	162	559	166	43	149	0	0	0	125	822	2389	2389
60	218	62	207	891	225	72	196	0	0	0	112	1049	3032	3032
65	143	54	210	1144	252	60	183	0	0	0	92	1117	3255	3255
70	90	42	181	1470	284	77	157	0	0	0	74	1476	3851	3851
75	37	26	110	1243	247	63	107	0	0	0	52	1528	3412	3412
80	7	12	49	769	120	30	48	0	0	0	26	1285	2347	2347
85	2	9	27	603	113	28	34	0	0	0	22	1406	2244	2244

NOTE: SUM -- Is sum of rows,
DEA -- Is marginal age death rate.

ALSO SEE CODE ON TABLE 1.

TABLE 6 -- **TOTAL NUMBER OF ESTIMATED DEATHS BY AGE FOR PORTUGAL IN 1930**

Age at Start of Interval	Method of Estimation: Model Life Tables West	North	East	South	R.A.S.	Actual Deaths
0	12292	10602	15852	12656	15099	15980
1	6735	8279	6200	9818	7134	7519
5	1788	1712	2167	2681	1474	1073
10	1104	1472	808	910	783	793
15	1621	1776	1331	1412	1141	1274
20	2068	2276	1744	1938	1918	1850
25	1854	1948	1446	1571	1843	1751
30	1748	1703	1320	1391	1658	1614
35	1951	1811	1523	1463	1838	1669
40	2200	1980	1765	1659	1850	1617
45	2380	2090	2027	1808	2054	1798
50	3035	2570	2633	2332	2296	2040
55	3277	2730	3007	2643	2744	2387
60	4078	3372	3748	3450	3122	3032
65	4080	3495	3966	3660	3813	3255
70	4274	3854	4334	4202	4275	3851
75	3693	3384	3829	3898	3751	3412
80	5149	5001	5115	5075	2752	4591
SUM	63327	60055	62815	62567	59506	59506

TABLE 7 -- **THE EFFECTS OF ELIMINATING CHARGAS DISEASE** (Number of Survivors of 100,000 births

Age at Start of Interval	If Chargas Eliminated	With Chargas	Extra Years of Life	Extra Years in Labor Force
(1)	(2)	(3)	(4)	(5)
0	100,000	100,000		
1	98,559	98,559		
5	96,845	96,845		
10	96,101	96,101		
15	95,267	95,267	390	66
20	94,183	94,027	1,235	1,103
25	92,896	92,558	2,780	2,685
30	91,751	90,977	5,110	4,936
35	90,415	89,145	8,403	8,117
40	88,932	86,841	13,100	12,445
45	86,946	83,797	17,698	16,459
50	83,534	79,604	22,132	19,454
55	78,838	73,915	24,958	20,316
60	71,792	66,732	25,522	18,529
65	62,172	57,023	24,000	14,712
70	49,669	45,218	19,358	9,679
75	31,894	28,602	14,208	5,683
80	20,487	18,096	10,405	2,081
85+	13,139	11,368	8,855	886

NOTE: Column (4) is the result of an interpolation from Column (3) and covers five year interval. Column (5) is Column (4) times the appropriate labor force participation rate.
Gains per member of original cohort of 100,000 birth.
Extra Years of Life: 1.98
Extra Years in Labor Force: 1.37

TABLE 8 -- **TRANSITION MATRIX FOR RIO DE JANEIRO SHOWING THE DISTRIBUTION OF 10,000 DEATHS IN 1940 AND 1970 BY AGE AND DISEASE**

			AGE AT DEATH AND CAUSE												
			0-4			5-19			20-59			60+			
Age at Death	Cause		P/I	C/H	O	P/I	C/H	O	P/I	C/H	O	P/I	C/H	O	Leavers
1940		Sum	269	21	2147	109	119	581	350	1625	995	175	2784	826	
0-4	P/I	588	269	5		50		264							314
	C/H	14		14											-
	O	2332		2	2147	29		154							183
5-19	P/I	393				29	24		67	83	189				340
	C/H	75					75								-
	O	454					20	163	53	66	152				271
20-59	P/I	1251							230	161		9	767	83	859
	C/H	1060								1060					-
	O	2264								254	654	14	1209	131	1354
60+	P/I	151										151			-
	C/H	808											808		-
	O	611												611	-
		Arrivals	-	-	-	79	-	418	120	149	341	23	1976	214	3320

NOTE ON SYMBOLS:
P/I -- Parasitic and Infectious Diseases.
C/H -- Neoplasms and Cardiovascular Diseases.
O -- Other Causes of Death Including Unknown.

Columns may not add due to rounding.