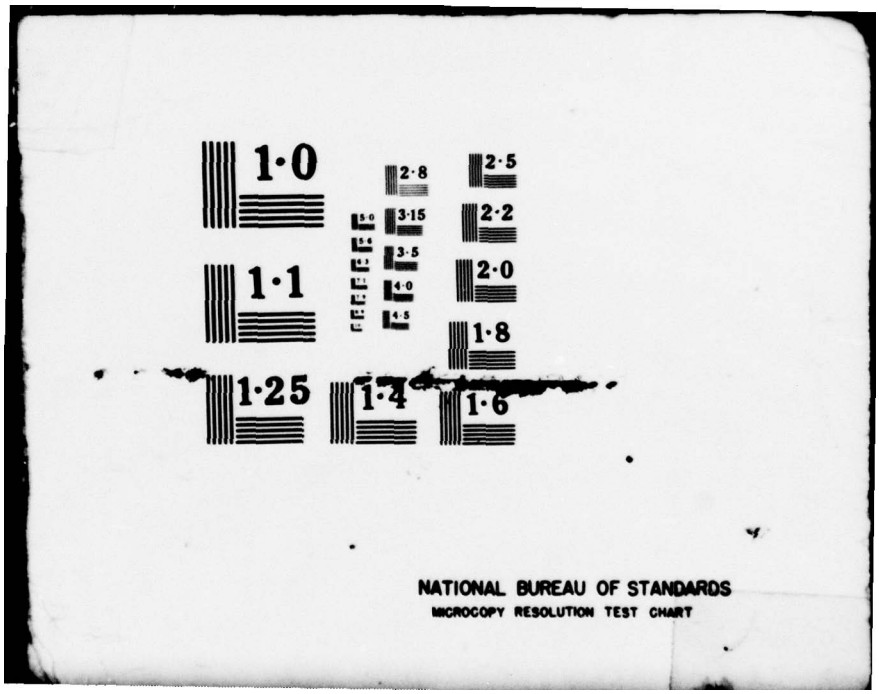
AD-A070 307 PRINCETON UNIV N J DEPT OF STATISTICS F/G 6/16
AN EMPIRICAL HIGHER-RANK ANALYSIS MODEL OF THE AGE DISTRIBUTION--ETC(U)
MAY 78 M B BRECKENRIDGE, J W TUKEY DAAG29-76-G-0298
UNCLASSIFIED TR-143-SER-2 ARO-14244.6-M NL

1 OF 2
AD
A070307





1·0
1·1
1·25
2·8
3·15
3·5
4·0
4·5
2·5
2·2
2·0
1·8
1·4
1·6
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (19) 14244.6-M (18) ARO	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) (6) AN EMPIRICAL HIGHER-RANK ANALYSIS MODEL OF THE AGE DISTRIBUTION OF FERTILITY.	5. TYPE OF REPORT & PERIOD COVERED (9) Technical report.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Mary B. Breckenridge, John W. Tukey (15)	8. CONTRACT OR GRANT NUMBER(s) DAAG29-76-G-0298
9. PERFORMING ORGANIZATION NAME AND ADDRESS Princeton University Department of Statistics Princeton, New Jersey 08540	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (12) 11 p.
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709	12. REPORT DATE (11) May 78 13. NUMBER OF PAGES 108
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

AD A070307

LEVEL II

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(14) TR-143-SER-2

DDC RECEIVED JUN 22 1979 C

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DDC FILE COPY

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

time series model
exploratory data analysis
unifying patterns
distributions of fertility
demographic data
fertility models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Empirical higher-rank (EHR) analysis of a 185-single-year sequence of Swedish age-specific overall fertility rates, and of a related 68-single-year marital fertility sequence, develops a time series model of the changing age distribution of births in cross-sectional and cohort perspectives. The procedure combines the data-guiding and flexibility of Tukey's exploratory data analysis (EDA) approach with robust/resistant methods of estimation to detect unifying patterns underlying the diverse distributions of fertility. The centrality of examination of residuals, in achieving optimal fits to the data, in detecting singular departures from fit, and in interpretation of

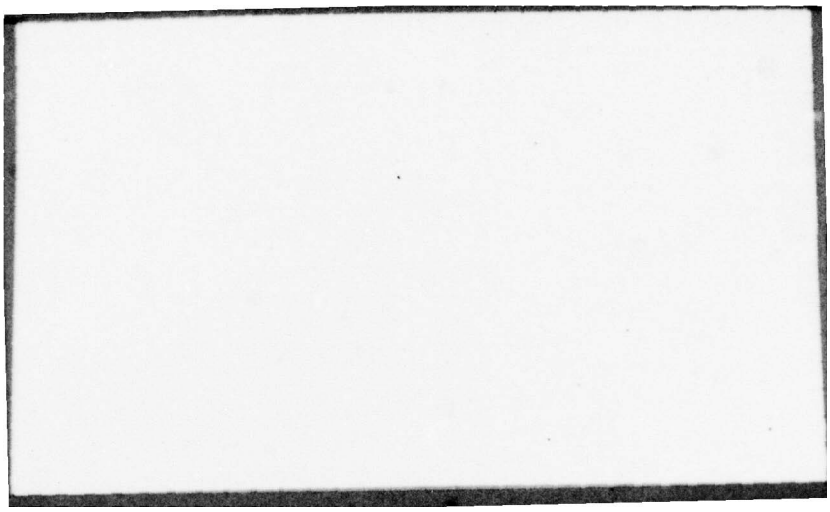
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

406 873 JOB

Unclassified

20. ABSTRACT CONTINUED

the fitted descriptions, is emphasized. Techniques, new in their detailed application to demographic data, are described in some detail.



AN EMPIRICAL HIGHER-RANK ANALYSIS MODEL OF THE AGE DISTRIBUTION OF FERTILITY

by

Mary B. Breckenridge
Department of Statistics
Princeton University

with

THE RELATIONSHIP OF EMPIRICAL ANALYSIS TO MORE NARROWLY MODELLED ANALYSIS

by

John W. Tukey
Princeton University and
Bell Laboratories

Technical Report No. 143, Series 2
Department of Statistics
Princeton University
May 1978

Research for this report was supported in part through a contract with the U. S. Army Research Office, No. DAAG29-76-G-0298, awarded to the Department of Statistics, Princeton University, Princeton, New Jersey.

Abstract

Empirical higher-rank (EHR) analysis of a 185-single-year sequence of Swedish age-specific overall fertility rates, and of a related 68-single-year marital fertility sequence, develops a time series model of the changing age distribution of births in cross-sectional and cohort perspectives. The procedure combines the data-guiding and flexibility of Tukey's exploratory data analysis (EDA) approach with robust/resistant methods of estimation to detect unifying patterns underlying the diverse distributions of fertility. The centrality of examination of residuals, in achieving optimal fits to the data, in detecting singular departures from fit, and in interpretation of the fitted descriptions, is emphasized. Techniques, new in their detailed application to demographic data, are described in some detail.

The close fits obtained for all sequences, coupled with guided choice of a standard form in which to use the fitted descriptions, reduce the variability in the fertility data to a concise and coherent demographic picture which differs in important ways from the descriptions other aggregate fertility models have provided, while having some significant relations to other models. Ways of refining still further the EHR-fitted descriptions to provide additional insight into the underlying structure of aggregate fertility distributions are suggested. Use of this analytic approach to identify and deal with errors in such data is discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

AN EMPIRICAL HIGHER-RANK ANALYSIS MODEL OF THE AGE DISTRIBUTION OF FERTILITY*

Introduction

In efforts to understand the historical decline of fertility in the now industrialized countries and the potential for fertility decline in the less developed countries, a variety of ingenious measures of fertility have been constructed from the available data.[1] Such measures simplify comparisons across time and place, and they can sharpen the exploration of associations between fertility change and social and economic change. The value of multiple approaches applied to data from diverse sources has been repeatedly demonstrated.[2] The problems of dealing with incomplete and error-ridden data continue to be a challenge for new methodology.

While age-marriage duration-specific and age-parity-specific fertility schedules are highly desirable for analysis of fertility change, data needed to calculate such schedules have not been available for most time periods and populations. The relative abundance of simple age-specific fertility schedules (both current and for earlier periods, and usually by

five-year age groups) has encouraged efforts to use, instead, this type of data to construct models containing a small number of meaningful parameters for comparison. Two types of such models, both based on pre-selected patterns, have been prominent. The first of these seeks to express the net maternity function in terms of specific functional forms.[3] The second relates the age distribution of fertility to selected demographic patterns, usually for an idealized population with specified constraints, such as the absence of illegitimate births and the absence of widowhood and divorce in the childbearing years. One of the most prominent demographic models has been that of Coale which then became the basis of the Coale-Trussell model fertility schedules.[4] This model has had wide application and is clearly superior in its descriptive capacity to any of the single function models so far proposed.[5]

The present study of the age distribution of fertility reports an approach complementary to that of Coale and Trussell. It begins without a priori assumptions about the exact shape or the causes of unifying patterns which may underlie diverse age distributions of births in an actual population rather than an idealized one. Using time histories of single-year Swedish overall and marital fertility in both cross-sectional and cohort perspectives, it adopts empirical higher rank (EHR) analysis,[6] including a particular type of re-expression of the fractional fertility up to some cut-off age, and shows

- how well this method of analysis develops a description of the Swedish data;

- how well the EHR results, though quite empirical, fit into a demographic model;
- some ways in which the EHR approach accommodates aspects of fertility distributions that have been problematic for other models when dealing with actual populations.

EHR analysis, as initially developed in McNeil and Tukey, combines

- the emphasis on data-guiding and flexibility, typical of Tukey's exploratory data analysis (EDA) approach,[7] with
- the avoidance of trouble from exotic values, and the relatively high efficiency under any of a wide variety of circumstances, typical of robust/resistant methods of estimation.

The work reported here could be the first steps in a complete EDA of the diverse age distributions of fertility found in time series and across populations.

A. Preliminary considerations

A1. Features of EHR analysis which may make it particularly appropriate to the task of describing fertility distributions.

EHR analysis begins without detailed assumptions about the distribution of the data (in this case, widely varying age patterns of births, reflecting both biological and social factors). Its approach is exploratory, guided at each step by what is left, the residuals, after some additive or multiplicative factor has been removed from one dimension or another of the data or of the residuals from the preceding step of the analysis. This iterative identification of regularities in the data (here, in each of the

two-way tables of fertility distributions in time sequence) is continued until the amount of additional regularity removed in a step seems negligible.

In the effort to identify as fully as possible the patterns which account for the variability in the data, EHR analysis allows for the fact that some data points may be far outside the "true" underlying pattern in the set, because of inaccuracies or singular circumstances. Use of a form of robust/resistant estimation (RRE) in the iterative fitting procedure gives cell-wise weights to the residuals at each iteration before searching for further adjustments to the fit. The choice of weight function is important if one is simultaneously to ensure resistance to "outliers" and also avoid giving undue weight to small residuals. At the present time, one weight function of choice appears to be the bisquare function of the residuals.[8] The iterative fitting procedure using this "biweight" is described in Appendix A.

Re-expression of data to improve linearity before analysis (a practice already well demonstrated as valuable in demographic research) is a usual preliminary to EHR analysis. Without altering the order relations of the members of the data set, this takes advantage of the greater ease of examining linear (here, additive-multiplicative) relations and departures from them. Whether log, reciprocal, power, or more complex re-expression has been used to simplify the data's behavior, de-transformation can readily return results to the original, raw, scale.

Because data are almost always the result of indirect or imperfect measurement, residuals of varying sizes are expected in the analysis. A distinctive feature of the EDA approach is that nothing is discarded. The EHR analysis does not stop with a model and a statement of percent

of the total variation in the data explained by the model. Instead, the residuals are not only examined for further pattern at each stage of the fitting procedure but are also retained and examined in detail at the end, to see where and in what way the data depart from the fitted description (e.g., in specific years? at specific age cuts?). This examination directs both interpretation and further exploratory analysis. It often enhances identification of major transitions in underlying pattern. It may add to understanding of the effects of singular events. (For example, within the context of a population's trend in age distribution of fertility, what effects has a war or a period of increased emigration had on childbearing patterns?) Examination of the residuals may also aid in identifying departures which are due to errors, and then in estimating appropriate corrections. In a complete EDA, such a process may need repetition, particularly after "fine tuning" the expression of the data.

A2. Focus of this EHR analysis of fertility

The biological and social factors acting on fertility may affect its age distribution, its level, or both. Consistent with the current body of demographic work on fertility models, we will focus first on the age distribution --and specifically on the proportion of all births, for a year or a cohort, which is attributable to women a given age and younger. The advantages of using the cumulative distribution, rather than the frequency distribution, include:

- the greater ease of fitting the class of distributions which differ in location and scale;

. the fact that the nature of any systematic bias in the data will be more apparent in a cumulative distribution at the same time that the influence of singular departures will be diminished.

We shall speak throughout of "age cuts" of this cumulative distribution, for example "cut at 24/25" to indicate a separation of the proportion of births to women aged 24 and below from that to women aged 25 and above. In the final sections, the change in overall level of fertility will be related to the parameters of the fitted fertility time sequences.

A3. Choice of data

A time series with diverse fertility distributions allows examination not only of the diversity but also of the dynamics of change. Choice of the Swedish series was based on:

. the length of the series of single-year schedules available;[9]

. the high quality of the data;[10]

. the striking changes in both the age pattern of marriage and the level of fertility, which should be reflected in the age distribution of births;

. the occurrence of singular events (wars,[11] periods of high emigration, periodic severe crop failures) which might be expected to affect temporarily the level and/or the age distribution of births; and

. the rich detail in the recorded marriage and fertility data to aid examination of demographic significance of the model parameters.[12]

To test the use of EHR analysis, 1775-1959 overall fertility and 1892-1959

marital fertility were selected.[13]

Of the demographic factors which affect the age distribution of births, some, in addition to parity-dependent limitation and spacing of births, are prominent for the Swedish data selected. A late marriage pattern prevailed, particularly from the second quarter of the 19th century until 1935.[14] The proportion married at all childbearing ages above 19 then increased dramatically within 15 years (for example, from 0.17 to 0.40 for those aged 20-24, and from 0.48 to 0.72 for those aged 25-29).[15] The proportion aged 15-19 married had nevertheless increased only from 0.02 to 0.05 by 1959. With significant adoption of parity-dependent limitation of births, the pattern of marital fertility decline with age (both in cohort and in cross-sectional perspectives) therefore reflects two influences:

- the tendency for later-marrying women to have low-order births well beyond age 25, and
- the tendency for earlier-marrying women to restrict their childbearing to younger ages.[16]

New prominence of divorce and remarriage in the final years of the 1775-1959 histories may also have influenced the aggregate fertility patterns.

The traditionally large proportion of first births premaritally conceived is of particular importance for the age distribution of marital fertility.[17] Entry into marriage becomes a better proxy for entry into childbearing than is possible with high incidence of post-marital delay of a first birth (due either to contraception or subfecundity). To the extent to which marriage customs select for those of proven fecundity, however,

not only at ages 15-19 but also at ages 20-24 in this late-marrying population, the age distribution of marital fertility shows some distortion toward the lower ages. This may go far in explaining the historical absence in Sweden of a "natural" marital fertility pattern (as this has been identified and defined by Henry[18]) even considering only births to women age 20 and above. Illegitimate fertility has also been non-negligible for the data in the present study. In each of the years from 1892 to 1959, between 9% and 15% of the births contributing to the age pattern and total rate of overall fertility are not included in the age pattern and total rate of marital fertility.

B. Procedure

B1. Preparation of data

Yearly age-specific overall fertility rates were calculated for the years 1775-1959 from recorded confinements[19] by age of mother in five-year age groups, 15-19 to 45-49, and the female population in each of these age groups as this is either recorded in the census or recorded as an intercensal estimate.[20] These schedules were considered in cross-sectional perspective in the 185-year sequence. They were also considered separately for 155 complete cohorts, using the customary procedure of identifying a group of women by the year in which they were aged 15-19, 1775-1929, and following their childbearing by five-year age groups until they reached age 45-49, 1805-1959.[21]

Each schedule was first cumulated to give the fertility rates for women a given age and younger, with age cuts at 19/20, 24/25...49/50. Each schedule was then normalized, i.e., divided by the rate for women

at age cut 49/50, to give the proportion of the year's (or cohort's) births achieved by a given age.[22] Cross-sectional and cohort time sequences of overall fertility for ages 20-49 alone were also prepared for analysis in the same way. For brevity, these four sets of data will be referred to as: X15 and X20 (cross-sectional, age 15-49 and age 20-49, overall); C15 and C20 (cohort, age 15-49 and age 20-49, overall).

For comparative analysis of overall and marital fertility distributions, four time sequences were used: cross-sectional marital fertility for ages 15-49 and ages 20-49 for the years 1892-1959; and the corresponding overall fertility sequences for these 68 years. These data sets will be referred to as MX15, MX20, XX15 and XX20.[23,24]

The several sequences truncated to age 20-49 were included for two reasons:

- to enhance comparisons with those studies and models which exclude age 15-19 fertility, directly, or indirectly through the choice of external standards;
- to test the relative capacity of the EHR approach to describe, empirically and demographically, fertility histories with full variability and histories with some known sources of irregularity removed.[25]

B2. EHR analysis

Preliminary EHR analysis of 20th century U.S. data had suggested that this analytic approach can describe changing overall fertility patterns more or less well in several forms.[26] No consideration was given in this

earlier work, however, to the possible demographic significance of such a description. The present analysis of Swedish overall and marital fertility time sequences tested in detail a variety of combinations of re-expression of data and weighting of residuals in the iterative removal of additive and/or multiplicative components from the data and from successive sets of residuals.[27,28] At this stage of exploration, descriptions were restricted to five parameters.[29] Each combination was tested for goodness-of-fit to the data by two criteria:

- the extent to which the final set of residuals appeared to be free of pattern or structure;
- the proportion of the absolute variation in the data not explained[30]

$$\frac{\sum |z_{ij}|}{\sum |F_{ij} - \text{median } F_{ij}|}$$

where z_{ij} = residual for year i (or cohort i) and age cut j

F_{ij} = cumulated, normalized re-expressed fertility rate for year i (or cohort i) and age cut j

Results are presented here for one effective re-expression-weighting combination (as judged by this pair of criteria) combined with the iterative selection of two time-by-age components for each series.

The selected re-expression is the folded square root of the cumulated, normalized fertility schedule[31]

$$F_{ij} = (f_{ij})^{\frac{1}{2}} - (1 - f_{ij})^{\frac{1}{2}}$$

This centers the distribution on the mean of the schedule and stretches and straightens both ends of the cumulated fertility distribution's sigmoid

configuration. The values for a given schedule are thus transformed from the range (0.0 to 1.0) to the range (-1.0 to 1.0) for analysis.

Setting $c = 12$ in the biweight function used in the robust/resistant estimation (so that residuals whose absolute value is greater than 12 times the median absolute value of residuals are given zero weight) proves to be an appropriate choice. This includes in the fitted parameters a high proportion of the systematic variability in the data, while still avoiding difficulties with "outliers."

The description developed of each two-way fertility table can be expressed:

$$F_{ij} = \alpha_i A_j + \beta_i B_j + z_{ij}$$

where F_{ij} = re-expressed fertility rate for year i (or cohort i) and age cut j

$\sum A_j^2 = \sum B_j^2 = 1$ (for standardization)

z_{ij} = residual for year i (or cohort i) and age cut j

The order of fitting for this expression emphasizes the patterns in the age dimension of the fertility matrix:

- the iterative selection of a central value for the distribution of fertility within each age cut across time, the formation of an age vector from these six "biweight centers", and determination of this vector's variability in the time dimension;
- from the residuals, iterative selection of a second central value for each age cut, formation of a second age vector from these

six "biweight centers", and determination of this vector's variability in the time dimension;

- iterative improvement of the two age vectors (and their scalar multipliers) from successive sets of residuals until convergence occurs for A_j and B_j.[32]

Since EHR analysis places few constraints on the fitting procedure (for example, not requiring that the age vectors be orthogonal) it is possible for the scalar multipliers of the iteratively selected vectors to be more or less linearly related. This occurs when the distribution shows appreciable variation at the first age cut (as with the X20, MX15 and MX20 time sequences or in cross-population comparisons)[33] in contrast to variation largely confined to the central age cuts (as with the X15 sequence where age 15-19 fertility is always low). This outcome gives emphasis to the importance of standard form re-presentations of fits as discussed below, page 26.

In trying to develop as sharp as possible a mathematical description of the regularities underlying the data's variability, one starts with EHR analysis. Knowledge of the data directs re-presentation of the selected fit in a standard form; the re-presented fit then directs further analysis and interpretation.

C. How well has the EHR analysis developed descriptions of the Swedish fertility time sequences?

The motivation of this study of the age distribution of fertility was exploratory--and successful exploration relies heavily on examination of

the residuals, not only their size, but more importantly their distribution. Later, interpretation of parameters is also enhanced by attention to idiosyncracies in the residuals. Our discussion therefore starts with the residuals, using those from the X15 and C15 histories to illustrate some of the considerations.

C1. Size of residuals

Size of the residuals is mentioned first, only because it will be important for the reader to have in mind how relatively small, in these cases, the residuals are that are then being examined for structure. The familiar expression of proportion of squared variation explained:

$$1 - \frac{\sum (z_{ij})^2}{\sum (F_{ij} - \text{mean } F_{ij})^2}$$

showed fits of 99.96-99.99% to be not uncommon for these fertility histories when comparing various combinations of re-expression and fitting conditions. All fits are reported here (Table 1) in terms of the less extreme, and somewhat more useful proportion of the linear variation explained, based on sum of the absolute deviations

$$1 - \frac{\sum |z_{ij}|}{\sum |F_{ij} - \text{median } F_{ij}|}$$

One reason for the greater usefulness is reduced attention to a few large residuals, likely to come from specific events or specific errors in data collection. In slightly different contexts, it may be desirable to use even more resistant measures of quality of fit.[34, 35]

Reported and fitted distributions, for representative years with very different age distributions of fertility[36] and for some years of least good fit, are given in Table 2 on both re-expressed and raw fraction scales.[37] Except at two of the 1110 cross-sectional age cuts and 63 of the 930 cohort age cuts, the difference between reported and fitted values is in no more than the third decimal place on the raw fraction scale. Even working with the full marital fertility distribution, including age 15-19 births, results in very close fits.[38, 39]

C2. Structure of residuals

In practice, examination of the residuals for structure should come first in deciding between various combinations of re-expression and fitting conditions. For data in time sequence, large residuals with truly irregular distribution suggest

- the presence of errors in the data, or
- the impact of singular events.

At another extreme, large and highly patterned residuals, perhaps for the end of a time sequence after an extended period of very close fit, indicate a major transition in the underlying patterns, and suggest the need for

- alteration in the specified combination of re-expression and fitting, to accommodate the new pattern also, and/or
- division of the series at that period to analyze the portions separately.

Table 1. Features of Residuals from EHR Fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the Age Distribution of Fertility, Expressed on the Folded Square Root Scale, in Selected Time Sequences: 1775-1959

Sequence	% Variation Explained (linear scale)	s^2_{bi}	Residuals Median Absolute	Upper 90%
X15	99.0	4.14×10^{-5}	.0046	.0101
X20	98.9			
C15	98.5	8.07×10^{-5}	.0057	.0160
C20	98.3			
MX15	98.9	1.76×10^{-5}	.0026	.0072
MX20	99.1			

Table 2. Reported and EHR Fitted Fertility Distributions on Re-expressed and Raw Fraction Scales: Selected Sequences and Selected Years, 1775-1959

Source of distribution	Age Cut 19/20	24/25	29/30	34/35	39/40	44/45
X15 1890						
Re-expressed						
Reported	-0.8774	-0.5528	-0.1774	0.1869	0.5424	0.8536
Fitted	-0.8781	-0.5579	-0.1801	0.1933	0.5467	0.8467
Residual	0.0007	0.0051	0.0027	-0.0064	-0.0043	0.0069
Raw fraction						
Reported	0.0134	0.1402	0.3755	0.6310	0.8542	0.9813
Fitted	0.0133	0.1375	0.3737	0.6354	0.8565	0.9795
Residual	0.0001	0.0027	0.0018	-0.0044	-0.0023	0.0018
X15 1950						
Re-expressed						
Reported	-0.6663	-0.2035	0.1929	0.5131	0.7709	0.9412
Fitted	-0.6720	-0.2029	0.2023	0.5189	0.7682	0.9376
Residual	0.0057	-0.0006	-0.0094	-0.0058	0.0027	0.0036
Raw fraction						
Reported	0.0844	0.3576	0.6352	0.8381	0.9570	0.9967
Fitted	0.0819	0.3580	0.6416	0.8414	0.9561	0.9963
Residual	0.0025	-0.0004	-0.0064	-0.0033	0.0009	0.0004
X15 1792 (a "worst fit")						
Re-expressed						
Reported	-0.8410	-0.5345	-0.1789	0.1829	0.5610	0.8267
Fitted	-0.8516	-0.5378	-0.1697	0.1928	0.5352	0.8253
Residual	0.0106	0.0033	-0.0092	-0.0099	0 0258	0.0014
Raw fraction						
Reported	0.0219	0.1501	0.3745	0.6283	0.8641	0.9743
Fitted	0.0193	0.1483	0.3809	0.6351	0.8503	0.9739
Residual	0.0026	0.0018	-0.0064	-0.0068	0.0138	0.0004

Table 2 - cont'd.

Table 2 - cont'd.

C15 1890						
Re-expressed						
Reported	-0.8685	-0.5007	-0.0854	0.2947	0.6352	0.8947
Fitted	-0.8655	-0.4925	-0.0875	0.2890	0.6308	0.9047
Residual	-0.0030	-0.0082	0.0021	0.0057	0.0044	-0.0100
Raw fraction						
Reported	0.0153	0.1689	0.4397	0.7038	0.9013	0.9899
Fitted	0.0160	0.1735	0.4383	0.7001	0.8992	0.9917
Residual	-0.0007	-0.0046	0.0014	0.0037	0.0021	-0.0018
C15 1920						
Re-expressed						
Reported	-0.7256	-0.2852	0.1075	0.4238	0.6934	0.9342
Fitted	-0.7264	-0.2926	0.1067	0.4327	0.7033	0.8837
Residual	0.0008	0.0074	0.0008	-0.0089	-0.0099	0.0505
Raw fraction						
Reported	0.0596	0.3025	0.5758	0.7859	0.9273	0.9959
Fitted	0.0593	0.2976	0.5752	0.7913	0.9314	0.9878
Residual	0.0003	0.0049	0.0006	-0.0054	-0.0041	0.0081
C15 1929 (a "worst fit")						
Re-expressed						
Reported	-0.7625	-0.3925	0.0023	0.4608	0.7777	0.9480
Fitted	-0.7681	-0.2969	0.1314	0.4772	0.7616	0.9472
Residual	0.0056	-0.0956	-0.1291	-0.0164	0.0161	0.0008
Raw fraction						
Reported	0.0459	0.2334	0.5016	0.8081	0.9593	0.9974
Fitted	0.0440	0.2947	0.5925	0.8176	0.9538	0.9974
Residual	0.0019	-0.0613	-0.0909	-0.0095	0.0055	0.0000

Table 2 - cont'd.

Table 2 - cont'd.

MX15 1905

Re-expressed						
Reported	-0.2868	0.0298	0.2789	0.5023	0.7130	0.9081
Fitted	-0.2888	0.0325	0.2789	0.4983	0.7129	0.9094
Residual	0.0020	-0.0027	0.0000	0.0040	0.0001	-0.0013
Raw fraction						
Reported	0.3014	0.5211	0.6933	0.8320	0.9354	0.9923
Fitted	0.3001	0.5230	0.6934	0.8298	0.9354	0.9925
Residual	0.0013	-0.0019	-0.0001	0.0022	0.0000	-0.0002

MX15 1930

Re-expressed						
Reported	-0.1107	0.2162	0.4395	0.6215	0.7903	0.9349
Fitted	-0.1089	0.2091	0.4386	0.6280	0.7941	0.9288
Residual	-0.0018	0.0071	0.0009	-0.0065	-0.0038	0.0061
Raw fraction						
Reported	0.4219	0.6511	0.7954	0.8947	0.9634	0.9960
Fitted	0.4232	0.6462	0.7948	0.8979	0.9646	0.9953
Residual	-0.0013	0.0049	0.0006	-0.0032	-0.0012	0.0007

MX15 1955 (a "worst fit")

Re-expressed						
Reported	-0.0345	0.2999	0.5397	0.7211	0.8652	0.9620
Fitted	-0.0254	0.3029	0.5328	0.7147	0.8635	0.9731
Residual	-0.0091	-0.0030	0.0069	0.0064	0.0017	-0.0111
Raw fraction						
Reported	0.4756	0.7072	0.8528	0.9386	0.9839	0.9986
Fitted	0.4820	0.7092	0.8490	0.9361	0.9835	0.9993
Residual	-0.0064	-0.0020	0.0038	0.0025	0.0004	-0.0007

Frequently encountered, also, are residuals of intermediate size and irregular enough distribution that existing pattern isn't readily seen. Then, detailed examination of the residuals is needed

- to bring out the hidden regularities, and
- to indicate what change in data re-expression or fitting will accommodate the detected additional pattern.

Two useful procedures are scatter plots, illustrated below (see page 20), and diagnostic plots.[40]

In these examples of irregular and patterned residuals, the sum of the absolute deviations from fit may actually be the same; it is the location of deviations that determines the analyst's response and leads to better understanding of the data. Figure 1 shows a plot of residuals which suggests alteration in specified re-expression and fitting, and also leads to consideration of the period around 1910-1920 as a major transition in the fertility history. Identified structure in even very small residuals, as in the present report, can suggest directions to move in seeking still sharper descriptions of the data.

C3. Ways of looking at residuals

Of a number of informative ways of looking at residuals, three are illustrated here for the X15 and C15 analyses:

- schematic plots,
- scatter plots for pairs of age cuts, and
- time sequence plots for each age cut.

They demonstrate, in different ways, the extent to which the developed descriptions have captured the underlying structure of the fertility histories.

(a) Schematic plots order, by size, the residuals for each age cut, and summarize their distribution within an age cut. These plots give a convenient picture of the degree of symmetry in the spreading of residuals around the median residual for each age cut, and the degree of agreement between age cuts in the amount of spreading of the residuals--both indicators of the extent to which pattern remains in the residuals.[41] For C15 (Fig. 2), 90% of the residuals (all of those within the range from one interquartile distance of the upper quartile to one interquartile distance of the lower quartile) are quite similarly and evenly distributed for all age cuts; however, the outliers, indicated by the open and shaded circles, are predominantly positive for the first two and last two age cuts, predominantly negative for the two central age cuts. In contrast, the X15 residuals (Fig. 3) are more compact. While symmetrically distributed around the median for the first four age cuts, they show dissimilarity between age cuts in the amount of spread, however, particularly in the interquartile range. At even this early stage of examination, one can conclude, then, that

- some structure remains in both cohort and cross-sectional residuals;
- what structure is most visible lies in the outliers for the cohorts, but appears in the main body of the residuals for the cross-sectional series.

(b) Scatter plots summarize the size and sign relations of residuals

for pairs of age cuts. They use the residuals, ordered in time sequence for any age cut, and plot them against the residuals ordered in time sequence for each other age cut. This reveals any tendency for the residuals for a pair of age cuts not to vary randomly with each other, but to vary together by year in some systematic way--for example, to have the same sign and magnitude for any selected year. To illustrate, the X15 residuals for age cut 24/25

- show no systematic variation over time with those for age cut 29/30 (Fig. 4A),
- tend to be of opposite sign by year from those for age cut 39/40 (Fig. 4B),
- tend to be of the same sign by year as those for age cut 44/45 (Fig. 4C).[42]

(c) A time sequence plot of residuals by age cut can provide more detail about remaining structure. We present such plots here with the residuals expressed on the raw fraction scale so that the relative importance of any residual can be judged directly. To emphasize the nature of the small deviations, the residuals have also been coded

0	—	+
< (M - 1/4 I)	(M - 1/4 I) to (M+1/4 I)	> (M+1/4 I)

where M = median residual across all years and age cuts

I = range of values between the residual at the lower 25% point and the residual at the upper 25% point.

When the 1110 residuals for X15 are ordered by value alone, I is then

the range between the value of residual number 278 and the value of residual number 832. Coding values are shown in Table 3 for all three sets of residuals to be examined here in time sequence plots.

For the X15 residuals, such a plot (Fig. 5) confirms the schematic plot impression of few singular departures from fit at any age cut. The years 1783 and 1792 stand out as deviants. The plot also locates in the time dimension some departures from random distribution which were detected in the scatter plots of the pairs of X15 residual vectors. Age cuts 34/35 and 39/40 tend to have residuals above the central interval before 1855, then below the central interval until about 1937. Age cuts 24/25 and 44/45 show the reverse, i.e., a tendency for residuals below the central interval before 1855, above the central interval until about 1937. This means, for example, that before 1855 a slightly higher proportion of births is reported to have occurred between the ages of 30 and 40 in most years than the fit would have predicted. Residuals by age cut for X20 have the same systematic variations as those for X15, indicating that departure from fit is not greatly influenced by age 15-19 fertility.

The lesser accuracy of the data in the early years of the series may contribute to the observed residual pattern. The distribution of residuals suggests also that a more complex model may remove the remaining pattern.[43] In the present report we want to show how much can be learned from a relatively simple EHR description of the fertility time history. The

Table 3. Basis of Coding of Raw Fraction Scale Residuals in Time Sequence Plots (Figs. 5, 6 and 7)

Sequence	Median residual	Lower 25% point	Upper 25% point	I	M-1/4 I	M+1/4 I
X15	-.000058	-.00214	.00191	.00404	-.00107	.00095
C15	-.000057	-.00215	.00256	.00471	-.00124	.00112
M15	.000047	-.00099	.00121	.00220	-.00050	.00060

decision on whether to go to a more complex model seeks appropriate balance between parsimony and completeness of description. A model sufficiently complex to give an excellent mathematical description of almost all variation in a data set can usually be found, but is at least somewhat more likely to have lost demographic significance of its parameters.[44]

Time sequence plots of residuals by age cut for C15 (shown in Fig. 6 in coded form on the raw fraction scale) present a very different picture from that of X15 residuals. When the vector of residuals for any age cut (e.g. 29/30) is compared to the vector of residuals for the next age cut (34/35 in this example) shifted forward by five years, the fluctuations appear highly correlated. This visual impression is reinforced by calculated correlation coefficients (.63-.83) for the pairs of lagged residual vectors on the folded square root scale used in the fitting. The fluctuations are less pronounced, however, for about a 40-year span (e.g. from 1870 to 1910 for cohorts at age cut 24/25).

Residuals by age cut for C20 have the same pattern as that observed for C15. Residuals from EHR analysis of the cohort marital fertility history for the last 38 cohorts in the overall fertility history, using the same combination of re-expression and fitting, also exhibit the lagged pattern.

At least two questions about these systematic variations need to be explored:

. Has the use of cross-sectional data in five-year age groups to

construct the cohort histories contributed to the variation?[45]

- Does the lagged pattern represent a period effect on cohort distribution--either the influence of real events which have affected the childbearing of all age groups to some extent in particular years, or the dissemination of cross-sectional errors to a number of cohorts?[46]

Further exploration indicates that other re-expressions of the data coupled with a more complex model can bring at least part of the lagged pattern from the residuals into the fitted components when this is desirable.[47]

The very small residuals for MX15 (shown in Fig. 7 in coded form on the raw fraction scale) and for MX20 include no singular departures from fit, but some tendency at all age cuts for stretches either above or below the central interval.

C4. Conclusions about residuals

1. Some structure remains in the small residuals of each of the overall fertility time histories, X15, X20, C15 and C20. Extensions of this EHR analysis would naturally ask:

- Are more complex descriptions of fertility distributions appreciably better?
- Do other re-expressions of the data result in still better fits?

2. The description of each of the histories is sufficiently good to proceed to

- re-presentation of the fits in a standard form, and

. examination and interpretation of the resulting parameter vectors which describe the major regularities inherent in each of the histories.

D. Re-presentation of fits in a standard form

As suggested above (see page 12) the freedom from most constraints in the EHR fitting procedure means that the final fitted representation of the data is not unique but is one of many exactly equivalent linear combinations of the parameters. To direct further analysis and to permit comparison and interpretation of the fitted parameters of different fertility sequences, it is important to re-present each rank-two fit[48]

$$\alpha_i A_j + \beta_i B_j$$

as an identically equivalent standard form constrained rank-two fit

$$K_1 \alpha_i^{**} A_j^{*} + K_2 \beta_i^{**} B_j^{*}$$

where $\alpha_i^{**} = a\alpha_i + b\beta_i$,

$A_j^{*} = cA_j + dB_j$,

a, b, c, d, K_1, and K_2 are constants,

$\sum A_j^{*2} = \sum B_j^{*2} = 1$ for standardization

and the total number of constraints equals 4 to ensure uniqueness.

For example, for the X15 history, the fitted rate (as expressed on the folded square root scale) at age cut 34/35 in 1910 can be presented in the identically equivalent forms:

$$F_{136,4} = \alpha_{136} A_4 + \beta_{136} B_4$$

$$= (1.4548)(.1620) + (.0890)(.4772)$$

$$= .2781$$

$$F_{136,4} = K_1 \alpha^{**}_{136} A^{*}_4 + K_2 \beta^{**}_{136} B^{*}_4$$

$$= (1.4759)(.9995)(.1170) + (.9992)(.2222)(.4754)$$

$$= .2781$$

Constraints for standard forms of rank-two fits are of three types:

- "fixing" a vector (e.g. to be a constant or to be linear) or requiring it to be as nearly as possible of a given form;
- making two vectors orthogonal;
- maintaining absence of a mixed row-by-column term, $\alpha_i B_j$ or $\beta_i A_j$.

The four constraints chosen here for a standard form re-presentation of the fitted fertility distributions were:

(a) to make $\alpha^{**}_i = a\alpha_i + b\beta_i$ as nearly constant as possible, by LS regression (with 0 intercept) of a vector of repeating constants on α_i and β_i. The two regression coefficients then equal a and b.

(b) to make $A^{*}_j = cA_j + dB_j$ as nearly linear as possible, by a canonical regression of two linearly independent straight lines on A_j and B_j.[49] The two elements of the eigenvector associated with the largest canonical correlation are then equal to c and d. (A_j and B_j for all of the time sequences will be found in Appendix B, Table B1.)

(c) and (d) to maintain diagonality, i.e., the absence of both $\alpha_i^{**} B_j^{*}$ and $\beta_i^{**} A_j^{*}$ in the expression.

Calculation (from A_j^{*}, $\alpha_i^{**} B_j$, β_i, a, b, c, and d) of the β_i^{**} and B_j^{*} to meet these requirements is shown in Appendix B. Detailed consideration of alternative standard form re-presentations in the two-way case will be found in unpublished work of Tukey.[50]

For all of the fertility time sequences the regressions to fix α_i^{**} are quite clear, with a fit of at least .983 based on sum of the absolute deviations from a vector of constants (Appendix B, Table B2). The greater degree of departure from constancy for the C15, C20, and MX15 sequences is attributable to a small number of members of each sequence (see Figs. 8 and 15 below, pages 86 and 93).

The canonical correlation analysis to fix A_j^{*} close to linearity produces eigenvalues of at least .999 for the first canonical variate for each time sequence (Appendix B, Table B2). This means that each of the A_j^{*} vectors, composed of A_j and B_j in proportion to the corresponding elements of the first eigenvector, departs from linearity by no more than 0.1%.

The "strength" of a constraint might be considered its relative immunity to sampling/measurement fluctuations as transmitted by the fitting process. It is possible that a different set of constraints with as great or greater overall strength could have been chosen (e.g. by fixing one vector exactly, instead of as close as possible to linearity: or by making one vector orthogonal to a fixed one, instead of fixing two vectors;

or by formal orthogonality, with $A_j^* \perp B_j^*$ and $\alpha_i^{**} \perp \beta_i^{**}$). There is some reason to expect, however, that demographic interpretation of parameters would be more difficult in many of these cases.

For comparisons between time sequences, we incorporate the constants K_1 and K_2 in the time parameters:

$$\alpha_i^* = K_1 \alpha_i^{**}, \quad \beta_i^* = K_2 \beta_i^{**}$$

so that each of the elements of the ith fitted fertility distribution is expressed as

$$F_{ij} = \alpha_i^* A_j^* + \beta_i^* B_j^*$$

The age vectors A_j^* and B_j^* for all of the time sequences will be found in Appendix B, Table B3.

E. Relations between the age distribution of births and the components of a standard form EHR fit

We are now ready to look closely at the two components of a fit, $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$, viewing each as a vector composed of one element for each age cut. We ask what changing each component vector, by varying α_i^* or β_i^*, does to the age distribution of births.

The fitted X15 history, which includes all births and all women aged 15-49, is selected for this examination. The extreme values which the elements of each component vector assume for this sequence are shown in Table 4.

Table 4. Range of Contribution of Components $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$ to the Fitted Description of the X15 Sequence, Folded Square Root Scale

Time Parameter	Component	Age Cut					
α_i^*	$\alpha_i^* A_j^*$	19/20	24/25	29/30	34/35	39/40	44/45
Low 1.420		-.8543	-.5559	-.1953	.1661	.5112	.8066
Median 1.477		-.8886	-.5782	-.2031	.1728	.5317	.8389
High 1.518		-.9132	-.5943	-.2087	.1776	.5465	.8622
β_i^*	$\beta_i^* B_j^*$						
Low -.0386		-.0111	-.0196	-.0213	-.0184	-.0127	-.0056
High .8427		.2419	.4276	.4656	.4006	.2773	.1212

* * * * * * * * * * * * * * * *

Table 5. Age-Specific Fertility Distributions (raw scale) Based on the Age Parameters A_j^* and B_j^* for X15 and Increments in the Time Parameters α_i^* and β_i^*

α_i^*	1.47	1.47	1.47	1.47	1.42	1.52	Change by 0.1
β_i^*	Change by 0.1	0	0.1	0.05	0.05	0.05	0.05
Age Group	Change in f(a)	f(a)	f(a)	f(a)	f(a)	f(a)	Change in f(a)
15-19	.0063	.0120	.0183	.0150	.0222	.0091	-.0131
20-24	.0209	.1163	.1372	.1267	.1301	.1222	-.0079
25-29	.0110	.2303	.2413	.2359	.2301	.2415	.0114
30-34	-.0055	.2622	.2567	.2596	.2508	.2683	.0175
35-39	-.0150	.2263	.2113	.2188	.2131	.2244	.0113
40-44	-.0140	.1295	.1155	.1224	.1244	.1196	-.0048
45-49	-.0037	.0235	.0198	.0216	.0294	.0149	-.0145

The $\alpha^*_i A^*_j$ vector represents (on the folded square root scale used in the fitting) an underlying pattern of cumulation of births by age. It is centered on zero at the median age (here, 32.7 years) that the maternity schedule would have if no $\beta^*_i B^*_j$ were added. The more negative the value of $\alpha^*_i A^*_j$ at an age cut below the median age, the smaller the proportion of total births included by that age cut; the more positive the value of $\alpha^*_i A^*_j$ at an age cut above the median age, the higher the proportion of total births included by that age cut.

The addition of the $\beta^*_i B^*_j$ vector makes a non-linear alteration in the cumulative maternity schedule expressed in $\alpha^*_i A^*_j$. The result is a change in shape, and in median age, of the schedule. A positive value of β^*_i lowers the median age, a negative value raises it.

To see readily the effects of change in α^*_i and/or β^*_i we

- construct a series of fertility distributions based on A^*_j and B^*_j, and uniform increments in α^*_i and β^*_i;
- de-transform these distributions from their folded square root scale to the raw fraction scale;
- express the distributions in non-cumulated form (Table 5).

A positive change in β^*_i decreases the proportions of births in the four highest age groups, increases the proportions in the three lowest. On the other hand, a positive change in α^*_i increases the proportions in the central part of the childbearing years at the expense of the two lowest and two highest age groups. Since α^*_i is a near-constant, any shift of

births toward or away from the central ages is limited to a narrow range. In fact, the full range for the X15 sequence is covered in Table 5.[51]

To understand still better what sort of description we are developing, let us shift our emphasis from the births, the outcome, to the distribution of exposure to this outcome. If we consider each woman as a potential activity state, and a birth as evidence of a functioning activity state, the systematic variations expressed in $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$ seem natural ones for fertility distributions. Whatever proportion of all potential activity states is functioning, this activity could be distributed evenly across age groups if women of all ages had equal opportunity of becoming active. The proportion of women both biologically susceptible and exposed to the possibility of pregnancy increases over the earliest childbearing ages, however, and declines over the highest ages. A first natural variation from uniform distribution is, therefore, some degree of concentration of functioning activity states around the median age. This is accommodated in $\alpha_i^* A_j^*$.[52]

A second natural variation in the age distribution of active states is asymmetry or skewness. High incidence of delay in the onset of functioning of potential activity states and/or extension of functioning to higher ages would contribute negatively to skewness; high incidence of early onset of functioning and/or curtailment of functioning at higher ages would contribute positively to skewness. The net effect of these opposed influences on the asymmetry of the distribution is accommodated in $\beta_i^* B_j^*$.

Diversity between women in the age patterns of activity may contribute

to a third natural type of variation in overall fertility distributions--a more even spreading of functioning activity states across a portion of the reproductive span. A simple example might be the aggregate distribution for two sub-populations, one of which attempts to limit its childbearing to early ages, while the other delays all childbearing to higher ages.[53] This third type of variation is explored in later work which fits a third component to the fertility time history.[54]

We proceed now to examine the two components $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$ developed in the present work for each of the Swedish time sequences analyzed. The next two sections concentrate on the 68-year sequences for marital and overall fertility. We ask what the components indicate about the distribution of functioning activity states and about changes in their distribution in this population.

F. Some demographic implications of the fertility components derived by EHR analysis (examination of the 1892-1959 data)

We approach the question of demographic significance in several stages:

- examine the distinctive features of the EHR components of MX15, MX20, XX15 and XX20;
- compare the EHR model parameters for marital fertility with the parameters of the Coale marital fertility model, giving particular attention to the "natural fertility" standard and to the measure of "degree of control";
- dissect marital fertility further, to examine separately the changes

in distribution of functioning activity states not associated with change in level of fertility;

- dissect from the overall distribution of functioning activity states that portion of change not associated with change in the distribution of marital fertility;
- derive level-compensated distributions of functioning activity states in overall and marital perspectives;
- approximate age-specific proportions of married plus cohabiting active women from these level-compensated distributions.

F1. EHR components of marital and overall fertility histories

For each of the 68-year marital and overall fertility sequences, the first component, $\alpha_i^* A_j^*$, reveals the median age

MX15	30.3
MX20	31.0
XX15	32.6
XX20	33.5

that the cumulated maternity schedule would have if none of the component $\beta_i^* B_j^*$ were added.

Sample fertility distributions constructed from the components $\alpha_i^* A_j^*$ (with α_i^* equal to the mean of its near constant value for each sequence) and $\beta_i^* B_j^*$ (with increasing values of β_i^*) can be examined in Table 6 for two of the sequences, MX15 and XX15.[55] The distributions are expressed in three forms: the cumulated form on the folded square root scale used in the fitting; and both cumulated and non-cumulated forms on the raw scale.

The extent to which the distribution of functioning activity states for

Table 6. Fertility Distributions Constructed from the Components $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$ for Selected Marital and Overall Fertility Sequences, 1892-1959

Sequence MX15 Components	Age Cut 19/20	24/25	29/30	34/35	39/40	44/45
Folded Square Root Scale						
$1.226A_j^*$	-.5903	-.2743	-.0071	.2565	.5477	.8442
$1.226A_j^* + 0.2B_j^*$	-.4892	-.1715	.0886	.3374	.6028	.8657
$1.226A_j^* + 0.6B_j^*$	-.2869	.0341	.2802	.4990	.7130	.9088
$1.226A_j^* + 1.0B_j^*$	-.0847	.2397	.4717	.6607	.8233	.9518
Raw Scale, Cumulated						
$1.226A_j^*$	.1207	.3097	.4950	.6784	.8570	.9789
$1.226A_j^* + 0.2B_j^*$	.1755	.3797	.5626	.7317	.8856	.9841
$1.226A_j^* + 0.6B_j^*$	.3013	.5241	.6942	.8302	.9354	.9924
$1.226A_j^* + 1.0B_j^*$	.4402	.6671	.8144	.9131	.9733	.9978
	Age Group 15-19	20-24	25-29	30-34	35-39	40-44	45-49
Raw Scale, Noncumulated							
$1.226A_j^*$	.1207	.1890	.1853	.1834	.1786	.1219	.0211
$1.226A_j^* + 0.2B_j^*$	.1755	.2042	.1829	.1691	.1539	.0985	.0159
$1.226A_j^* + 0.6B_j^*$	.3013	.2228	.1701	.1360	.1052	.0570	.0076
$1.226A_j^* + 1.0B_j^*$	.4402	.2269	.1473	.0987	.0602	.0245	.0022

Sequence XX15 Components	Age Cut 19/20	24/25	29/30	34/35	39/40	44/45
Folded Square Root Scale						
$1.468A_j^*$	-.8792	-.5544	-.1897	.1701	.5289	.8546
$1.468A_j^* + 0.2B_j^*$	-.8232	-.4575	-.0797	.2697	.5983	.8795
$1.468A_j^* + 0.6B_j^*$	-.7112	-.2637	.1404	.4690	.7370	.9293
$1.468A_j^* + 1.0B_j^*$	-.5991	-.0698	.3605	.6683	.8757	.9791
Raw Scale, Cumulated						
$1.468A_j^*$	.0131	.1394	.3671	.6194	.8469	.9815
$1.468A_j^* + 0.2B_j^*$	.0267	.1939	.4438	.6872	.8833	.9870
$1.468A_j^* + 0.6B_j^*$	.0653	.3168	.5988	.8129	.9448	.9953
$1.468A_j^* + 1.0B_j^*$	.1163	.4507	.7465	.9165	.9862	.9996
	Age Group 15-19	20-24	25-29	30-34	35-39	40-44	45-49
Raw Scale, Noncumulated							
$1.468A_j^*$	.0131	.1263	.2277	.2523	.2275	.1346	.0185
$1.468A_j^* + 0.2B_j^*$	.0267	.1672	.2499	.2434	.1961	.1037	.0130
$1.468A_j^* + 0.6B_j^*$	.0653	.2515	.2820	.2141	.1319	.0505	.0047
$1.468A_j^* + 1.0B_j^*$	.1163	.3344	.2958	.1700	.0697	.0134	.0004

Note: α_i^* is set at its mean value for each sequence.

each sequence shows some underlying concentration in the central portion of the reproductive span is most easily examined in the raw scale non-cumulated distributions implied by $\alpha_i^* A_j^*$ alone with no $\beta_i^* B_j^*$ added. With addition of $\beta_i^* B_j^*$, the decline in median age of functioning activity states is clear in the cumulative distributions on either scale. For XX15, increments in β_i^* over the observed 0.04 to 0.82 range results in more than a five-year decline in this median age. The degree of asymmetry contributed to the distribution by $\beta_i^* B_j^*$ at each increment of β_i^* reflects the net imbalance in the forces which discourage or intensify the functioning of potential activity states at lower and higher ages. Change in asymmetry may or may not be accompanied by change in the level of fertility; and conversely, change in level may or may not be accompanied by change in asymmetry.

The limited year to year variation in the time parameter α_i^* resulting from re-presentation of fits for the MX15, MX20, XX15 and XX20 histories can be examined in Figs. 8 and 9. While the α_i^*'s differ in level by sequence, the small departures of α_i^* from constancy are similar for all sequences. The most prominent variation is the peak at 1942-45. The MX15 and XX15 α_i^*'s also show a slight rise after 1950 not seen for the MX20 and XX20 histories.

Year to year variations in the skewness parameter β_i^* for the four time sequences include some distinctive features (Figs. 8, 9):

- a lower rate of increase in skewness of the distribution toward the younger ages for all four sequences before 1920, higher for

all four between 1920 and 1935;

- a divergence of marital from overall fertility skewness patterns between 1935 and 1950--the continued rise in positive skewness for XX15 and XX20 while skewness for MX15 and MX20 fluctuates.

What would these results mean in demographic terms? If the degree of positive skewness of the marital fertility distribution is considered to be due mainly to a decline in functioning activity states at higher ages, the fitted parameter suggests acceleration of this process for 15 years, then leveling or even some regression for a 15-year period before resumption of the trend. If the degree of positive skewness of the overall fertility distribution is due mainly to the combined effects of limitation of marital fertility at higher ages and the proportion married and/or entering childbearing at the younger ages, the divergence of overall β_i^* from marital β_i^* has picked up the significant increases in age-specific proportions married which occurred between 1935 and 1950. These possibilities are examined below at several levels of detail.

F2. Comparison of the EHR and Coale descriptions of marital fertility

For a first broad examination of demographic significance of the EHR-derived marital fertility parameters, comparison with a model whose parameters are considered to have demographic meaning may be useful. We shall therefore compare the parameters of the EHR model of marital fertility

$$F_{ij} = \alpha_i^* A_j^* + \beta_i^* B_j^*$$

with those of the Coale model of marital fertility

$$r(a) = Mn(a)e^{m \cdot v(a)}$$

which becomes, with omission of M, the core of the Coale-Trussell model fertility schedules for an idealized population.

Each model of marital fertility concentrates on the age pattern of fertility apart from level, but

- $r(a)$ refers to the fertility rate for women within an age group (e.g. 35-39), while
- F_{ij} uses the normalized cumulative distribution, and therefore refers to proportion of total fertility achieved by an age cut (e.g. 39/40).

Each model relates its own expression of an observed fertility schedule to a standard schedule:

- $n(a)$, an arithmetic mean of ten of the schedules identified by Henry as having a "natural" marital fertility pattern,[56]
- A_j^*, a cumulative marital fertility pattern extracted from the fertility history or other group of schedules analyzed.

Each model has a multiplier of its standard schedule:

- M, a variable scale factor,[57]
- α_i^*, a positive-valued near-constant which is a measure of the extent to which the cumulation of births expressed in A_j^* accelerates toward the median age of A_j^* and then decelerates over higher ages.

Each expresses deviations from its standard pattern, $n(a)$ or A_j^*, in terms of a standard pattern of departure with age:

- $v(a)$, a logarithmic departure from $n(a)$ above age 24, based on 43 recent schedules and interpreted to reflect the age pattern of conscious introduction of behavior to control fertility after some desired number of children is reached.[58]
- B_j^*, specific to the fertility history or other group of schedules analyzed and including all age cuts in its roughly quadratic pattern.[59]

Each model provides a measure of the degree of this deviation:

- m, interpreted as a measure of "degree of control" of fertility at higher ages and expressed in the extent to which the marital fertility distribution above age 24 is positively (or occasionally negatively) skewed according to the age pattern of $v(a)$.
- β_i^*, the extent to which the whole age distribution of marital fertility is skewed according to an age pattern determined by B_j^*.

We ask first whether the pattern empirically derived for Sweden and expressed in $\alpha_i^* A_j^*$ is a reference standard at all comparable to the tightly clustered family of patterns referred to as natural marital fertility distributions. The similarity of (1) the normalized age-specific schedules implied by $\alpha_i^* A_j^*$ for MX20 and (2) some of Henry's schedules, normalized to concentrate on pattern, is immediately apparent in the upper portion of Table 8 (see page 50 below) which relates mainly to analyses in a later section. In fact, $\alpha_i^* A_j^*$ is more like some of the "natural" distributions than the latter are like each other.

The relation is clarified when each of Henry's schedules, in cumulated normalized form, is approximated by a weighted sum of the

Swedish A_j^* and the Swedish B_j^*. (That is, each schedule is re-expressed on the folded square root scale used in the EHR analysis, and regressed on the A_j^* and the B_j^* found for MX20.) The degree of central concentration, α_i^*, ranges from 1.083 to 1.20, not unlike the range of 1.080 to 1.138 for MX20; degree of skewness, β_i^*, ranges from 0.035 to 0.218, compared to a low of 0.204 for MX20.[60,61] Closeness of fit of some of these EHR approximations to the reported "natural" distributions can be examined in raw scale non-cumulated form in Appendix C. The relation between natural fertility distributions and EHR-derived age patterns can of course be more fully examined when a more general A_j^* and B_j^* have been derived from a full range of fertility schedules representing a variety of populations and periods. That fertility distributions identified as having a "natural" pattern may be culturally determined slight variations of a more general pattern underlying all fertility distributions is given further support in a later section.

We next test whether m and β_i^* appear to measure the same phenomenon in a fertility time sequence. We use the expression

$$m = \frac{\ln[r(a)/Mn(a)]}{v(a)}$$

with the values of $n(a)$ and $v(a)$ underlying the Coale-Trussell model schedules, and with $M = \frac{r(20\text{-}24)}{n(20\text{-}24)}$, as in the Coale model of marital fertility alone; and we calculate

- m for the reported 1892-1959 marital fertility schedules,
- m for the fitted schedules resulting from the EHR analysis of this 68-year time sequence.

In order to obtain for each year a single value of m to compare with β_i^*, a weighted average m is calculated, weighting the m at each age by the proportion of the year's births occurring to that age group.[62] This emphasizes the shape of the schedule over the central ages of childbearing, the portion of the schedule which Coale and Trussell consider to be of primary importance in estimating their "degree of control."

Plotting β_i^* for MX15 and for MX20 in the same figure with weighted average m for fitted schedules and weighted average m for reported schedules (Fig. 10), demonstrates that averaged m and β_i^* have the same pattern of variations over time and differ principally in scale.[63]

F3. Changes in the marital fertility distribution not accounted for by changes in level of marital fertility

Now that the demographic significance of the EHR description of marital fertility has been examined broadly, we shall go further in dissecting the age distribution of marital fertility and in associating change with underlying demographic factors. Specifically, we separate from the time parameters $MX\alpha_i^*$ and $MX\beta_i^*$ that portion of change which is not accounted for by the changes in total rate of marital fertility (MT) over the 68 years of the history.[64]

To accomplish this separation, we use regression as exclusion, regressing the $MX\alpha_i^*$ and $MX\beta_i^*$ for each time sequence on the corresponding MT.[65] The regression residuals, referred to as $MX\alpha^*_{i \cdot MT_i}$ and $MX\beta^*_{i \cdot MT_i}$ to denote the time parameters linearly compensated for MT, are shown in Fig. 11. The $MX\alpha^*_{i \cdot MT_i}$ sequences suggest that very little,

if any, of the small amount of variation in $MX\alpha_i^*$ can be accounted for by change in MT. (Compare spread and pattern over time in the upper portions of Figs. 8 and 11.) The regression coefficient of MT for each sequence (Table 7) confirms this impression.

In contrast, a significant amount of the change in $MX\beta_i^*$ can be accounted for by changes in MT. The relation can be judged both from the regression coefficients (Table 7) and from comparison of the lower portion of Fig. 8 ($MX\beta_i^*$ over time) with the lower portion of Fig. 11 ($MX\beta_{i \cdot MT_i}^*$ over time).

At the same time, the portion of change in $MX\beta_i^*$ not accounted for by MT has some significant features as expressed in $MX\beta_{i \cdot MT_i}^*$:

- the transition from positive to negative at 1910,
- the long largely negative stretch from 1911 to 1943 (with one notable exception, 1920),
- the increasingly positive values after 1950.

These residuals can be interpreted in the following way:

0 indicates that change in $MX\beta_i^*$ is proportional to change in MT in the opposite direction (e.g. that increase in skewness of the distribution toward the younger ages, according to the pattern determined by MXB_j^*, parallels the decline in MT and could therefore be entirely accounted for by such a decline in functioning activity states at the higher ages).

Table 7. Regression Coefficients and Constants in Linear Compensation of EHR Time Parameters for the Level of Fertility and for Other Selected Factors

Compensation of	For	Regression Coefficient	Constant
$MX15\alpha_i^*$	MT15	0.002	1.223
$MX20\alpha_i^*$	MT20	0.004	1.102
$MX15\beta_i^*$	MT15	-0.536	1.664
$MX20\beta_i^*$	MT20	-0.400	0.817
$XX15\alpha_i^*$	$MX15\alpha_i^*$	0.470	0.892
$XX20\alpha_i^*$	$MX20\alpha_i^*$	0.713	0.399
$XX15\beta_i^*$	$MX15\beta_i^*$	1.101	-0.535
$XX20\beta_i^*$	$MX20\beta_i^*$	1.394	-0.153
$XX15\alpha_i^*$	XT15	0.018	1.458
$XX20\alpha_i^*$	XT20	0.009	1.182
$XX15\beta_i^*$	XT15	-0.966	0.953
$XX20\beta_i^*$	XT20	-1.098	0.990
$C15\alpha_i^*$	CT15	-0.013	1.489
$C15\beta_i^*$	CT15	-1.008	0.934

+ indicates that change in $MX\beta_i^*$ is greater than change in MT in the opposite direction (e.g. that skewness toward the younger ages according to the pattern determined by MXB_j^* shows greater increase than corresponds on the average to the decline in MT, and therefore indicates a disproportionate shift of functioning activity states from higher to lower ages).

- indicates that change in $MX\beta_i^*$ is less than corresponds on the average to change in MT in the opposite direction (e.g. that the age distribution of births is not skewed toward the younger ages according to the pattern determined by MXB_j^* as much as would have been predicted from a decline in MT due entirely to decrease in functioning activity states at the higher ages).

A demographic view of the observed variations over time may be that:

. From 1910 to 1943, the slightly disproportionate fraction of total fertility accounted for by women at higher ages reflects the childbearing behavior of cohorts with a high incidence of late marriage.[66]

. After 1950, older women were limiting their fertility disproportionately more highly than younger women, members of the post-1935 earlier-marrying cohorts.

The derived parameter may represent, then, a separation of effects of marriage entry pattern from the effects of birth limitation per se, in a cross-sectional view of the aggregate age distribution of activity states functioning within marriage.

F4. Changes in the overall distribution not accounted for by changes in the marital distribution of functioning activity states

Now we look at one of the possible dissections of the age distribution of overall fertility. By regression, we separate from the time parameters α_i^* and β_i^* for overall fertility that portion of change which cannot be accounted for by change in the corresponding marital fertility parameters.

Coefficients from regression of the $XX\alpha_i^*$'s on the corresponding $MX\alpha_i^*$'s signal that their variations over time, while covering a narrow range, are quite highly related (Table 7, page 43 above). The post-1950 tail for $XX15\alpha_i^*$ can be fully accounted for by $MX\alpha_i^*$. (Compare spread and pattern over time of $XX\alpha_i^*$, in the upper portion of Fig. 9, and $XX\alpha^*_{i \cdot MX\alpha_i^*}$, in the upper portion of Fig. 12.) A portion of the peak in $XX\alpha_i^*$ in the 1940s remains in the regression residuals, $XX\alpha^*_{i \cdot MX\alpha_i^*}$, however, and invites further investigation.

Coefficients from the regression of the $XX\beta_i^*$'s on the corresponding $MX\beta_i^*$'s indicate a high relation between these parameters (Table 7). At the same time, the regression residuals $XX\beta^*_{i \cdot MX\beta_i^*}$ show distinctive variations over time (Fig. 12):

- the transition from + to - at 1915,
- the decline to the most negative level of 1925-1935,
- the sharp rise between 1935 and 1945, with transition to + at 1940,
- the divergence of $XX15\beta^*_{i \cdot MX15\beta_i^*}$ from $XX20\beta^*_{i \cdot MX20\beta_i^*}$ after 1949.

These residuals can be interpreted in the following way:

0 indicates that change in $XX\beta_i^*$ is proportional to change in $MX\beta_i^*$ in the same direction (e.g. that any increase in positive skewness of the overall fertility distribution could be entirely accounted for by change in degree of positive skewness of the marital fertility distribution);

\+ indicates that change in $XX\beta_i^*$ is greater than corresponds on the average to change in $MX\beta_i^*$ in the same direction (e.g. that the combined effects of change in proportion entering marriage by age and the proportion by age having an illegitimate child have contributed to greater positive change in skewness of the overall fertility distribution than can be accounted for by positive change in skewness of the marital fertility distribution).[67]

\- indicates that change in $XX\beta_i^*$ is less than corresponds on the average to change in $MX\beta_i^*$ in the same direction (e.g. that positive skewness of overall fertility shows lesser increase than does positive skewness of marital fertility, suggesting that the proportion of women married and/or entering childbearing at younger ages is not increasing as rapidly as the distribution of marital fertility is shifting toward the younger ages).

A demographic view of the variations pictured in Fig. 12 may be that:

. To a small extent before 1915, and to a greater extent after 1940, the age distribution of overall fertility in these time sequences is positively influenced by the age pattern of entry into cohabitation

(and, in the final years of the sequences, is possibly also influenced by recent increases in marriage dissolution);

- The domination of the age distribution of overall fertility by the age distribution of marital fertility, extending from 1915 to 1940 and at its greatest between 1925 and 1935, declines rapidly with the 1935-1948 shift of the marriage pattern to younger ages;
- After 1948, a sustained positive influence of the age pattern of entry into cohabitation for XX15, but not for XX20, may result from increased relative importance of illegitimate births at age 15-19 when total fertility rate is low.

An approach for approximating age-specific proportions of married plus cohabiting active women is developed below, beginning with level-compensated distributions of overall and marital functioning activity states.

F5. Level-compensated distributions of functioning activity states in overall and marital perspectives.

In a preceding section (see page 41 above), we used regression to exclude from the time parameters, $MX\alpha^{*}_{i}$ and $MX\beta^{*}_{i}$, the effects of change in level of marital fertility over the 68 years. Level-compensated variation was expressed in the resulting vector pairs of regression residuals, $MX\alpha^{*}_{i \cdot MT_i}$ and $MX\beta^{*}_{i \cdot MT_i}$. We found the variation in $MX\alpha^{*}_{i}$ to be highly independent of change in MT, and found $MX\beta^{*}_{i}$ to have systematic variations not associated with change in MT. Now, the pairs of residual vectors become the basis for construction of year-by-year distributions of functioning activity states, freed of association with change in the level

of fertility.

First, the level-compensated time vector $MX\alpha^*_{i \cdot MT_i}$ is centered on the constant, k_1, from the regression of $MX\alpha^*_i$ on MT, so that $MX\alpha^*_i$ is compensated to a standard level of fertility. Then each element of this vector $(k_1 + MX\alpha^*_{i \cdot MT_i})$ is multiplied by each element of the age vector A^*_j, and each element of the level-compensated vector $MX\beta^*_{i \cdot MT_i}$ is multiplied by each element of the age vector B^*_j to form the pairs of components for each year at each age cut. Summing the pairs of components within each cell of the matrix then gives a time sequence of standard-level-compensated distributions of functioning activity states for the marital fertility history. For each year i and age cut j, the complete level-compensated element is:

$$M_{ij} = (k_1 + MX\alpha^*_{i \cdot MT_i}) A^*_j + MX\beta^*_{i \cdot MT_i} B^*_j$$

where $MX\alpha^*_{i \cdot MT_i}$ and $MX\beta^*_{i \cdot MT_i}$ = $MX\alpha^*_i$ and $MX\beta^*_i$ linearly compensated for MT

k_1 = constant from regression of $MX\alpha^*_i$ on MT

We can construct the corresponding standard-level-compensated distributions for each overall fertility history after regression of the $XX\alpha^*_i$'s and $XX\beta^*_i$'s on the corresponding XT's. For each year i and age cut j, the overall fertility level-compensated element is

$$X_{ij} = (k_2 + XX\alpha^*_{i \cdot XT_i}) A^*_j + XX\beta^*_{i \cdot XT_i} B^*_j$$

where $XX\alpha^*_{i \cdot XT_i}$ and $XX\beta^*_{i \cdot XT_i}$ = $XX\alpha^*_i$ and $XX\beta^*_i$ linearly compensated for XT

k_2 = constant from regression of $XX\alpha^*_i$ on XT

These expressions, M_{ij} and X_{ij}, on the folded square root scale used in the EHR analysis, are more useful to us once expressed on the raw fraction scale, and de-cumulated to give the standard-level-compensated age-specific overall and marital fertility distributions X'_{ij} and M'_{ij}.

For marital fertility, we can compare some of these level-compensated patterns with those implied by $\alpha^*_i A^*_j$ alone, and with some of those identified as having a "natural" fertility pattern. Inspection of Table 8 reveals both similarities (e.g., Greenland 1901-1930 and Sweden 1956; Hutterite 1921-1930 and Sweden 1951) and some small but distinctive differences (e.g., the lower proportion of functioning activity states at ages 25-29, also higher proportion at ages 40-44, in the earlier Swedish schedules than in all others, even the later Swedish). We appear to be dealing with a close family of distributions reflecting culture-specific slight variations of a more general underlying pattern--on which, then, the pattern of decline of marital fertility with age is imposed with varying intensity.

For overall fertility, the level-compensated patterns (examples in Table 9) should reflect both

- the timing of entry into childbearing, and

Table 8. EHR Standard Distributions, EHR Level-Compensated Distributions and "Natural" Distributions of Marital Fertility, Selected Examples

Source of Distribution	Age Group						$5\sum_{j=1}^{7} f(a)$
	20-24	25-29	30-34	35-39	40-44	45-49	
EHR Standard $\alpha_i^* A_j^*$							
(with lowest α_i^*, 1.080)	.2473	.2110	.2001	.1891	.1284	.0241	
(with highest α_i^*, 1.138	.2348	.2212	.2107	.1961	.1238	.0134	
Hutterites[1]							
Marriages 1921-1930	.2514	.2294	.2043	.1856	.1015	.0279	10.9
Marriages before 1921	.2425	.2302	.2169	.1909	.1046	.0148	9.8
Canada[1]							
Marriages 1700-1730	.2358	.2293	.2242	.1899	.1070	.0139	10.8
Norway[1]							
Marriages 1874-1876	.2434	.2336	.2096	.1776	.1106	.0252	8.1
Greenland[2]							
1851-1900	.246	.230	.223	.189	.098	.015	7.70
1901-1930	.271	.229	.208	.169	.104	.018	7.84
Taiwan (women born c. 1900)[1]	.2626	.2403	.2201	.1892	.0820	.0058	6.95
Sweden							
$f(A_j^*, B_j^*)$ 1895	.2476	.2180	.2044	.1891	.1228	.0181	7.76
$f(A_j^*, B_j^*)$ 1921	.2431	.2122	.2024	.1915	.1289	.0221	5.50
$f(A_j^*, B_j^*)$ 1935	.2372	.2093	.2025	.1946	.1331	.0233	3.28
$f(A_j^*, B_j^*)$ 1951	.2534	.2198	.2035	.1860	.1195	.0177	3.07
$f(A_j^*, B_j^*)$ 1956	.2723	.2280	.2020	.1759	.1071	.0147	3.12

(1) Data from Henry, loc. cit. in footnote 18.

(2) Data from Hansen, H.O. "From Natural to Controlled Fertility: Studies in Fertility as a Factor of the Process of Economic and Social Development in Greenland c. 1851-1975," presented at the IUSSP Seminar on Natural Fertility, Paris, March 1977.

. those culture-specific factors which determine the timing of functioning of activity states subsequent to a first birth.

These constructed distributions, freed of association with level of activity while retaining association with timing of activity, become the basis of approximations of age-specific proportions of married plus cohabiting active women.

F6. Use of standard-level-compensated distributions of births to approximate age-specific proportions of married plus cohabiting active women

Actual populations often depart from the hypothetical idealized one in experiencing premarital pregnancy and illegitimacy at significant levels. Social customs of the time and place determine the size and age composition of the non-married group exposed to pregnancy, and, for members of this group, strongly influence the outcome of pregnancy, in terms of abortion, or an illegitimate birth, or a conception legitimated by marriage before the birth. With a given proportion of unmarried women exposed to pregnancy, quite different age distributions of marital fertility can result, depending on:

. the separation by age into those marrying and those not marrying before giving birth to a non-maritally conceived child, and

. the relation, in magnitude and age distribution, between the pre-maritally conceived legitimate births and post-maritally conceived births.

Table 9. EHR Level-Compensated Distributions of Cross-Sectional Overall Fertility, Selected Years

Source of Distribution	Age Group						
	20-24	25-29	30-34	35-39	40-44	45-49	$5\sum_{j=1}^{7} f(a)$
EHR Standard $\bar{\alpha}_i^* A_j^*$	.1044	.2136	.2571	.2471	.1556	.0222	
$f(A_j^*, B_j^*)$ 1895	.1020	.2118	.2572	.2489	.1570	.0226	4.07
$f(A_j^*, B_j^*)$ 1921	.0988	.2034	.2531	.2513	.1656	.0280	2.83
$f(A_j^*, B_j^*)$ 1935	.0771	.1903	.2560	.2673	.1811	.0282	1.61
$f(A_j^*, B_j^*)$ 1951	.1471	.2390	.2517	.2179	.1259	.0183	2.11
$f(A_j^*, B_j^*)$ 1956	.1701	.2535	.2494	.2027	.1096	.0147	2.00

Resulting problems for aggregate fertility models can be severe, both in description of the age distribution of marital fertility and in expression of age-specific proportions of cohabiting women. Age 15-19 births, or births at marriage durations of less than two years, are often omitted to lessen or avoid the effects of non-maritally conceived births on model descriptions of fertility. The demographic importance of both the timing of first birth and the length of the first interbirth interval argues strongly, however, for seeking ways to include in analysis these portions of aggregate data which cover the beginning of childbearing for a significant proportion of women in many populations.

We have already seen that EHR analysis can provide excellent fits to very diverse distributions, including marital fertility distributions highly influenced by premarital pregnancy. Further exploration suggests informative ways of dealing, in the aggregate, with those functioning activity states which lead to illegitimate births rather than to precipitated marriages. At this stage, we use level-compensated distributions of functioning activity states as the basis for approximating age-specific proportions of married plus cohabiting active women (referred to as MPA women for brevity). That is, of the possible combinations of exposure and activity status:

	Married	Not married but cohabiting	Not cohabiting
Functionally active	1	2	X
Functionally inactive	3	4	5

we approximate the age-specific proportions of (1 + 2 + 3). We illustrate the approach with the conventional overall and marital fertility sequences analyzed above (and then suggest other pairs of sequences whose analysis may further aid understanding of the phenomenon).

The total rates of overall births and legitimate births for each year are dispersed across the corresponding level-compensated patterns X'_{ij} and M'_{ij} derived for that year. This provides, first of all, a level-compensated overall fertility rate and a level-compensated marital fertility rate for each age group in each year. An approximation of the proportion of MPA women in year i and age group j is then calculated as

$$P_{ij} = \frac{X'_{ij}(XT)(5)}{M'_{ij}(MT)(5)}$$

where (XT)(5) = total rate of overall fertility expressed as mean number of children per woman over the reproductive span of a synthetic cohort

(MT)(5) = total rate of marital fertility expressed as mean number of children per married woman over the reproductive span of a doubly synthetic cohort.[68]

Approximations based on XX20 and MX20 parameters are shown in Fig. 13 for three age groups which have experienced striking alteration of marriage pattern over the 68 years of the time sequence analyzed. The significant change in age-specific proportions married, beginning in 1935, and the deceleration in change after 1945 are clearly picked up by the

procedure.

The approximations can be improved, however, by EHR analysis of at least two additional sequences:

- a legitimate fertility sequence (calculated from legitimate births by age of mother, and the total female population in each age group) to pair with marital fertility; and
- a "married or actively cohabiting" fertility sequence (which simply includes all non-marital functioning activity states on the same basis, whether or not they precipitate marriage before the birth of a non-maritally conceived child) to pair with overall fertility.[69]

The immediate goal in refining and extending the analysis is to develop as concise and useful a description as the data will allow of a population's experience of childbearing, within, outside of, and influencing legal marriage. Taking advantage of the richness of the data for this one population and of the strengths of the EHR approach to the data, the larger goals are at least two:

- to expand understanding of the outcome of pregnancy in the early years of childbearing in the context of a particular society changing over time;
- to tease out more information on the age-specific proportions cohabiting which are apt to underlie the observed fertility rates for the early portion of the reproductive span within the context of a population's complete age pattern of childbearing.

G. Changes across time in cohort and cross-sectional overall fertility patterns (selected observations on the full 1775-1959 fertility histories)

Demographic implications of the EHR components have been examined at several depths, using the data of extraordinarily high quality for the 68-year cross-sectional overall and marital fertility sequences. Now we turn to the 185-year overall fertility history in cross-sectional and cohort perspectives. The analyses of the X15, X20, C15 and C20 time sequences provide several unusual opportunities:

- to examine in EHR model terms, the relation between changes in cohort and cross-sectional distributions of functioning activity states over an extended period;
- to discover how much can be deduced about change, from EHR analysis of overall fertility alone;
- to see how EHR analysis handles data of somewhat lesser accuracy in combination with data of very high quality.

G1. Comparability of cohort and cross-sectional EHR components

A fertility model may reasonably be questioned on its relative capacity to handle cross-sectional and cohort data. Each of these perspectives on fertility distributions provides its own challenge to appropriate description. A cross-sectional slice is the composite of a small portion of the child-bearing experience of each of many cohorts whose diverse histories influence their behavior at any given period. The period in which the cross-sectional slice is taken then has its own influence on the experience of all of the

cohorts of women passing through the period at various ages.

For the fertility rate matrix, submitted to EHR analysis in cross-section (by year) and on the diagonal (by cohort),[70] three of the possible outcomes are that

- the underlying similarities of the data in cohort and cross-sectional forms may dominate the fitted parameters, leaving in the residuals most of the divergence of the two;
- the systematic differences may be well captured in the separate sets of fitted parameters for cohort and cross-sectional sequences;
- either "true" similarities or "true" differences between cohort and cross-sectional underlying regularities may be obscured by diffusion through fitted parameters and residuals for one or both sequences.

We found above (see pages 12 to 25) that the present EHR analysis

- has given the overall fertility sequences close fitted descriptions in cohort perspective, and still closer fitted descriptions in cross-sectional perspective;
- has left a different type of regularity in the small residuals from the cross-sectional analysis than in those from the cohort analysis.

Looking for systematic differences in the cohort and cross-sectional fitted parameters is more profitable after fine-tuning the fits.[71] Here, we concentrate on the similarities of the pairs of components $\alpha^*_i A^*_j$ and $\beta^*_i B^*_j$ so far derived using the folded square root re-expression. We give

particular attention to the X15 and C15 time histories, and consider those for X20 and C20 in less detail.

The components $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$ underlying the X15 sequence have already been examined above (see pages 29 to 33) to establish the function of each component in describing a fertility time history. Here we construct the same model distributions for the C15 sequence so that Table 10 can be compared directly with Table 4, and Table 11 with Table 5.

We see that $\alpha_i^* A_j^*$ has a slightly wider range for C15 than for X15, and $\beta_i^* B_j^*$ has a slightly lower and narrower range for C15 than for X15 (Tables 10 and 4). When we turn to Tables 11 and 5, however, we find pronounced similarity for X15 and C15 in:

- the raw scale schedules defined by a given level of α_i^* and β_i^*, and in
- the magnitude and direction of change in the distribution as each component is varied in even increments while the other is held constant.

This means that change in these two histories over time can reasonably be compared below by means of these EHR time parameters.[72]

G2. Some intersections of cohort and cross-sectional age distributions of functioning activity states, viewed through EHR components

Before superimposing the cohort and cross-sectional time parameters, we look at the parameters for each sequence individually, to note the

Table 10. Range of Contribution of Components $\alpha_i^* A_j^*$ and $\beta_i^* B_j^*$ to the Fitted Description of the C15 Sequence, Folded Square Root Scale

Time Parameter	Component	Age Cut					
α_i^*	$\alpha_i^* A_j^*$	19/20	24/25	29/30	34/35	39/40	44/45
Low 1.396		-.8325	-.5424	-.1866	.1699	.5079	.7994
Median 1.480		-.8830	-.5753	-.1979	.1802	.5387	.8479
High 1.544		-.9209	-.6000	-.2064	.1879	.5618	.8843
β_i^*	$\beta_i^* B_j^*$						
Low -.0624		-.0126	-.0298	-.0350	-.0313	-.0233	-.0100
High .6327		.1282	.3019	.3545	.3177	.2362	.1018

* * * * * * * * * * * * * * * * * * *

Table 11. Age-Specific Fertility Distributions (raw scale) Based on the Age Parameters A_j^* and B_j^* for C15 and Increments in the Time Parameters α_i^* and β_i^*

α_i^*	1.47	1.47	1.47	1.47	1.42	1.52	Change by 0.1
β_i	Change by 0.1	0	0.1	0.05	0.05	0.05	0.05
Age Group	Change in f(a)	f(a)	f(a)	f(a)	f(a)	f(a)	Change in f(a)
15-19	.0046	.0135	.0181	.0157	.0230	.0097	-.0133
20-24	.0211	.1169	.1380	.1274	.1307	.1231	-.0076
25-29	.0131	.2319	.2450	.2386	.2327	.2443	.0116
30-34	-.0043	.2631	.2588	.2611	.2523	.2699	.0176
35-39	-.0145	.2247	.2102	.2175	.2118	.2229	.0111
40-44	-.0159	.1282	.1123	.1202	.1224	.1170	-.0054
45-49	-.0039	.0216	.0177	.0196	.0272	.0132	-.0140

distinctive features (Figs. 14,15).

The pairs of α_i^*'s (for X15 and X20, for C15 and C20) differ principally in level. The narrow range of departures of α_i^* from constancy is similar in all of the histories (except for some of the post-1914 cohorts).[73] The pairs of β_i^*'s (for X15 and X20, for C15 and C20) also differ principally in level.

Two periods of transition are suggested for X15 and X20, three for C15 and C20:

- a prominent increase in positive skewness of the distribution at 1885-1895 for X15 and X20 (Fig. 14 , lower section) and for those cohorts starting their childbearing at ages 15-19 about 1860-1865 and therefore in the latter half of their childbearing years in 1885-1895 (Fig. 15, lower section) --suggesting parity dependent limitation of births;
- a turn toward negative skewness at 1835-1845 for X15 and X20 (Fig. 14, lower section) and for those cohorts starting their childbearing at ages 15-19 in 1830-1840 and therefore aged 20-35 in 1835-1845 (Fig. 15, lower section)--suggesting increase in age of marriage around 1840;
- an abrupt change in cohort behavior beginning with those cohorts aged 15-19 after 1914 (as judged by both α_i^* and β_i^*).

This last transition has already been signaled in the cohort residuals (Fig. 6).[74] That these variations may reflect new patterns in the cohort age distribution of functioning activity states is suggested by the combination

of stability in post-1915-cohort total rate of fertility after the earlier fairly steady decline, and the rapid evolution which these cohorts experienced in age pattern of first marriage, in striking departure from the cohorts aged 15-19 before 1913.[75] (That specifying a different combination of data re-expression and fitting may better accommodate such variations has already been suggested.)

In the context of the full 185-year sequence of α_i^*'s and β_i^*'s for the X15 and X20 histories, the periods 1910-1920 and the period around 1940 might be thought small aberations in an ongoing trend. One will recall, however, that these periods were identified in earlier portions of the analysis as transitions, appearing to involve both change in marital fertility patterns and change in the age pattern of entry into cohabitation. (Recall Fig. 1, evidence as early as the EHR fitting; also Fig. 11 on linear compensation of the marital fertility sequence parameters for total rate of fertility and Fig. 12 on linear compensation of the overall fertility sequence parameters for marital fertility.)

For a more vivid impression of some intersections of cohort and cross-sectional experience, we superimpose the pairs of time vectors. For example, with $C15\beta_i^*$ centered on year at age 30-34 (Fig. 16A), one sees that, before 1908, the extent to which the cross-sectional age distribution of functioning activity states was changing in degree and direction of skewness was very similar to the extent to which cohorts then at the central ages of childbearing were shifting their age pattern of activity to younger or older ages. The patterns over time then diverge sharply. If

one moves the post-1907 $C15\beta_i^*$'s forward another 10-15 years on the time line--(Fig. 16B shows a shift of 12 years, so that these cohort values are then superimposed on those for $X15\beta_i^*$ for the years 1920 and after)--the patterns of change in the two skewness vectors are again extraordinarily similar for about 25 years before parting abruptly once more about 1945.

We learn more about the relation of these patterns by a further dissection of the cohort parameter β_i^*. By LS regression of $C15\beta_i^*$ on CT, we separate into the residuals that portion of change in β_i^* not associated with change in level of fertility (Fig. 17). (The regression coefficient and constant are included in Table 7, page 43.) The cohorts with positive residuals are those whose distribution of functioning activity states is skewed more toward the younger ages than would correspond on the average to the level of fertility; the cohorts with negative residuals are those with later age distributions of activity than would correspond on the average to the level.

If we direct our attention to the cohorts aged 15-19 in 1892-1917 (and aged 30-34 in 1907-1932), we see that the divergence of cohort from cross-sectional pattern over time in Fig. 16A begins precisely with those cohorts which adopted an earlier than average age pattern of level-compensated activity. The almost-superimposable behavior for 25 years in Fig. 16B includes all of these cohorts with the earlier pattern, and the divergence at 1945 is precisely with those cohorts which reverted to a later-than-usual age pattern of level-compensated activity. (One may note that for the cohorts covered by

Fig. 16B, total rate of fertility, expressed as mean number of children per woman, dropped from 3.50 to 1.88, while cross-sectional total rate in the years 1920-1945 dropped from 3.23 to 1.70 and rose again to 2.63.)

Comparison of the cohort and cross-sectional sequences of full level-compensated distributions of functioning activity states (as for the 68-year marital and overall fertility sequences, pages 47 to 51 above) is reserved for a later discussion of the parameters resulting from fine-tuning the fits.

G3. Cohort vs. cross-sectional evidence for data points of lesser accuracy

A major benefit of EHR analysis of fertility distributions was expected to be the capacity to extract the underlying regularities while revealing the points or periods of departure from the trend or the usual pattern. We have concentrated in preceding sections on the regularities and their further dissection, in order to learn more about the dynamics of change over the time sequences. Here we open the question of the handling of data of lesser accuracy.

Some errors, particularly in the population data before about 1840, are thought to exist in the yearly data from Grunddragen used to calculate the pre-1875 age-specific fertility distributions.[76] (The possibility of other common types of error --omissions of births, misplacement of births in time, misplacement of women by age--must also be considered.) From the EHR analysis, what general evidence of such errors is there?

The EHR-derived time parameters for X15 and X20 do not exhibit

much year-to-year irregularity (Fig. 14). The only singular departures are in β_i^* for 1783 and 1792. (Fig. 5 shows these two years to be aberrant in the residuals also.)[77] In contrast, the total rate of fertility fluctuates rather widely from year to year before 1870. It appears that the age distribution of births in cross-section may be fairly accurately represented by the recorded data, even though the level of fertility may be variously in error for the early years.

Cohort histories constructed from the cross-sectional age-specific rates would reflect any year-specific errors in level in two ways:

- The less accurate rates for one year would be disseminated across seven cohorts at one five-year age group in each cohort. Thus the age distribution of births in each of the cohorts would be distorted in a different way.
- The cohort total rate, the sum of the more or less accurate age-specific rates across the seven age groups for a cohort, would serve to average out errors in cross-sectional total rate. Cohort total rate should thus follow a smoother course than does cross-sectional total rate.

The picture given by the EHR cohort analysis is consistent with such a dissemination of cross-sectional error. In conjunction with the five-year lagged pattern in the residuals (Fig. 6) jumpiness is evident in both α_i^* and β_i^* at about five-year intervals over the early portion of the C15 and C20 histories (Fig. 15). At the same time cohort total rate is much less irregular than cross-sectional total rate.

This outcome does not, of course, prove error, since circumstances which actually altered fertility rates in a period would have a similar impact on aggregate data. If volition is significant in determining when potential activity states are functioning, cohort-to-cohort variation in the timing of births over the reproductive span would lead to irregularity in cross-sectional total rate even when differences between cohorts in average number of births per woman is small. The lagged pattern seen earlier in the cohort residuals (Fig. 6) may, then, have different causes, or a combination of causes, in different portions of this time history. One approach, in this instance, may be to work backwards--reconstructing cross-sectional rates from cohort rates after fine-tuning the EHR cohort fits. Broad understanding of fertility data and knowledge of the population's social history can be important aids in exploring choices in such a process.

H. Conclusions

Empirical higher-rank (EHR) analysis proves to be a powerful means of extracting the patterns which unify a long and varied time series of age-specific fertility schedules. Analyses of data in single-year time sequence demonstrate that dynamics of change as well as variety of pattern can be captured in the fitted descriptions.

By emphasizing the centrality of examination of residuals in achieving optimal fits and in interpretation of the fitted descriptions, the robust/resistant and data-guided analyses reported here

- provide unusually close fitted descriptions of the diverse age distributions of overall and marital fertility in both cross-sectional

and cohort perspectives;

- take some major steps in reducing the variability in the fertility data to a concise and coherent demographic picture which differs in important ways from the descriptions other aggregate fertility models have provided, while having some significant relations to other models;
- suggest ways of refining still further the fitted descriptions to provide additional insight into the underlying structure of aggregate fertility distributions, and ways of identifying and dealing with error-ridden data.

Demographically guided choice of a standard form in which to use the fitted descriptions so far developed leads to separation of the fertility distributions in each time sequence (overall or marital) into three components:

- a nearly-fixed pattern of cumulation of births with age, on which are imposed the major variations in the distribution, due to change in level of fertility or to the timing of births;
- a component which comprises the association of change in the level of fertility with change in its age distribution;
- a component which encompasses effects of the timing of births, apart from level, on the age distribution of fertility.

This separation proves to be an effective one in efforts to discern in the aggregate the relative contributions of the age-specific proportions of women cohabiting and the age patterns of childbearing of cohabiting women.

We present here not a "finished" model but informative and provocative steps in the continuing search for useful and more refined ways of looking at the full diversity of fertility patterns in changing social milieux. The success in developing a sound description of a long and varied fertility history encourages a full exploratory analysis with extension of this EHR-based work to cross-population comparisons of fertility distributions.

FOOTNOTES

* The work on which this paper is based was begun at the Office of Population Research and continued in the Department of Statistics at Princeton University. I received useful criticisms from Ansley J. Coale, Norman B. Ryder, and Barbara A. Anderson at various stages of the work, and Donald R. McNeil gave generously of technical advice during the early stages of the analysis. I am particularly indebted to John W. Tukey for his advice and sustained interest.

1. Some notable examples are found in the "Brass methods" (brought together in Brass, W., 1975, Methods for Estimating Fertility and Mortality from Limited and Defective Data, An Occasional Publication, Chapel Hill: University of North Carolina, International Program of Laboratories for Population Statistics); and in the Coale indices (Coale, A.J., 1967, "Factors associated with the development of low fertility: an historic summary," in New York: United Nations, World Population Conference, 2, pp.205-209) subsequently used in a series of monographs on the decline of fertility in Europe (Coale, A. J., Anderson, B.A., and Harm, E.; Knodel, J.E.; Lesthaeghe, R.; Livi-Bacci, M.; Van de Walle, E., Princeton: Princeton University Press; Forrest, J.D., Ph.D. dissertation, Princeton University.)

2. An example of productive reappraisal of admittedly flawed data, using new techniques, is found in Barclay, G.W., Coale, A.J., Stoto, M.A., and Trussell, T.J., 1976, "A reassessment of the demography of traditional rural China," Population Index, 42, pp.606-635.

3. See Keyfitz, N., 1977, Introduction to the Mathematics of Population, with revisions, Reading, Mass.: Addison-Wesley, pp.140-169, and Brass, W., 1974, "Perspectives in population prediction: Illustrated by the statistics of England and Wales," Jour. Royal Statist. Soc. A 137, pp.532-583, for discussions of the most widely studied functions and the extent to which they fall short of describing a range of fertility distributions accurately.

4. Coale, A.J., 1971, "Age pattern of marriage," Population Studies 25, pp.193-214.

   Coale, A.J. and McNeil, D.R., 1972, "The distribution by age of the frequency of first marriage in a female cohort," J. Amer. Stat. Assoc. 67, pp.743-749.

   Coale, A.J. and Trussell, T.J., 1974, "Model fertility schedules: variations in the age structure of childbearing in human populations," Population Index 40, pp.185-258.

5. Since a given population may depart, to a greater or lesser degree, from the assumptions and external standards on which the model schedules are based, the model parameters may not, of course, retain precise demographic meaning in fitting actual fertility schedules. Coale and Trussell report that the discrepancy is especially pronounced when marriage or childbearing patterns are changing (loc. cit. in footnote 4, p. 193).

6. McNeil, D.R., and Tukey, J.W., 1975, "Higher-order diagnosis of two-way tables, illustrated on two sets of demographic empirical distributions," Biometrics 31, pp.487-510.

7. Tukey, J.W., 1977, Exploratory Data Analysis, Reading, Mass.: Addison-Wesley.

8. See Mosteller, F., and Tukey, J.W., 1977, Data Analysis and Regression, Reading, Mass.: Addison-Wesley, particularly pp.351-358, for discussion of the desirable properties of the biweight in such procedures.

9. Single-year schedules were preferred because of the belief that the variability in data at this level of detail contains useful demographic information not captured in ten-year or five-year averages even when systematic long-term changes over time are gradual.

10. For a concise description of the sources of Swedish population statistics from the earliest times, and for a discussion of the quality of the data and the adjustments that have been made to early data, see Hofsten, E. and Lundstrom, H., 1976, Swedish Population History, Stockholm: National Central Bureau of Statistics.

11. Although 1814 was the last year in which Sweden was actively engaged in a war, subsequent conflicts have affected the country to a greater or lesser degree. (For example, the possible effects of World War II on Swedish fertility are considered in Hyrenius, H., 1946, "The relation between birth rates and economic activity in Sweden 1920-1944," Bulletin of the Oxford University Institute of Statistics 8, pp.14-21.)

12. The extensive records of economic and social variables also encourage later tests of the value of a derived fertility model's parameters in substantive research on economic and social change.

13. Sweden, 1878, Grunddrag af Sveriges Befolknings-Statistik för åren 1748-1875, Stockholm: National Central Bureau of Statistics.

Sweden, 1875-1910, Sveriges Officiella Statistik: Befolknings-Statistik, Stockholm: National Central Bureau of Statistics.

Sweden, 1911-1959, Sveriges Officiella Statistik: Befolkningsrorelsen, Stockholm: National Central Bureau of Statistics.

Data for single years through 1875, as adjusted for obvious omissions and published by the Bureau in the single appendix Grunddragen, were preferred to Sundbärg's later more extensively revised figures by five-year periods up to 1860. Even with some errors, a larger number of data points have advantages over aggregated data for the exploratory type of analysis proposed here.

The marital fertility history covers that portion of the overall fertility history for which recorded data allow calculation of yearly age-specific marital fertility rates. Before 1892, decennial reports of population by age and marital status combined are available, beginning with the census of 1870. Reporting of confinements by age of mother and legitimacy of birth combined began in 1868.

The series was stopped at 1959 because of the wish to include as full a variety of age patterns of childbearing as would be consistent with related analysis of overall and marital fertility, but to stop short of the recent increased dissociation of childbearing from marriage. There will be value now in extending the series to see how new cohabitation and marriage patterns are influencing the age distributions of overall and marital fertility as seen through EHR parameters.

14. Social and political factors contributing to eighteenth and early nineteenth century marriage patterns are discussed in Utterstrom, G., 1962, "Labour policy and population thought in eighteenth century Sweden," Scandinavian Economic History Review 10, pp. 262-279. A view of Swedish marriage changes in terms of proportion of years between ages 15 and 50 lived in the married state by a birth cohort of women, 1751-1901, will be found in Ryder, N. B., "The influence of declining mortality on Swedish reproductivity," Current Research in Human Fertility, Proceedings of a round table at the 1954 annual conference, Milbank Memorial Fund, pp. 65-81. An analysis of single-year marriage entry patterns of post-1850 birth cohorts will be found in Ewbank, D. C., 1974. An Examination of Several Applications of the Standard Pattern of Age at First Marriage, Ph. D. dissertation, Princeton University.

15. For a summary of government efforts, beginning in 1937, to encourage marriage in some segments of the population, see Glass, D. V., 1967, Population Policies and Movements, London: Cass, pp. 327-331.

16. Page has examined the effect of marriage duration, independent of age, on the childbearing pattern in Sweden since 1911. (Page, H. J., 1977, "Patterns underlying fertility schedules: A decomposition by both age and marriage duration," Population Studies 31, pp. 85-106.)

17. In each year from 1911, when recording of births by duration of marriage began, until 1959, the final year of the fertility sequences analyzed in the present report, approximately 75-85% of all legitimate births to women aged 15-19 and approximately 34-45% of all legitimate births to women aged 20-24 are reported to have been premaritally conceived. See Hofsten and Lundstrom, op. cit. in footnote 10, pp. 26-29, as well as Sveriges Officiella Statistik: Befolkningsrorelsen, 1913-1959.

18. Henry, L., 1961, "Some data on natural fertility," Eugenics Quarterly 8, pp. 81-91.

19. beginning in 1955, live births

20. Data by single year of age of mother are available after 1890, but will not be considered in the present analysis since our interest here is in demonstrating how much can be learned from more widely available five-year age group data by this exploratory approach. Considerable real irregularity is, of course, averaged out in the use of these five-year age groups. Significant change (for example, in marriage patterns) may also occur within such an age group.

21. These are not, of course, true cohorts, but overlapping approximations cut on the bias.

22. The sum of age-specific rates for the five-year age groups, $\sum_{j=1}^{7} f(a)$ expressed in the rate for women at age cut 49/50, will be referred to as "total rate." Multiplication of this rate by five gives, for each cohort, the mean completed fertility per woman, and gives for each cross-sectional schedule the conventional expression of TFR, or mean number of children per woman over the childbearing years of a synthetic cohort.

23. The parameters for XX15 and XX20 can be expected to differ slightly from those for X15 and X20 because the former are developed by fitting the fertility experience of the last 68 years alone, divorced from the experience of the preceding 117 years. The time parameters α_i^* and β_i^* for the XX15 and XX20 histories are seen below to follow the same patterns of variation as those of the X15 and X20 histories, however, and to differ mainly in level (see pages 87 and 92). Fitting the shorter sequence alone does contribute to fine-tuning of the fit for that portion of the longer sequence.

24. Results of a corresponding analysis of cohort overall and marital fertility sequences for the last 38 cohorts, those aged 15-19 in 1892-1929, will be referred to but not reported in detail here.

25. In an extended EHR analysis, the possible value of further truncation can be explored.

26. McNeil and Tukey, loc. cit. in footnote 6.

27. Comparison of results for the various combinations are included in Breckenridge, M. B., 1976, Time Series Model of Age-Specific Fertility: An Application of Exploratory Data Analysis, Ph.D. dissertation, Princeton University.

28. The iterative fitting procedures and display programs implemented by McNeil in APL for use with large data sets and an interactive computer are brought together in McNeil, D.R., 1977, Interactive Data Analysis, New York: Wiley.

29. E. J. Orav (1977), An Expanded Exploratory Data Analysis Study of Age-Specific Fertility, Senior thesis, Princeton University) has since tested a variety of five- and six-parameter models and one eight-parameter model on the X15 sequence, using the folded square root re-expression and various weightings.

30. While this measure is less sensitive to a small number of large residuals than is a squared variation criterion of fit, it can still produce a misleading impression of poor fit from a very good overall fit with a few "outliers." This criterion of fit is best used, therefore, in conjunction with detailed examination of residuals to determine the nature of departures from fit. A recently developed robust measure of variance of residuals proves, in many contexts, to be a more useful measure of fit, one less distorted by "outliers." This s^2_{bi}, described in Mosteller and Tukey, op. cit. in footnote 8, pp. 207-208, has been used, in a slightly modified form, in extensions of the present work to simplify choice between fits.

31. The similarity to the logit which Brass has used so productively will be noted.

32. If the fitting had started instead with the time dimension of the matrix, a similar, but probably not precisely equivalent fitted description would have been generated.

33. The general case includes variation at either end cut in relation to the other; but in a fertility distribution cumulated to age 45-49, the variation at the last age cut is always small under usually encountered circumstances.

34. About 97% of the squared variation, or about 86% of the absolute variation, is taken up by fitting a single time-independent cumulative distribution.

35. While the robust measure of variance, s^2_{bi}, (see footnote 30) was not used in the fitting process in the work reported here, values of s^2_{bi} of residuals for several sequences are included in Table 1 for comparison.

36. No attempt has been made to pick "best fits."

37. With a folded linearizing re-expression (such as the folded square root) which centers the distribution on its mean, the values at age cuts near the center are changed relatively less than those at the *ends* of the distribution in the process of re-expression. Therefore, a residual of a given size, when it appears at the lowest or highest age cuts, will have relatively less significance for the fitted distribution on the raw fraction scale than will a residual of the same size when it occurs at one of the central age cuts. For example, with the folded square root re-expression, a residual of 0.01 will have its highest de-transformed value, 0.007, at the center of the cumulated normalized fertility distribution and will have progressively lower de-transformed value toward either tail of the distribution.

38. This inclusion has generally been considered problematic because of high rates of premarital pregnancy.

39. The difficulties which many models have in fitting the tails of fertility distributions have often been dismissed as relatively unimportant. Good fit in the tails may be of particular importance, however, when total fertility is low or the age pattern of entry into cohabitation is changing.

40. Tukey, op. cit. in footnote 7, pp. 355-356, 398; Mosteller and Tukey, op. cit. in footnote 8, pp. 192-193.

41. These plots show as a box the interquartile range of the residuals for each of the six age cuts 19/20 to 44/45, with location of the median residual indicated by a bar. The relative positions of upper and lower values within one interquartile distance of the upper and lower quartiles are indicated by an x beyond each end of the box. Outlying values within 1.5 times the interquartile distance of each quartile are shown by empty circles, while residuals further out are shown by shaded circles.

42. When an appropriate compound non-linear smoothing procedure (see Tukey, op. cit. in footnote 7, pp. 205-264 and 523-542) is applied to the residual vectors by age group to deemphasize irregular fluctuations, thus providing more sensitive detection of patterns of co-variation, the tilts in the scatterplots persist, further supporting the sense of some remaining underlying structure in the small residuals.

43. A triple multiplicative model, in combination with the folded square root re-expression and c = 12 in the weight function, does remove the long stretches of residuals above or below the central interval, but does not remove inter-age group structure as seen in scatterplots of the diminished residuals (Breckenridge, M. B., and Orav, E. J., "An expanded EHR analysis of the age distribution of fertility" (in preparation), which will combine results from Orav, op. cit. in footnote 29, and parallel analysis and research by the present author).

44. Anscombe, F. J., 1967, "Topics in the investigation of linear relations fitted by the method of least squares," Jour. Royal Statist. Soc., B29, pp. 11-52.

45. In EHR analysis of post-1869 single-year-of-age cohort data for Sweden, lagged pattern persists in the residuals, suggesting that factors in addition to age grouping and method of constructing the cohort sequences must be sought.

46. This question receives further attention below (see page 64).

47. Breckenridge, unpublished.

48. where rank-two refers to the sum of two rank-one terms, and a rank-one term is the product of a constant by a function of row alone by a function of column alone.

49. A modification of a program written by Alison Pollack was used for this procedure.

50. Tukey, J. W., 1977, "Transfactorial fits. The linear geometry in the two-way case."

51. For any one age group, the amount of change in f(a) on the raw scale varies slightly and systematically over repeated increments in β_i^* to higher levels, holding α_i^* constant, or over repeated increments in α_i^* to higher levels, holding β_i^* constant--a consequence of having used a non-linear re-expression of the data in the fitting procedure.

52. When one is considering less than all potential and functioning activity states, e.g. the activity of married women only, or of women above age 19 only, $\alpha^*_i A^*_j$ expresses the tendency for activity to be pulled toward the median age of $\alpha^*_i A^*_j$ rather than to be distributed evenly beyond the first age group, whatever proportion of activity is attributable to that first age group.

53. At times, such sub-populations will be readily identifiable ethnic, religious, regional, or occupational groups. Widespread incidence of highly intermittent activity, with varying causes, may also be a significant source of diversity between women contributing to spread of the aggregate distribution.

54. Breckenridge and Orav, loc. cit. in footnote 43.

55. A^*_j and B^*_j for all sequences are shown in Appendix B, Table B3.

56. Henry, loc. cit. in footnote 18. These schedules are considered to represent the aggregate childbearing experience of couples when their behavior affecting fertility is not influenced by the number of children already born to them. Such distributions are, however, recognized to reflect cultural and biological variations which may not be age- and parity-independent.

57. In the absence of "control," M is interpreted as the level at which natural fertility is experienced; in the presence of "control," however, Trussell reports that M appears to be a composite of several factors: not only the level of underlying natural fertility, but also functions of total fertility or degree of control of fertility, and variations in the distribution due to spacing at high levels of birth limitation. These factors, he concludes, elude separation with the existing models of age-specific marital fertility. (Trussell, T.J., 1977, presented at the IUSSP seminar on natural fertility, Paris, March 1977.)

58. Trussell (loc. cit. in footnote 57) discusses the need for modification of v(a) to include a marriage-duration effect, and the problems of doing this except in an overall fertility model.

59. The fact that B^*_j is not pegged to any particular age may be of particular importance in describing fertility distributions such as those in the Swedish time sequences, in which low order births influence the distribution well above age 25 for most of the period analyzed.

60. Mean number of children per woman for the "natural" schedules ranges from 10.9 to 6.2, compared to a high of 7.76 for MX20.

61. When Sundbärg's estimated age-specific marital fertility rates for ages 20-49 for five-year periods, 1750-1890, are appended to the 68-single-year series for 1892-1959, and the whole sequence fitted by the same EHR procedures used in this report, the 29 values of α_i and β_i covering the first 140 years fluctuate slightly around the values of α_i and β_i for 1892-1894.

62. If the Coale model fits a schedule perfectly, the value of m will be the same at all ages, indicating that the population follows the standard age pattern of decline of fertility with uniform intensity. The calculated values of m for these Swedish histories do show variability with age for any given year and vary in different ways in different periods, probably due, at least in part, to the effect that changing age patterns of marriage and entry into childbearing have had on the cross-sectional schedules.

63. A new procedure for determining a single value of m by regression (A. J. Coale, personal communication) also emphasizes the shape of the schedule over the central ages of childbearing and omits some higher ages entirely. By removing dependence of M on age group 20-24 (but leaving v(a) pegged to that age group), this procedure provides for these Swedish schedules a time sequence of m with the same pattern of variations as those in Fig. 10 but with values ranging from .2 to 1.7. All of these procedures appear, then, to pick up the same pattern of change over time in these schedules. The EHR standard form parameter β_i^* appears, however, to register more fully in a single parameter (see pages 41 and 42) the change in age distribution of fertility associated with limitation of births than do the Coale-Trussell procedures, which variously divide the force of this change between m acting on v(a) and M acting on n(a), depending on the method of determining m. This division may, in some instances, be of consequence in the use of the Coale-Trussell model fertility schedules, which incorporate the Coale model of marital fertility except for omission of M:

$$f(a) = G(a)n(a)e^{m \cdot v(a)}$$

where $f(a)$ = age-specific overall fertility rate
$G(a)$ = age-specific proportion ever-married.

To the extent that variable aspects of marital fertility would have

been incorporated in M acting on n(a), these aspects would be absorbed by G(a), thus attributing to the age pattern of marriage some of the variation in the overall fertility distribution actually due to marital fertility.

64. Total rate of marital fertility refers to the sum of age-specific rates for the five-year age groups, $\sum_{j=1}^{7} f(a)$, expressed in the rate for women at age cut 49/50.

65. See Mosteller and Tukey, op. cit. in footnote 8, pp. 268-270, for discussion of this use of regression.

66. Impressions of change in marriage patterns are based on a summary of some of Ewbank's findings for single-year birth cohorts of 1851-1922 (Ewbank, loc. cit. in footnote 14). To be consistent with the designation for childbearing cohorts by five-year age group used throughout the present work, the mean and variance of age at marriage for cohorts aged 15-19 in a given year is taken here as the average of Ewbank's values for the five cohorts comprising those who became 15-19 in that year.

Year at Age 15-19	Age at Marriage	
	μ(range)	σ(range)
	years	years
1876-1879	27.71-27.61	6.62-6.60
1880-1891	27.54-27.28	6.60-6.46
1892-1911	27.22-27.09	6.53-6.34
1912-1917	27.25-27.63	6.64-7.09
1918-1921	27.70-27.75	7.07-6.75
1922-1929	27.64-26.59	6.56-5.47
1930-1941	26.41-24.46	5.37-4.58

67. Significant levels of widowhood or divorce at childbearing ages would, of course, also add to the degree of positive skewness of the overall fertility distribution.

68. This expression of mean number of children per married woman is synthetic first of all in the same sense as is the cross-sectional TFR (expressed by (XT)(5)): it sums the fertility experience of women in a number of different age cohorts at a given time as if this were the childbearing experience of a single cohort over time. The expression is synthetic in a second sense also: it sums at a given time the fertility experience by age of those women who were then married, as if all of these women had been in a single marriage cohort which started childbearing at age 15-19 and bore children at each age at the same rate as did those who were actually married at that age.

69. Further work shows that both a legitimate fertility sequence and a "married or actively cohabiting" sequence can be as well fit by EHR analysis as can the overall and marital fertility sequences analyzed here (Breckenridge, unpublished).

70. Recall that two triangles of incomplete cohorts are omitted--one at the beginning of the sequence, one at the end--so that a matrix of 185 years provides 155 complete cohorts.

71. Breckenridge and Orav, loc. cit. in footnote 43.

72. The comparability of X20 and C20 time parameters can be established in the same way from the appropriate pairs of A^*_j and B^*_j age vectors, which are shown for all sequences in Appendix B, Table B2.

73. The 1942-1945 aberration in α^*_i for both X15 and X20 and the post-1950 rise in this parameter for X15 have already been noted in the shorter time sequences, XX15 and XX20 (see page 36).

74. There, after 1909, only the cohorts of 1914, 1917, 1918, 1920 and 1921 fit the model without definite departure for at least one age cut.

75. Ewbank, loc. cit. in footnote 66.

76. Hofsten and Lundstrom, op. cit. in footnote 10.

77. The year 1792 is notable for the assassination of King Gustavus III after a period of political unrest. Severe famine is variously recorded for years from 1780 to 1785. (Thomas, D.S., 1949, Social and Economic Aspects of Swedish Population Movements, 1750-1933, New York: Macmillan, pp. 81-88, 102-108, identifies 1780-1783 and 1785 as years of major crop failures. Utterstrom, G., 1954, "Some population problems in pre-industrial Sweden," Scandinavian Economic History Review 2, pp. 103-165, questions the harvest index which was Thomas' criterion, and identifies 1783-1785 as famine years.) Whether these circumstances affected the actual level and age distribution of births in the years in question, or whether they led to errors in population estimates or in recording of births is open to investigation.

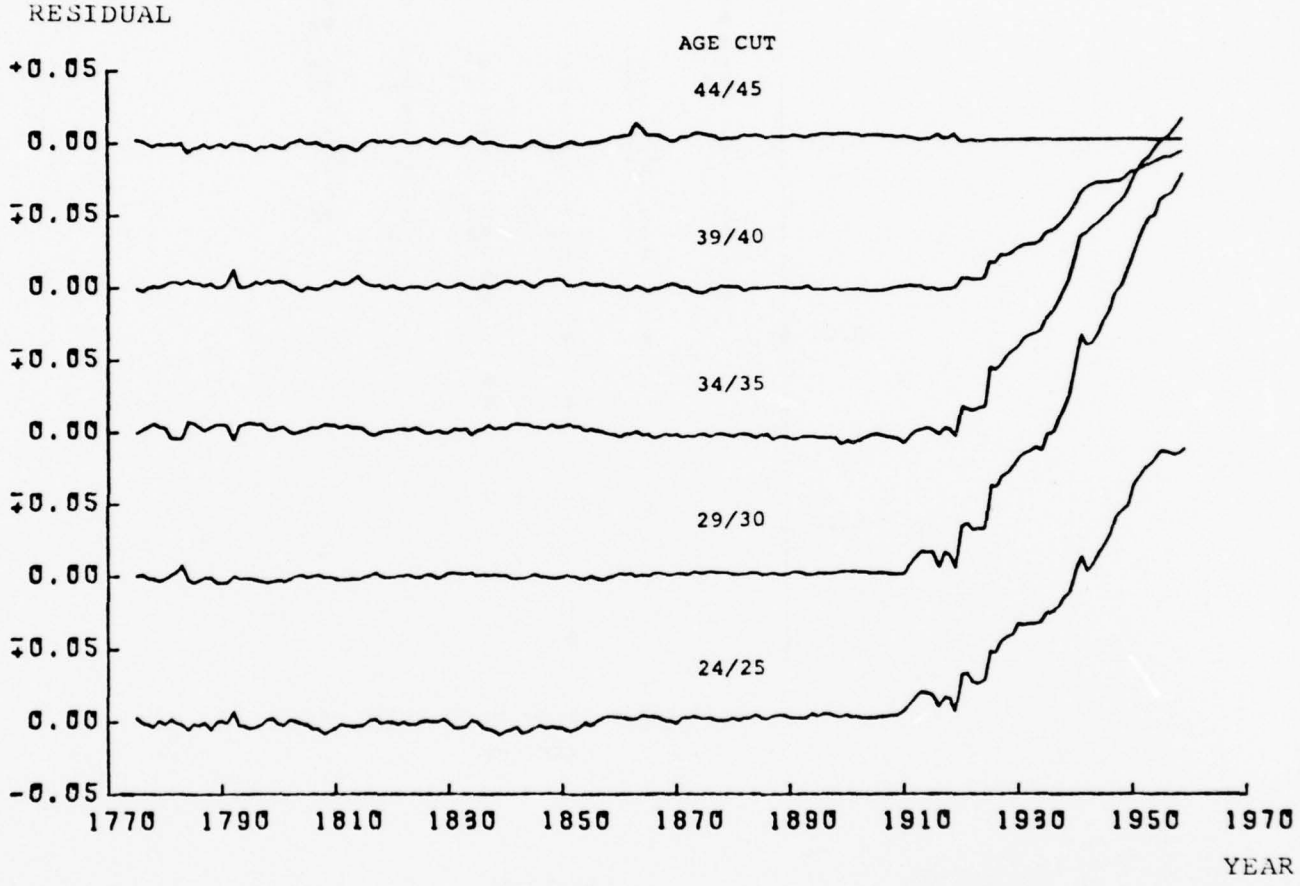
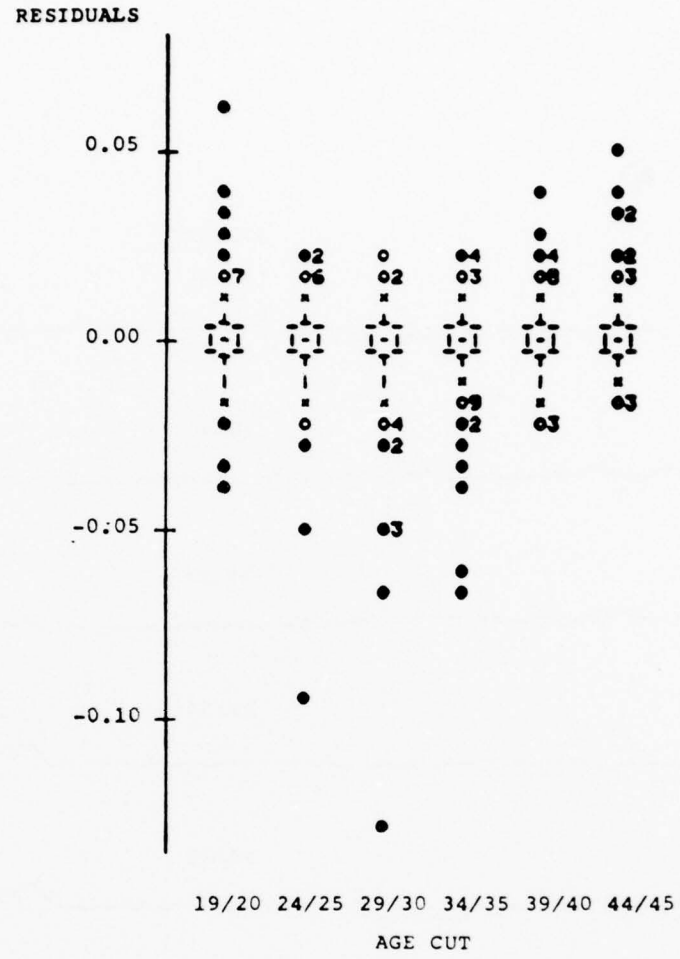
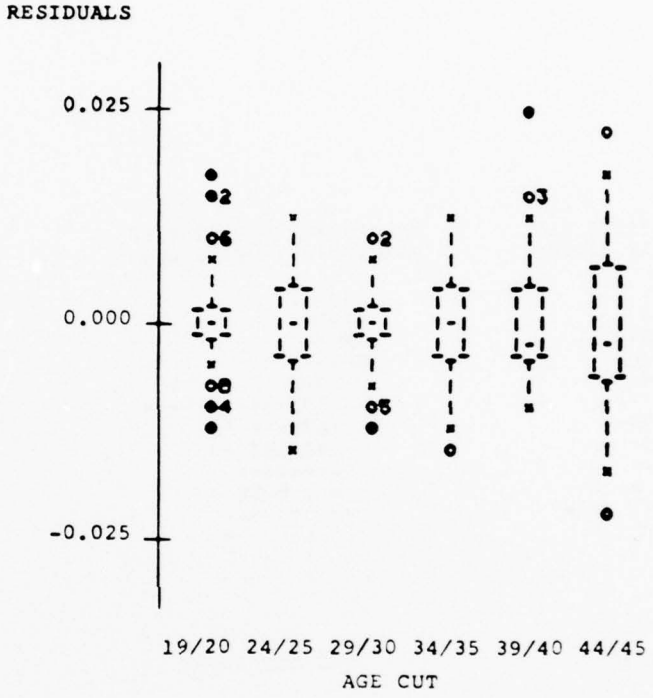


Fig. 1. Time sequence plot of residuals by age cut from EHR fitting (with c = 6 in the biweight) of $f_{ij} = \alpha_i A_j + \beta_i B_j$ to the cross-sectional age 20-49 overall fertility sequence, 1775-1959 (with data expressed on the raw fraction scale).



AGE CUT	RESIDUALS				
	Minimum	Lower quartile	Median	Upper quartile	Maximum
19/20	-.0366	-.0060	-.0012	.0042	.0637
24/25	-.0956	-.0062	-.0015	.0054	.0240
29/30	-.1292	-.0068	.0004	.0058	.0236
34/35	-.0670	-.0048	.0000	.0062	.0265
39/40	-.0223	-.0057	.0002	.0057	.0390
44/45	-.0180	-.0054	.0011	.0044	.0515

Fig. 2. Schematic plots of residuals by age cut (folded square root scale) from EHR fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the cohort age 15-49 overall fertility sequence, 1775-1929.



AGE CUT	RESIDUALS				
	Minimum	Lower quartile	Median	Upper quartile	Maximum
19/20	-.0138	-.0022	.0004	.0029	.0176
24/25	-.0157	-.0052	.0000	.0056	.0119
29/30	-.0136	-.0032	.0006	.0029	.0109
34/35	-.0152	-.0052	-.0008	.0051	.0128
39/40	-.0109	-.0051	-.0015	.0040	.0258
44/45	-.0234	-.0069	-.0017	.0068	.0241

Fig. 3. Schematic plots of residuals by age cut (folded square root scale) from EHR fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the cross-sectional age 15-49 overall fertility sequence, 1775-1959.

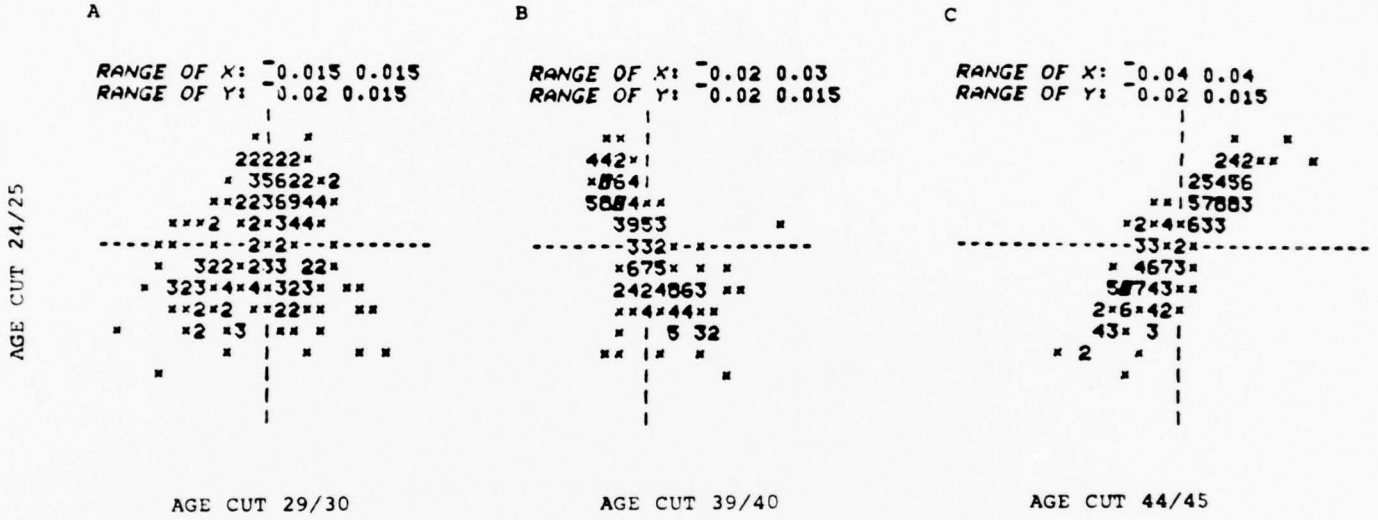


Fig. 4. Scatter plots of residuals for pairs of age cuts (folded square root scale) from EHR fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the cross-sectional age 15-49 overall fertility sequence, 1775-1959.

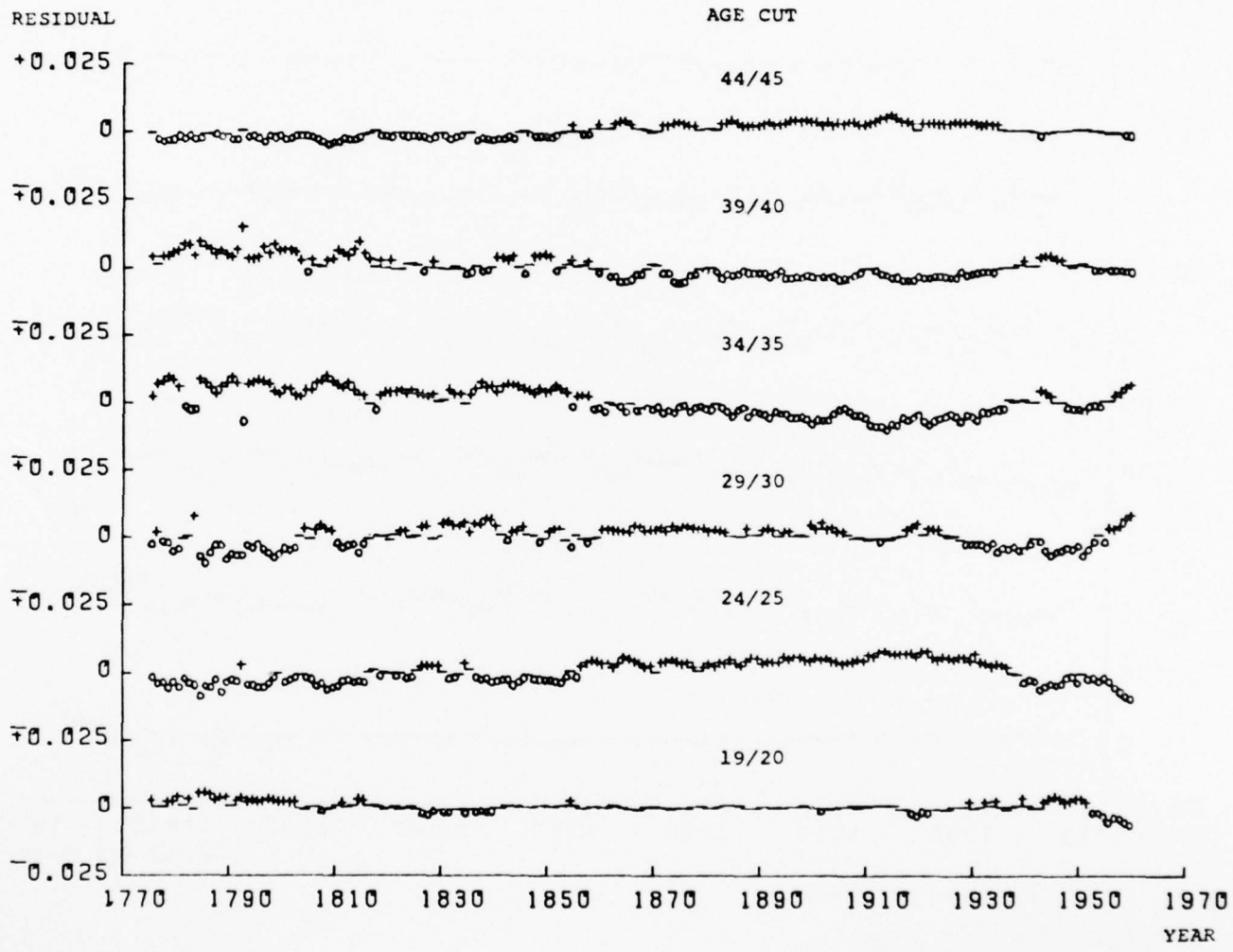


Fig. 5. Time sequence plot of coded residuals by age cut (raw fraction scale) from EHR fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the cross-sectional age 15-49 overall fertility sequence, 1775-1959.

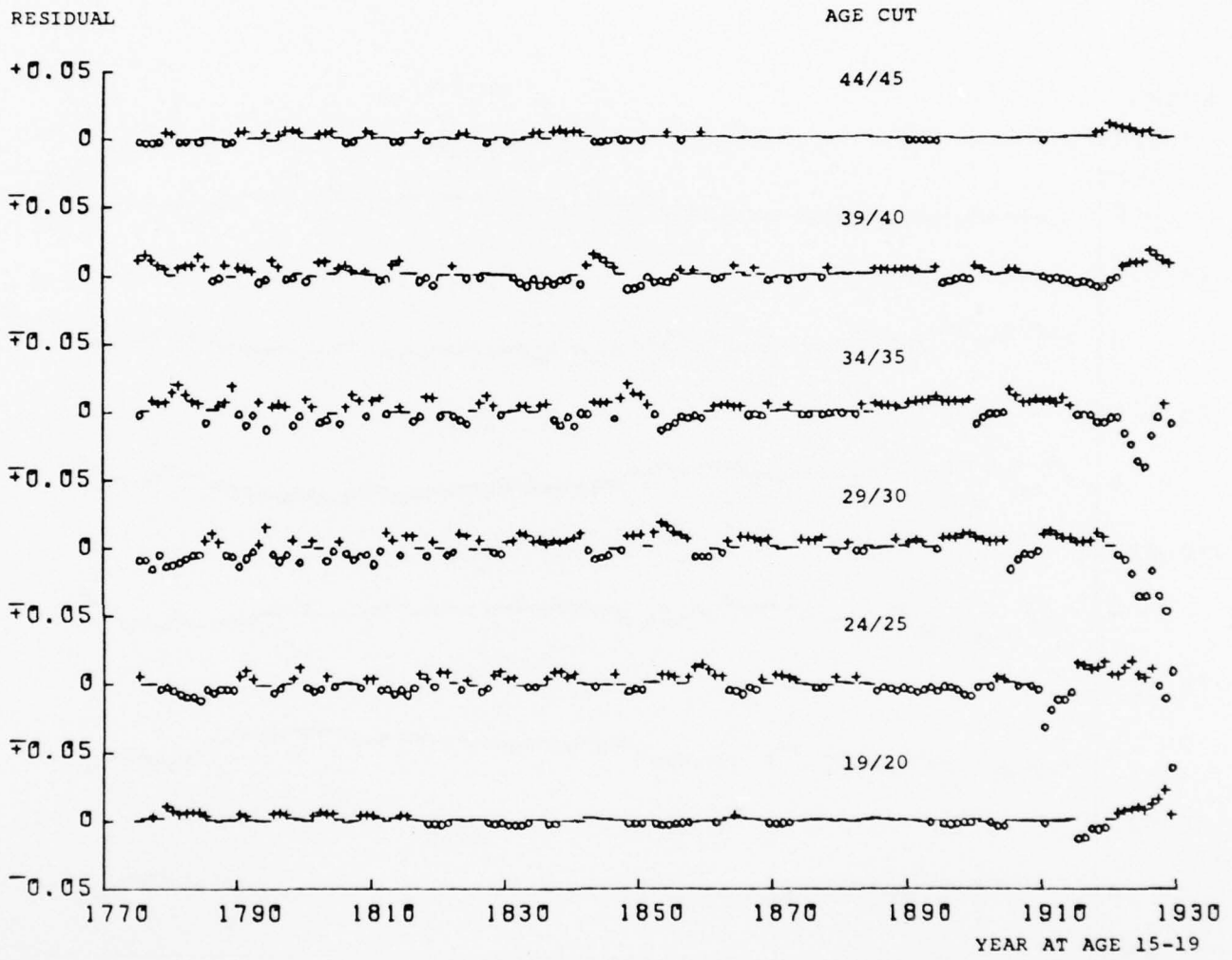


Fig. 6. Time sequence plot of coded residuals by age cut (raw fraction scale) from EHR fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the cohort age 15-49 overall fertility sequence, 1775-1929.

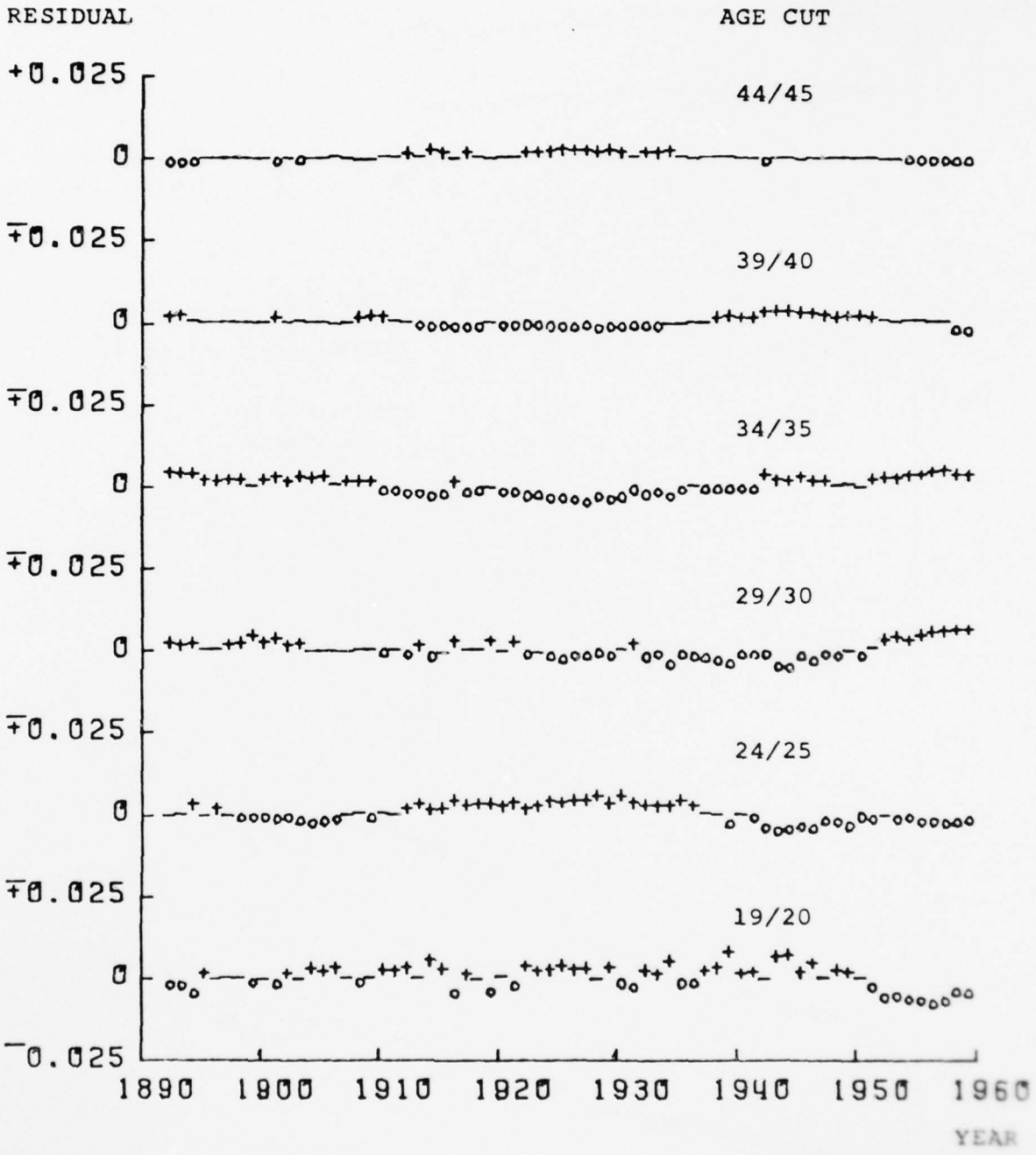
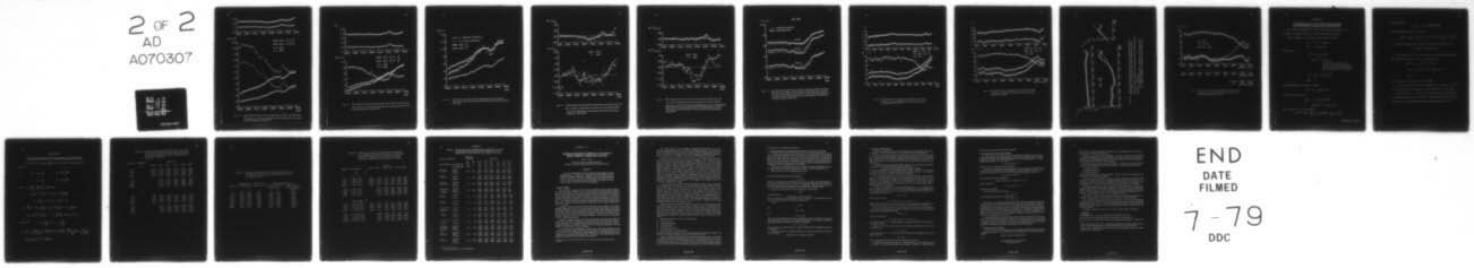


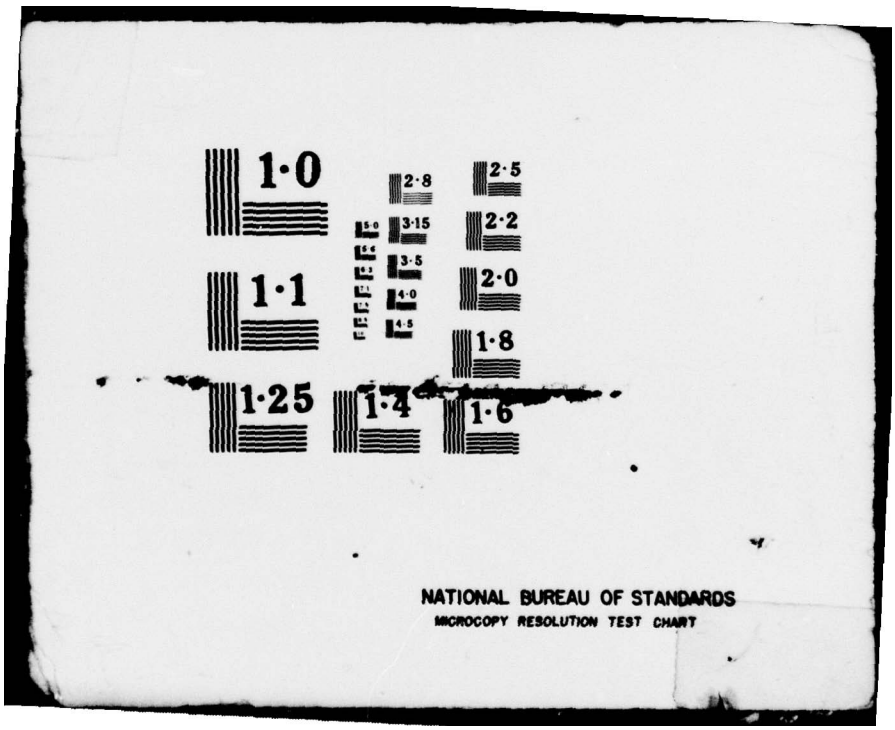
Fig. 7. Time sequence plot of coded residuals by age cut (raw fraction scale) from EHR fitting of $F_{ij} = \alpha_i A_j + \beta_i B_j$ to the cross-sectional age 15-49 marital fertility sequence, 1892-1959.

AD-A070 307 PRINCETON UNIV N J DEPT OF STATISTICS F/G 6/16
AN EMPIRICAL HIGHER-RANK ANALYSIS MODEL OF THE AGE DISTRIBUTION--ETC(U)
MAY 78 M B BRECKENRIDGE, J W TUKEY DAAG29-76-G-0298
UNCLASSIFIED TR-143-SER-2 ARO-14244.6-M NL

2 OF 2
AD
A070307



END
DATE
FILMED
7-79
DDC


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

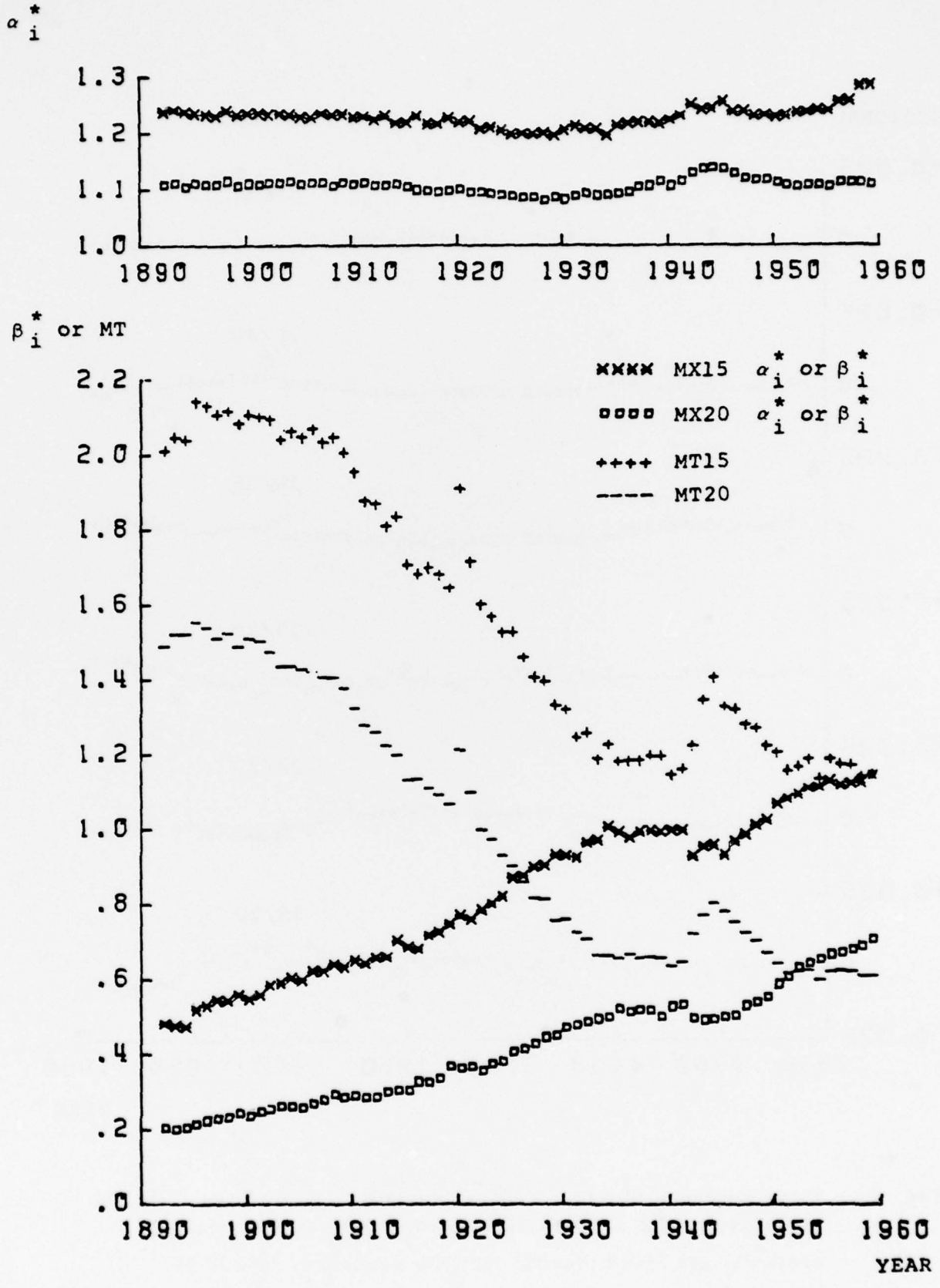


Fig. 8. EHR standard form time parameters α_i^* and β_i^*, and total rate of fertility, for cross-sectional age 15-49 and age 20-49 marital fertility sequences, 1892-1959.

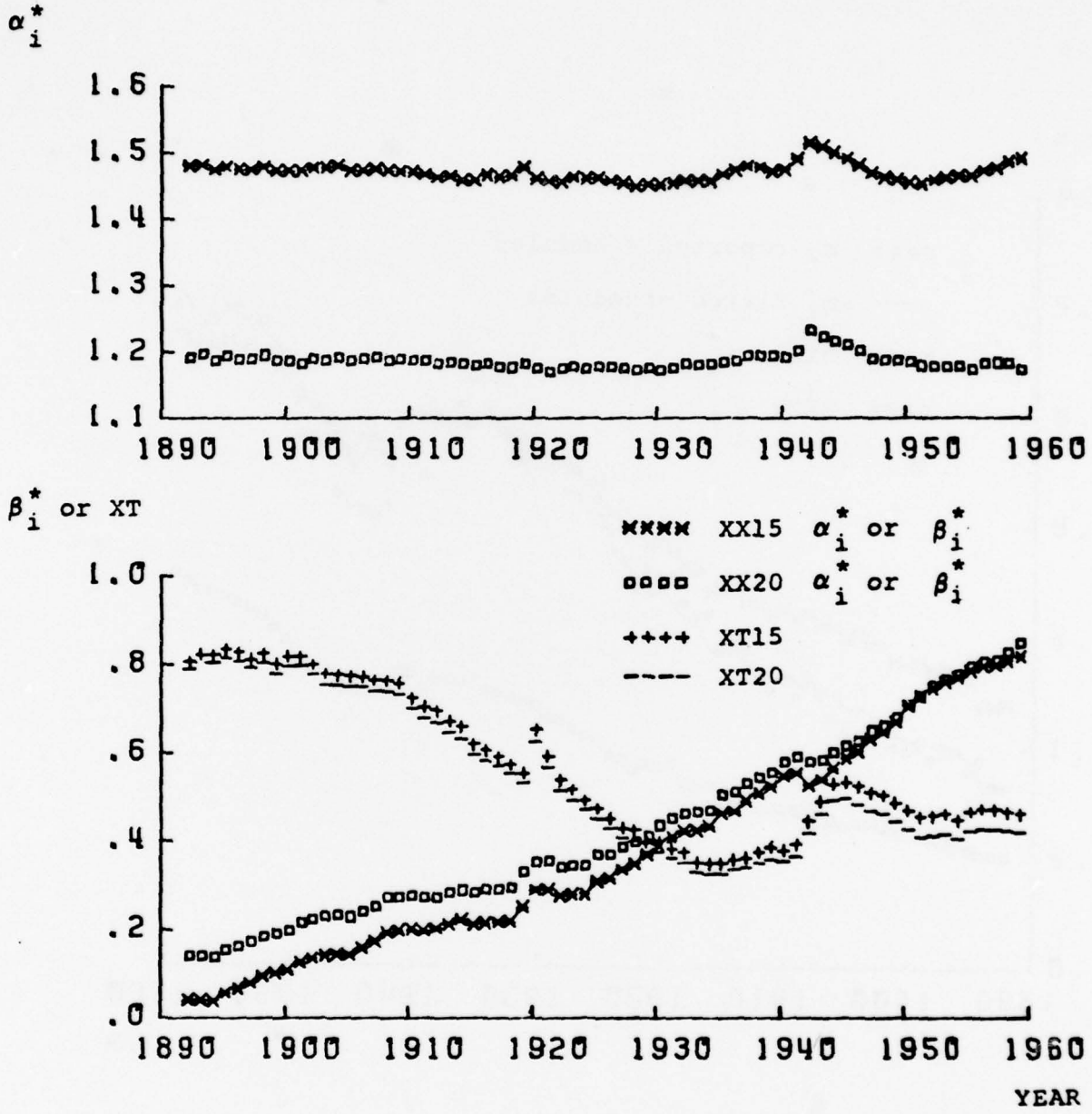


Fig. 9. EHR standard form time parameters α_i^* and β_i^*, and total rate of fertility, for cross-sectional age 15-49 and age 20-49 overall fertility sequences, 1892-1959.

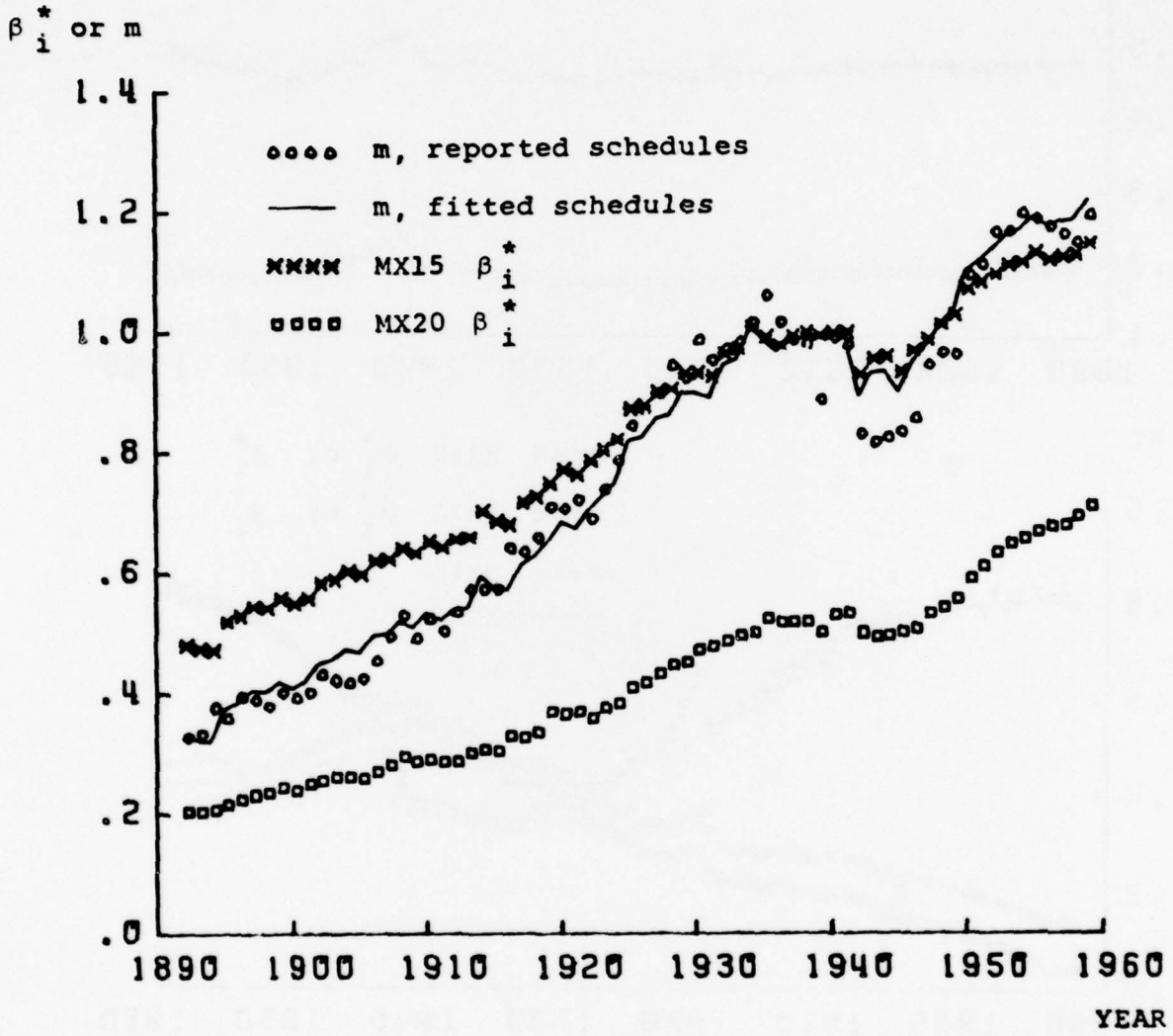


Fig. 10. *EHR model and Coale model expressions of the degree of* skewness of the cross-sectional marital fertility distributions, 1892-1959.

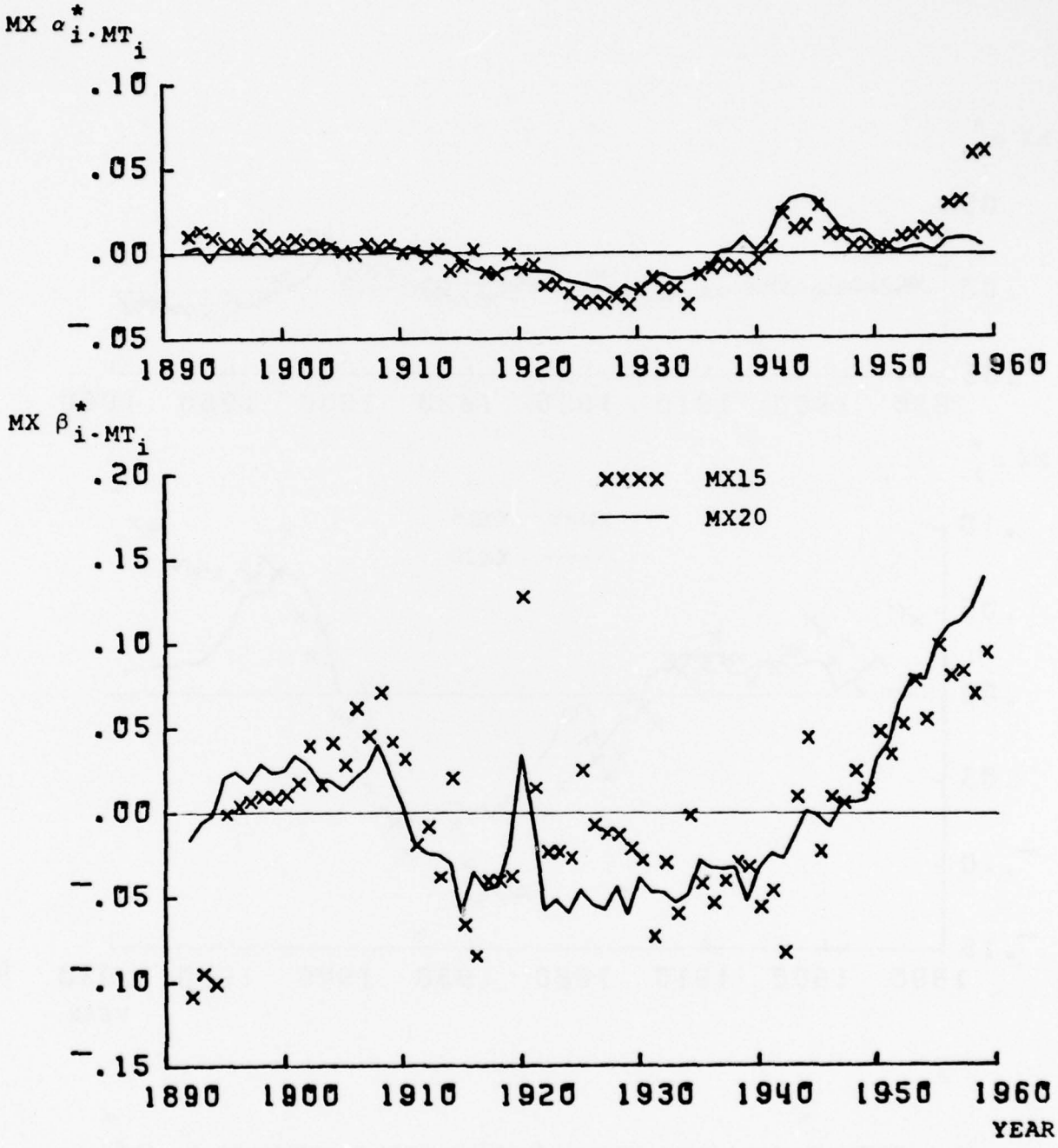


Fig. 11. EHR standard form marital fertility time parameters α^*_i and β^*_i, linearly compensated for total rate of marital fertility (cross-sectional age 15-49 and age 20-49 marital fertility sequences, 1892-1959).

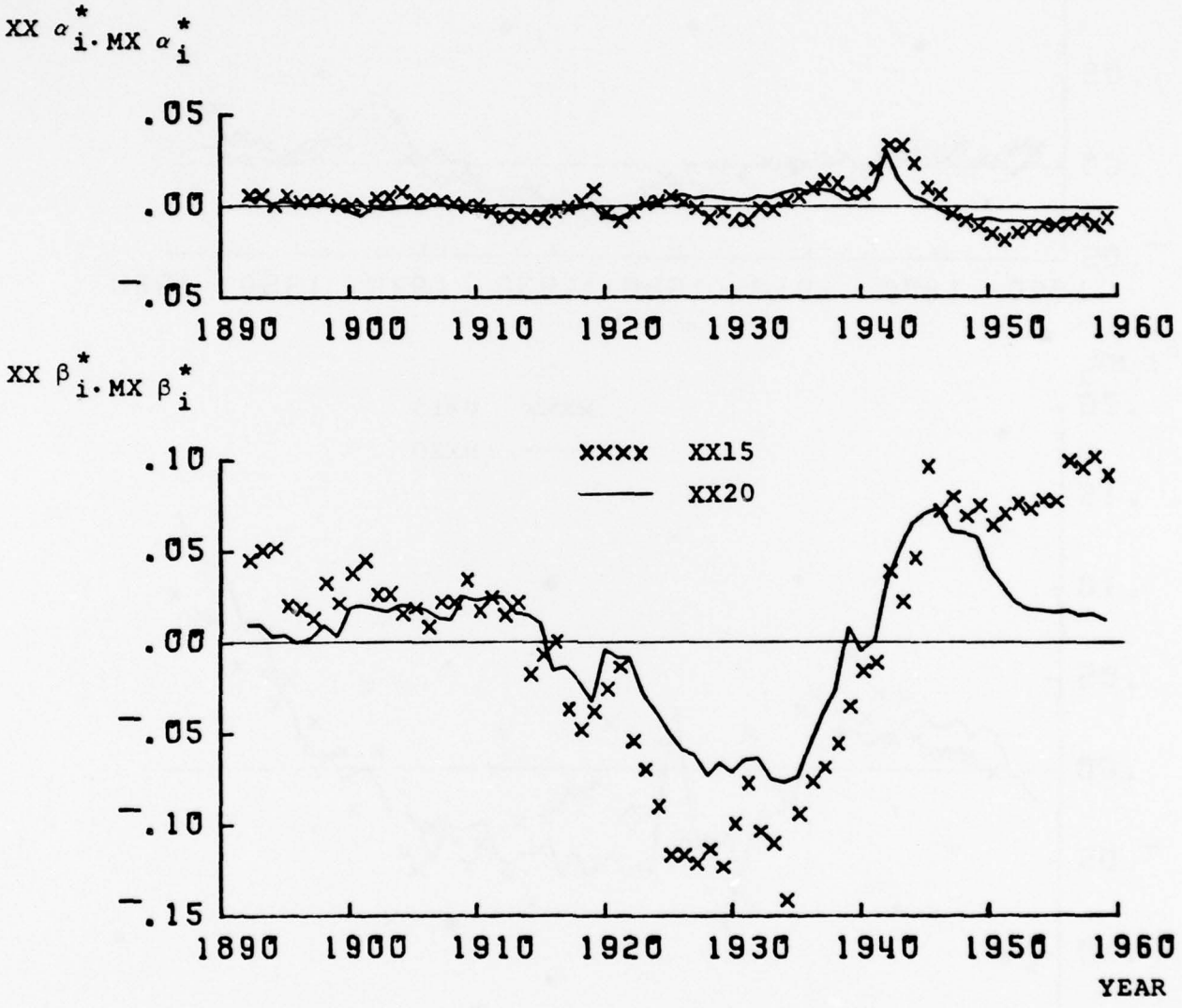


Fig. 12. EHR standard form overall fertility time parameters α^*_i and β^*_i (cross-sectional age 15-49 and age 20-49 overall fertility sequences, 1892-1959) linearly compensated for the corresponding EHR standard form marital fertility time parameters (cross-sectional age 15-49 and age 20-49 marital fertility sequences, 1892-1959).

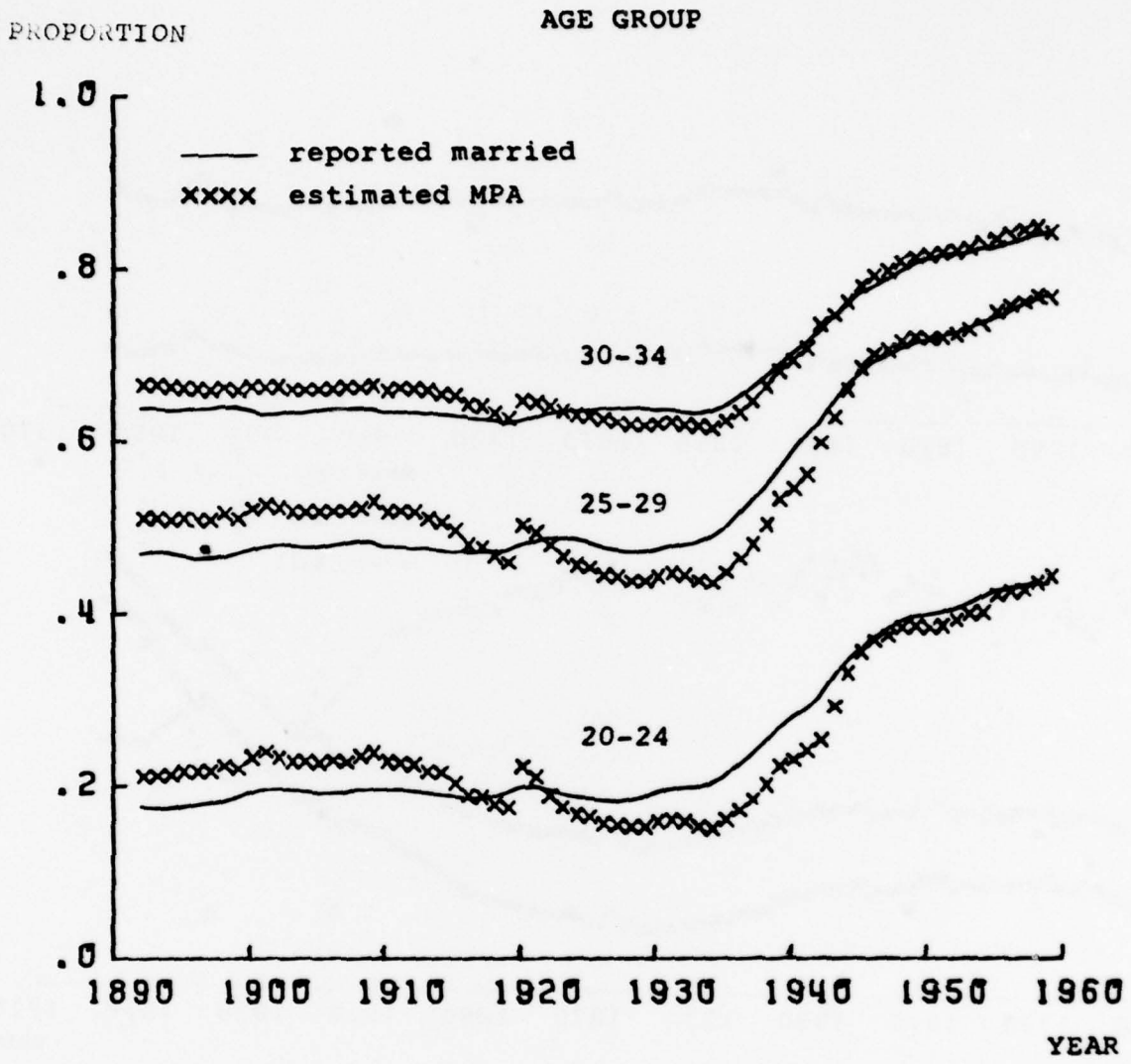


Fig. 13. Age-specific proportions of women reported married, and age-specific proportions of married plus cohabiting active women estimated from EHR-derived level-compensated distributions of overall and marital fertility and the total rates of overall and marital fertility, 1892-1959.

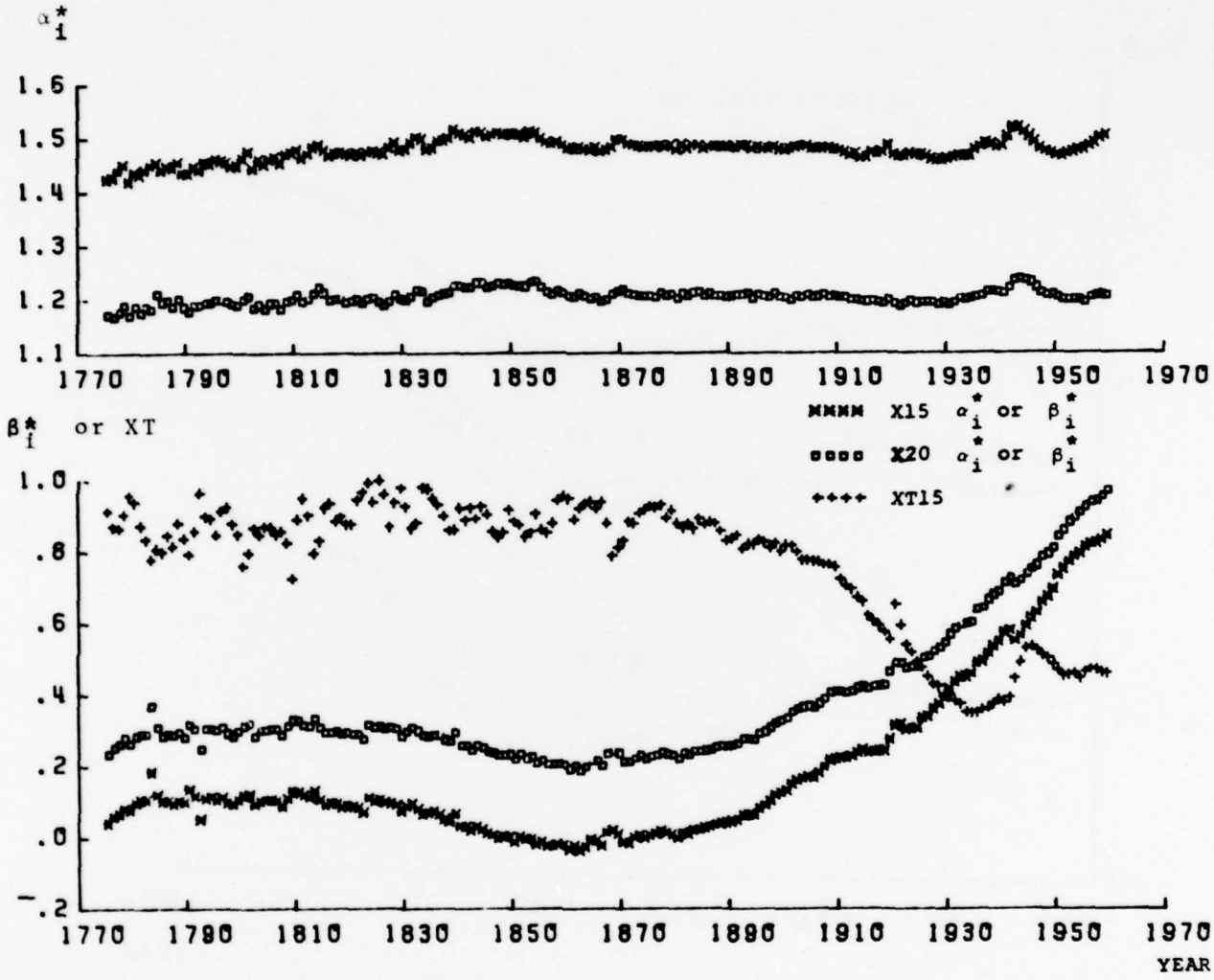


Fig. 14. EHR standard form time parameters α_i^* and β_i^*, and total rate of fertility, for cross-sectional age 15-49 and age 20-49 overall fertility sequences, 1775-1959.

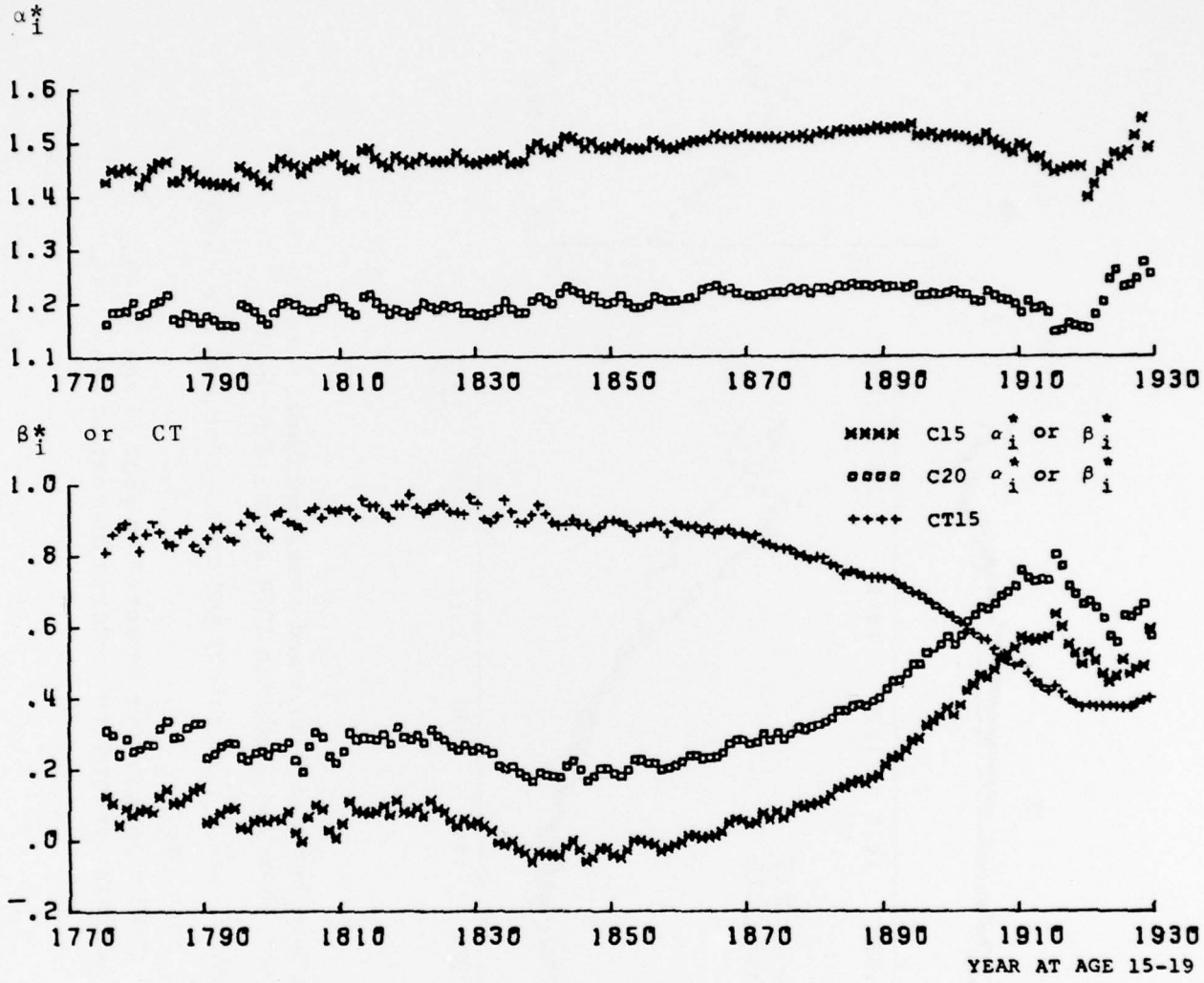


Fig. 15. EHR standard form time parameters α_i^* and β_i^*, and total rate of fertility, for cohort age 15-49 and age 20-49 overall fertility sequences, 1775-1929.

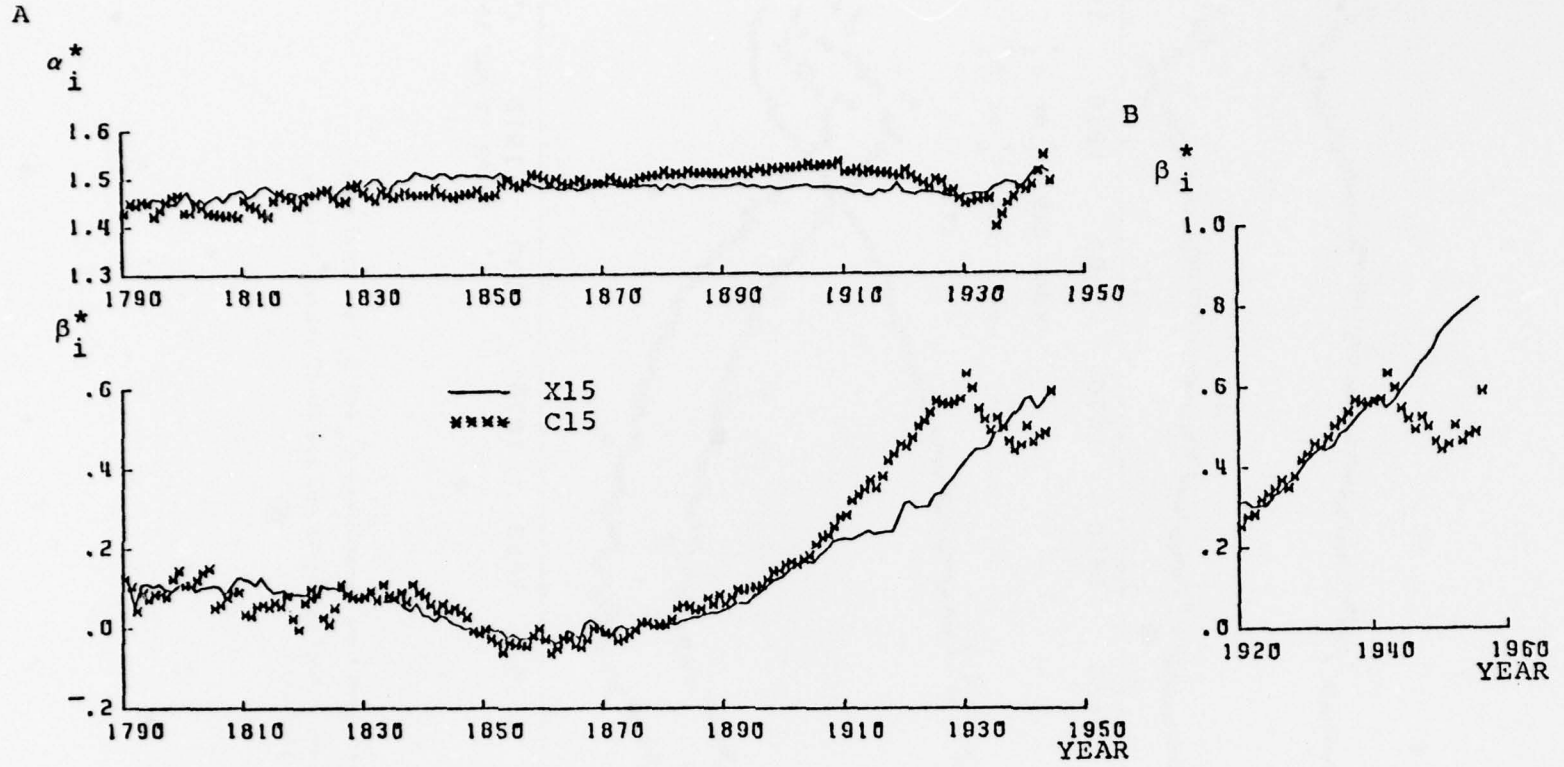


Fig. 16. Some relations between cohort and cross-sectional fertility distributions, demonstrated with EHR standard form time parameters α_i^* and β_i^* for cohort and cross-sectional age 15-49 sequences, 1775-1959.

A. Cohort parameters centered on year at age 30-34.

B. Cohort parameters centered on year at age 42-46.

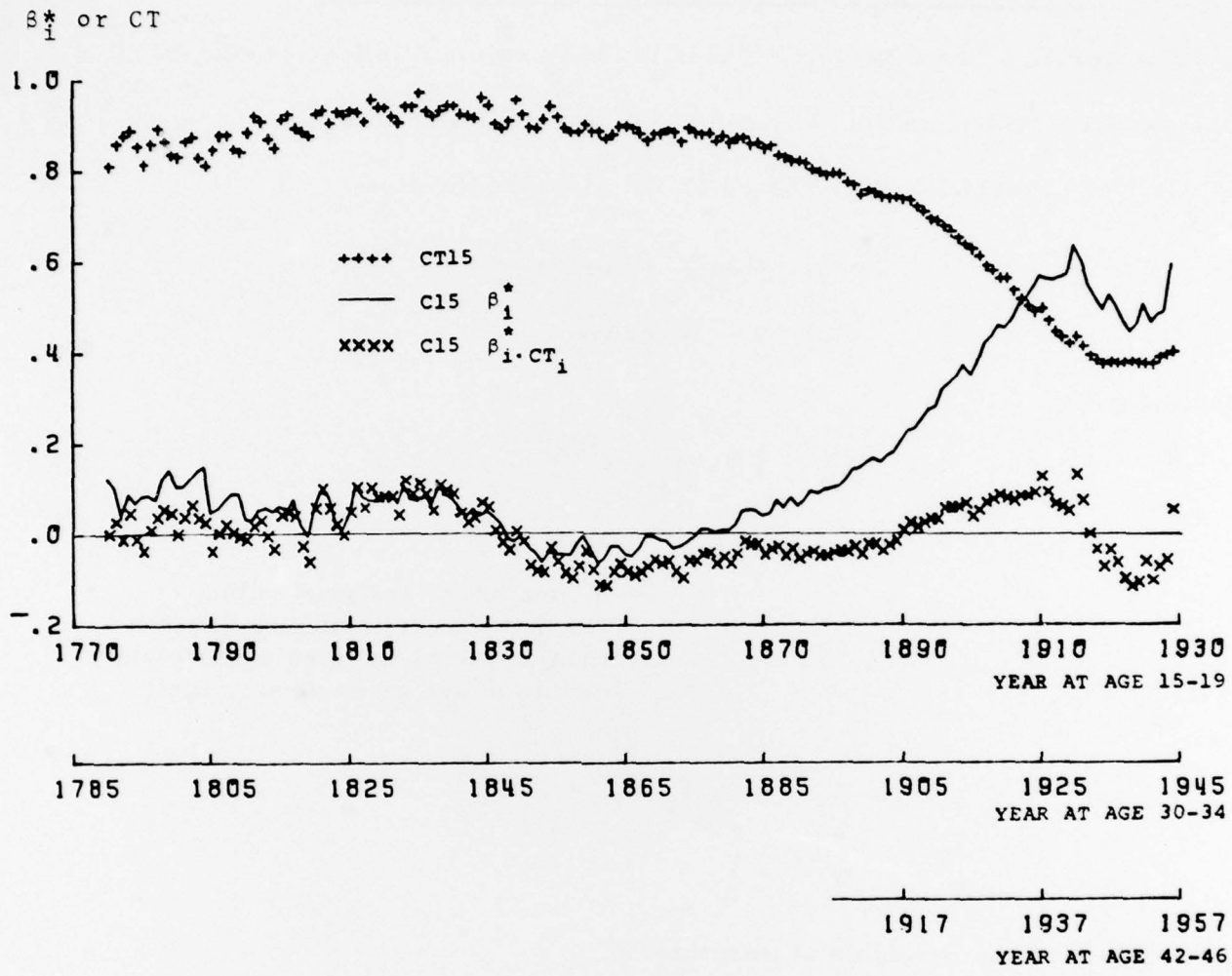


Fig. 17. EHR standard form time parameter β_i^* before and after linear compensation for the total rate of fertility, cohort age 15-49 overall fertility sequence, 1775-1929.

APPENDIX A

A Robust/Resistant Procedure for the Iterative Fitting of Two Multiplicative Components to an M x N Matrix

The iterative fitting procedure used in fitting two multiplicative components to a fertility matrix, F_{ij}, can be summarized as follows:

Cellwise weights w_{ij} are based on the bisquare function

$$w(u) = (1-u^2)^2 \quad \text{for } |u| \leq 1$$

$$w(u) = 0 \quad \text{otherwise}$$

Starting with

$$w_{ij}^{(0)} = [1 - (F_{ij}/cs^{(0)})^2]^2$$

where $s^{(0)}$ = median F_{ij}

c = a constant of assigned value (so that residuals of size greater than c times the median absolute deviation are given zero weight)

$$A_j^{(0)} = \sum_{j=1}^{c} F_{ij} w_{ij}^{(0)} \Big/ \sum_{j=1}^{c} w_{ij}^{(0)}$$

$$\alpha_i^{(0)} = 1$$

and designating the residuals at iteration m

$$z_{ij}^{(m)} = F_{ij} - (\alpha_i^{(m)} A_j^{(m)})$$

and weights at iteration m

$$w_{ij}^{(m)} = [1 - (z_{ij}^{(m)}/cs^{(m)})^2]^2$$

where $s^{(m)}$ = median $z_{ij}^{(m)}$

the estimators of A_j are improved

$$A_j^{(m+1)} = A_j^{(m)} + \sum_{j=1}^{c} z_{ij}^{(m)} w_{ij}^{(m)} \alpha_i^{(m)} \Big/ \sum_{j=1}^{c} w_{ij}^{(m)} \alpha_i^{(m)}$$

APPENDIX A - cont'd.

and standardized

$$A_j^{(m+1)} = A_j^{(m+1)} \Big/ \sqrt{\Sigma (A_j^{(m+1)})^2}$$

and the estimators of α_i are improved

$$\alpha_i^{(m+1)} = \alpha_i^{(m)} + \sum_{i=1}^{r} z_{ij}^{(m+1)} w_{ij}^{(m+1)} A_j^{(m+1)} \Big/ \sum_{i=1}^{r} w_{ij}^{(m+1)} A_j^{(m+1)}$$

Iterations continue until a selected convergence criterion is met

$$1 - [\Sigma (cs^{(m+1)})^2 / \Sigma (cs^{(m)})^2] < \varepsilon$$

The residuals $z_{ij}^{(m+n)}$ from this portion of the fitting procedure are then examined in the same way for B_j, beginning with

$$w_{ij}^{(m+n)} = [1 - (z_{ij}^{(m+n)} / cs^{(m+n)})^2]^2$$

$$B_j^{(0)} = \sum_{j=1}^{c} z_{ij}^{(m+n)} w_{ij}^{(m+n)} \Big/ \sum_{j=1}^{c} w_{ij}^{(m+n)}$$

$$\beta_i^{(0)} = 1$$

and iterating to convergence in $B_j^{(p)}$ and $\beta_i^{(p)}$.

The two-stage procedure is then repeated for the residuals $z_{ij}^{(m+n+p)}$, and so on iteratively to convergence in final estimates of A_j α_i, B_j, β_i.

Optimal values of c appear to vary somewhat with the data and the desired degree of resistance to outliers. Values between 6 and 9 are commonly useful. Least squares estimators are approached as $c \longrightarrow \infty$.

APPENDIX B

A Selected Standard Form Re-presentation of a Rank-Two Fit

To re-present $F_{ij} = \alpha_i A_j + \beta_i B_j$ as an identically equivalent $\alpha^{**}_i A^{*}_j + \beta^{**}_i B^{*}_j$,

let $\alpha^{**} = a\alpha + b\beta$ $\qquad \alpha = \dfrac{\alpha^{**} - b\beta}{a}$

$A^{*} = cA + dB$ $\qquad \beta = \dfrac{A^{*} - dB}{c}$

Then $F_{ij} = \left(\dfrac{\alpha^{**}}{a} - \dfrac{b\beta}{a}\right)\left(\dfrac{A^{*}}{c} - \dfrac{dB}{c}\right) + \beta B$

$$= \frac{1}{ac}\left(\alpha^{**}A^{*} - b\beta A^{*} - \alpha^{**}dB\right) + \left(1 + \frac{bd}{ac}\right)\beta B$$

If $\left(1 + \dfrac{bd}{ac}\right)q = \dfrac{d}{c}$ and $\left(1 + \dfrac{bd}{ac}\right)p = \dfrac{b}{a}$.

$$F_{ij} = \frac{\alpha^{**}A^{*}}{ac} - \frac{p}{c}\left(1 + \frac{bd}{ac}\right)\beta A^{*} - \frac{q}{a}\left(1 + \frac{bd}{ac}\right)\alpha^{**}B + \left(1 + \frac{bd}{ac}\right)\beta B$$

$$= \left[\frac{1}{ac} - \frac{pq}{ac}\left(1 + \frac{bd}{ac}\right)\right]\alpha^{**}A^{*} + \left(1 + \frac{bd}{ac}\right)\left(\beta - \frac{q}{a}\alpha^{**}\right)\left(B - \frac{p}{c}A^{*}\right)$$

Substituting $q = \dfrac{d}{c(1+\frac{bd}{ac})}$ and $p = \dfrac{b}{a(1+\frac{bd}{ac})}$,

$$F_{ij} = \left[\frac{1}{ac} - \frac{db}{a^2c^2(1+\frac{bd}{ac})^2}\left(1+\frac{bd}{ac}\right)\right]\alpha^{**}A^{*} + \left(1+\frac{bd}{ac}\right)\left[\beta - \frac{d\alpha^{**}}{ac(1+\frac{bd}{ac})}\right]\left[B - \frac{bA^{*}}{ac(1+\frac{bd}{ac})}\right]$$

$$= \left(\frac{1}{ac+bd}\right)\alpha^{**}A^{*} + \left(1 + \frac{bd}{ac}\right)\beta^{**}B^{*}$$

Table B1. Fitted Age Parameters (folded square root scale) Derived by EHR Analysis of the Age Distribution of Overall and Marital Fertility in Selected Time Sequences, 1775-1959

Sequence	Parameter	Age Cut					
	A_j	19/20	24/25	29/30	34/35	39/40	44/45
X15		-.5835	-.3504	-.0889	.1620	.3949	.5888
XX15		-.5331	-.2570	.0108	.2453	.4530	.6199
C15		-.5810	-.3478	-.0845	.1669	.3986	.5895
MX15		-.1093	.1154	.2802	.4190	.5449	.6510
X20			-.4620	-.1323	.1816	.4729	.7159
XX20			-.3393	-.0055	.2793	.5284	.7264
C20			-.4467	-.1149	.1969	.4829	.7178
MX20			-.1429	.1431	.3661	.5595	.7156
	B_j						
X15		.2820	.5047	.5522	.4772	.3329	.1493
XX15		.3579	.5229	.5502	.4639	.2818	.0349
C15		.2177	.4860	.5623	.4976	.3627	.1455
MX15		.6627	.5339	.3907	.2171	-.0120	-.2754
X20			.6255	.6079	.4404	.1954	-.0834
XX20			.5862	.6048	.4780	.2344	-.0850
C20			.5584	.6029	.4885	.2926	.0217
MX20			.6314	.5783	.4143	.1175	-.2855

Table B2. Results of Regressions to Fix EHR-Fitted Fertility Distribution Parameters for Re-presentation of Fits in a Standard Form

Sequence	LS Regression Coefficients		Degree of Fit (linear scale)	Canonical Regression Elements of First Eigenvector		Eigenvalue for First Canonical Variate
	a	b		c	d	
X15	.6867	.0062	.9903	.9875	-.0899	.9989
X20	.8403	-.2862	.9921	.8948	-.2724	.9996
XX15	.6880	-.0997	.9932	.9532	-.2536	.9997
XX20	.8061	-.2560	.9934	.9321	-.3562	.9999
C15	.9505	-.0249	.9828	.9934	-.0886	.9989
C20	1.075	-.2428	.9846	.9415	-.2808	.9997
MX15	.6951	-.4293	.9893	.8070	-.5936	.9994
MX20	.7808	-.5033	.9915	.9463	-.3288	.9999

Table B3. Fitted Age Parameters (folded square root scale) After Standard Form Re-presentation of EHR Fits to the Age Distributions of Overall and Marital Fertility in Selected Time Sequences, 1775-1959

Sequence	Parameter		Age Cut 19/20	24/25	29/30	34/35	39/40	44/45
	α_i^*							A_j^*
	μ	σ						
X15	1.476	0.019	-.6016	-.3915	-.1375	.1170	.3600	.5680
XX15	1.468	0.013	-.5989	-.3776	-.1292	.1158	.3603	.5821
C15	1.479	0.030	-.5965	-.3886	-.1337	.1217	.3639	.5728
MX15	1.226	0.018	-.4816	-.2238	-.0058	.2093	.4468	.6888
X20	1.205	0.013		-.5838	-.2839	.0425	.3699	.6633
XX20	1.187	0.011		-.5250	-.2206	.0898	.4087	.7075
C20	1.203	0.023		-.5774	-.2775	.0482	.3725	.6697
MX20	1.106	0.013		-.3428	-.0547	.2102	.4908	.7711
	range of β_i^*							B_j^*
X15	-0.039 to 0.843		.2870	.5074	.5525	.4754	.3291	.1438
XX15	0.037 to 0.815		.2801	.4846	.5502	.4983	.3468	.1245
C15	-0.062 to 0.633		.2025	.4771	.5603	.5022	.3733	.1610
MX15	0.473 to 1.144		.5056	.5140	.4788	.4042	.2757	.1075
X20	0.191 to 0.971			.4826	.5802	.5178	.3674	.1654
XX20	0.140 to 0.849			.4575	.5758	.5408	.3834	.1402
C20	0.168 to 0.802			.4527	.5709	.5273	.3975	.1819
MX20	0.204 to 0.711			.4521	.5622	.5451	.4008	.1474

APPENDIX C

Table C1. The Age Distributions of Natural Fertility, Reported[1] and Fitted as Weighted Sums of the A_j^* and B_j^* Derived by EHR Analysis of the Swedish Age 20-49 Marital Fertility Time Sequence for 1892-1959

Source of Distribution		EHR Time Parameter		Age Group					
		α_i^*	β_i^*	20-24	25-29	30-34	35-39	40-44	45-49
EHR Standard A_j^*	with lowest α_i^*	1.080	0	.2473	.2110	.2001	.1891	.1284	.0241
	with highest α_i^*	1.138	0	.2348	.2212	.2107	.1961	.1238	.0134
Hutterites	Reported			.2514	.2294	.2043	.1856	.1015	.0279
1921-1930	Fitted	1.083	.0576	.2633	.2177	.1990	.1807	.1180	.0213
	Residual			-.0119	.0117	.0053	.0049	-.0175	.0066
Canada	Reported			.2358	.2293	.2242	.1899	.1070	.0139
1700-1730	Fitted	1.151	.0354	.2422	.2274	.2119	.1920	.1161	.0105
	Residual			-.0064	.0019	.0123	-.0021	-.0091	.0034
Hutterites	Reported			.2425	.2302	.2169	.1909	.1046	.0148
before 1921	Fitted	1.141	.0495	.2482	.2274	.2098	.1887	.1145	.0114
	Residual			-.0057	.0028	.0071	.0022	-.0099	.0034
Europeans of	Reported			.2562	.2354	.2200	.1773	.1040	.0071
Tunis	Fitted	1.165	.0948	.2561	.2365	.2123	.1840	.1039	.0071
1840-1859	Residual			.0001	-.0009	.0077	-.0067	.0001	.0000
Crulai	Reported			.2643	.2523	.2252	.1682	.0841	.0060
1674-1742	Fitted	1.182	.1650	.2727	.2472	.2125	.1742	.0896	.0039
	Residual			-.0084	.0051	.0127	-.0060	-.0055	.0021
Norway	Reported			.2434	.2336	.2096	.1776	.1106	.0252
1874-1876	Fitted	1.091	.0449	.2577	.2179	.2009	.1836	.1197	.0201
	Residual			-.0143	.0157	.0087	-.0060	-.0091	.0051
Bourgeoisie of	Reported			.2602	.2421	.2187	.1839	.0823	.0127
Geneva	Fitted	1.155	.1296	.2683	.2385	.2092	.1773	.0991	.0075
before 1600	Residual			-.0081	.0036	.0095	.0066	-.0168	.0052
Taiwan	Reported			.2626	.2403	.2201	.1892	.0820	.0058
c.1900	Fitted	1.174	.1287	.2639	.2419	.2126	.1794	.0969	.0053
	Residual			-.0013	-.0016	.0075	.0098	-.0149	.0005
Bourgeoisie of	Reported			.2788	.2576	.2278	.1524	.0749	.0085
Geneva	Fitted	1.163	.2184	.2928	.2490	.2065	.1636	.0833	.0048
1600-1649	Residual			-.0140	.0086	.0213	-.0112	-.0084	.0037
Sotteville-	Reported			.2682	.2514	.2291	.1760	.0698	.0056
Les-Rouen	Fitted	1.200	.1849	.2744	.2526	.2147	.1725	.0837	.0021
1760-1790	Residual			-.0062	-.0012	.0144	.0035	-.0139	.0035
Iran	Reported			.2642	.2475	.2174	.1706	.0870	.0134
1940-1950	Fitted	1.141	.1460	.2762	.2377	.2061	.1733	.0978	.0089
	Residual			-.0120	.0098	.0113	-.0027	-.0108	.0045
India	Reported			.2609	.2326	.2278	.1712	.0808	.0267
1945-1946	Fitted	1.093	.2395	.2769	.2274	.2001	.1735	.1063	.0158
	Residual			-.0160	.0052	.0277	-.0023	-.0255	.0109

[1] Data from Henry, loc. cit. in footnote 18.

APPENDIX D

THE RELATIONSHIP OF EMPIRICAL ANALYSIS TO MORE NARROWLY MODELLED ANALYSIS

John W. Tukey

Princeton University* and Bell Laboratories
Princeton, New Jersey 08540 and Murray Hill, New Jersey 07974

ABSTRACT

If we are to make proper use of both empirical analysis and more narrowly modelled analysis -- in particular to make good use of both EHR on the one hand and the Coale-Trussell fertility schedules (and their analogs and generalizations) on the other -- we need to understand quite clearly both the characteristics of the two approaches and their interrelation. The discussion that follows is intended to be a step toward such understanding.

1. Kinds of "Models"

The word "model" is one of those which means quite different things to different people -- or to the same person at different times. At one extreme it may be both almost completely normative and very precise, as in the mathematical expressions which describe the motions of two (or three) bodies under Newtonian gravitation. Here the discovery of "unexplained" (meaning "beyond the narrow model") deviations can be of great importance, as when the advance of the perihelion of Mercury was vital in the assessment of Einstein's theory of relativity. The existence of such precise normative models almost always seems to depend upon a long series of interactions between experiment or experience on the one hand and concepts and theory on the other.

At another extreme lie "models", like those discussed in the next section, that are highly adaptable, because they involve so many more constants, which can be adjusted to give a good fit and which, because of the diverse kinds of behavior to which these constants are adapted, are thought of almost entirely as providing empirical descriptions. Here the emphasis is on the ability to describe very diverse phenomena in a single way, and the discovery of more or less systematic deviations is often a call to increased flexibility -- to the use of still more general "models" to absorb these deviations.

In general, a "model" seems to tend to contain two elements, the collection of things from which one is to be selected to describe a particular instance (the "stock") and, often, explicit or implied guidance to aid in interpreting the meaning of whichever element of the stock is selected (the "guidance"), although -- especially in the two extremes just discussed -- the latter element is often very weak -- or even nonexistant.

In the context of multiple regression, Mosteller and I (1977) have introduced the word "stock" (the word "posse" has also been used) for the collection of possibilities that are to be fit -- from which one is to be selected as a useful description. In the extremely flexible case just described we are concerned with "broad stocks". (By contrast, a stock involving only a few, hopefully well-selected, constants would be a "narrow stock".)

*Prepared in part in connection with research at Princeton University sponsored by the Army Research Office (Durham).

The second extreme, in which flexibility is emphasized but guidance has yet to be included, is reasonably referred to as involving "broad empirical models". (The fact that guidance is avoided initially need not mean that it cannot be added, as Breckenridge's paper illustrates. It may well be very desirable to start with an emphasis on adaptability and consequent good fit and then move to an emphasis on guidance in the interpretation of specific fits.

Still another extreme is given by relatively narrow (e.g., few-parameter) models where it is felt that the way in which the constants enter into the algebraic or other expressions is such that we can make useful interpretations of changes in any particular one. A good type instance might be compartment models in biology, in which the passage of a traceable substance through the body -- perhaps in and out of the blood stream -- is modelled in terms of very simple differential equations -- differential equations in which only the rate constants are to be fitted to whatever data has been observed. Here a change in one constant may be rightly given a different interpretation than a change in another. However, there need be no feeling that close similarity of actual occurrence to what can be modelled is essential (or even very likely). Deviations, if not too large, are often recognized as something to be anticipated and overlooked. We might call such models "separating models" since their main purpose is to separate information into pieces that, at least hopefully, bear upon separate aspects of what is being studied. Here guidance is an important part of the model. As a consequence, for example, the concern of economists for identifiability is natural.

A very important class of models (unfortunately frequent, as some would say; sometimes hard to separate from the previous class) are those well described as "narrow empirical models". Here we have found a way to describe most of the detail of some behavior in terms of a few constants. This offers two great advantages: first, we can often compare situations more effectively and more intuitively if we have each of them described in terms of only a few numbers. Second, we can usually gain precision by estimating only a few constants from the data, leaving the bulk of the impact of irregularities, deviations, and sampling fluctuations to the residuals. (As compared with broad empirical models, these models will involve fewer constants, perhaps many fewer.) Here, in contrast to separating models, guidance is prominent by its absence, and fit may be less than, or even far from, perfect -- imperfection of fit being accepted in return for the two advantages just cited.

The last class of "models" we shall choose to mention here is that of systems of "transferring models" by whose aid we hope we can effectively transfer what can be learned from observations of very different sorts into common terms. Economists hope that some of their models are of this kind, as when they compare the results of cross-sectional studies with those of studies of time series. Here guidance is likely to be much more important than stock. Bridgeman's (1927) discussion of operational constructs in physics, in which, for example, masses measured in different ways represent different concepts, illustrates the delicacy of such transfers.

In thinking about the list of model types just sketched:

- normative models
- broad empirical models
- separating models
- narrow empirical models
- transferring models (or model systems)

we need to be careful to remember that these have been isolated as characteristic extremes, and that many real situations are likely to be mixtures of at least two of them.

Finally, a mathematician might hope for very frequent successful occurrence of "interpolatory models" with whose aid careful measurement in a few well-selected situations tells us about what will happen in many other (intermediate) situations. Such models are at least very close to being precise normative ones. The models used in such fields as the strength of materials and rates of chemical reactions often come close to doing just this. Outside of classical physical

science, however, such models seem infrequent.

2. The general character of combinational broad stocks, such as those used in EHR

It is important to recognize that the expressions fitted in empirical higher rank analysis (in EHR) are selected from broad stocks and do belong to one or another class of very flexible stocks. These classes make very little, if any, use of our understandings of the mechanism underlying the data. They strive to mobilize their intrinsic flexibility and to be guided by the data in the way that they dispose this flexibility to provide relatively close description. As Breckenridge has emphasized, they are usually well adapted to relatively automatic generalization -- something that can be more difficult for narrower models.

The classes of such flexible models so far of greatest importance are defined in terms of the simplest arithmetic operations, beginning with a single "+" sign (or perhaps two such).

Additive fits with two crossed categories take the form

$$\hat{y}_{ij} = a + b_i + c_j$$

and are often made of the greatest use by writing

$$y_{ij} = q(data_{ij})$$

where q is a well-chosen monotone function (the choice $q(z) = \log z$ is but one frequent example). This sort of approach not only underlies the widespread ramifications of the analysis of variance (perhaps the most widely used of our nonelementary statistical procedures) but also plays a key role in axiomatized fundamental measurement (Luce and Tukey, 1964).

Multiplicative fits are almost a twin sib to additive fits, as the formula

$$Y_{ij} = AB_iC_j$$

and the transfer rule

lower case letter = log of capital letter

shows for multiplicative fits *involving only positive* A, B_i and C_j. Their importance here lies not in this twin-relation but in their facility for generalization.

Not only

$$A + B_iC_j$$

but

$$B_iC_j + D_iE_j$$

and

$$A + B_iC_j + D_iE_j$$

offer convenient generalizations of simple multiplicative models, conveniently described as higher-rank models. (We find using this general-sounding term fairly freely for a somewhat special-appearing class reasonable, in part because the twin here

$$a(b_i + c_j)(d_i + e_j)$$

has not, at least as yet, often proved to be a stage of description that was helpful on our way to understanding.)

We can go on easily to still higher complexities, to stocks in which we sum still more terms of the form

(function of row) × (function of column).

3. Comments on EHR analysis.

A number of comments about the use of EHR and the corresponding stocks deserve attention here. The most important have to do with re-expression and re-presentation.

Re-expression can greatly influence the satisfactoriness with which a well-selected example from a stock of a given kind describes the data. We are familiar with this in circumstances where we understand, in detail, what is happening. We tend to forget that it is almost equally likely to be so when we face less-understood (perhaps still-impenetrable) situations in a very empirical mood.

If we were given, for a collection of cylinders, volumes, cross-sectional areas, and lengths, we would shrink from anyone who proposed to fit volumes with

(a function of cross-section) PLUS (a function of length)

since we would recognize the need for TIMES instead of PLUS. Even in this case, we might not stop to think that, if we worked with log volume instead of volume -- a very simple re-expression -- we could use the additive broad stock very effectively.

If we had data on blood pressures, I fear we would be much less likely to shun the PLUS analysis of raw blood pressures, much less likely to pounce on the advantages of a PLUS analysis of *log* blood pressure, though such an advantage would be there. And as we move toward even less understood data, we are even more likely to "miss the boat" when re-expression would help. There is no intrinsic reason for this; we have only failed to learn to take advantage of our opportunities.

The issue of re-presentation, discussed by Breckenridge above, is of a very different kind. Where re-expression sought for us a way to find more useful fits, more useful by doing a better job of fitting, we now are trying to do a more useful job of looking at the exact same fit. As a simple example, consider a fit

$$2A_iB_i + 2C_iD_i$$

which can also be written

$$(A_i + C_i)(B_i + D_i) + (A_i - C_i)(B_i - D_i)$$

(Notice carefully that these two forms are algebraically identical, as can easily be seen by multiplying out the second form). Here there is no question of changing fit, only of rewriting it.

If we are to compare the results of such a fit applied to two or more sets of data, we badly need to seek out a distinguished re-presentation of each fit, at least so that the results will be conveniently comparable. If one looks like

$$2A_iB_i + 2C_iD_i$$

when the other looks like

$$(A_i + C_i)(B_i + D_i) + (A_i - C_i)(B_i - D_i)$$

we may miss an instance of a striking resemblance, something we ought take only the least possible chance of doing.

Another way in which re-presentation can be important arises when we can find a re-presentation, say,

$$E_iF_i + G_iH_i$$

in which one factor, say F_i, is very nearly constant. This offers us the opportunity to try a less general stock, say

$$E_i^* + G_iH_i$$

with E_i^* approximately given by E_i times the (nearly) common value of F_i.

Empirical fits with broad stocks need not -- and often should not -- be thought of as ends in themselves. Often they play important roles in leading us to simpler fits, simpler fits which

may or may not gain a more or less normative character.

4. The structure of the Coale-Trussell models.

Let us look next at the internal structure of what might be thought of, by some at least, as almost the antithesis of the higher-rank broad stocks we have been discussing. The Coale-Trussell model schedules are traditionally thought of as involving one constant, m, and one variable function of age a, $G(a)$, in the form

$$f(a) = G(a)\,n(a)\,e^{m\,v(a)}$$

where $n(a)$ and $v(a)$ are fixed functions of the age, a.

Once we think of dealing with a single population at several dates (or in several cohorts) or once we think of dealing with several populations, we need to subscript m and $G(a)$. We may as well also subscript a, since we will only be using discrete age-ranges. This gives us

$$f_i(a_j) = G_i(a_j)\,n(a_j)\,e^{m_i\,v(a_j)}$$

and, once we take the (natural) logarithm,

$$\log f_i(a_j) = \log n(a_j) + \log G_i(a_j) + m_i\,v(a_j)$$

which is of the form

$$y_{ij} = K_j + C_i D_j + E_i L_j$$

where K_j and L_j are fixed.

This is now obviously a special case of

$$B_j + C_i D_j + E_i F_j$$

an often useful, but special case of the rank-3 stock

$$A_i B_j + C_i D_j + E_i F_j$$

Thus there is no necessary antithesis between such models and empirical higher-rank analysis. There may well be a difference in purposes and in style. If we thought of the empirical higher-rank analysis as an end in itself, a vast gap might indeed open up between the approaches. But if, as we ought, we think of such analyses as the first step, in which the regular behavior of the data is to be encompassed as thoroughly as we can (going to still higher rank when necessary), so that we are ready to proceed to seek out as great a simplification of the EHR fit as we believe the data and our purposes, combined, will sustain, there will be no antithesis -- and the gap may be very small.

It need not happen that, as the Danes are reputed to put it, we "fall with our nose in the butter". The effective fewer-constant fits, if they exist, may not be such that they can be found in such a way; it may not be possible to convert them into higher-rank form by re-expressing the response. When this does happen, we will have to work with the facts as they are, but we should not, I would argue, accept that it has happened without careful inquiry.

5. Comparison of the two models.

A few words about the detailed differences between Breckenridge's EHR analysis and the Coale-Trussell schedules are in order. The first major difference is in the chosen response, between

fertility at an age (for an age-interval)

and

accumulated fertility up to an age-cut
(as a fraction of total)

The fact that one takes the logarithm of the former, but the folded square-root of the latter is also important, but perhaps not as important.

Beyond this, the question of how complicated -- or simple -- a stock one uses (how broad or how narrow) is mainly a matter of detailed purposes. (I argue strongly that the practical way to begin is to fit the broad stock, going on then to whatever degree of reduction is appropriate to the combination of data behavior and our purposes.)

What are the main issues of choice in this situation? I believe that the purposes toward which the Coale-Trussell schedules are directed combine, to various degrees, those typical of

- descriptive models,
- separating models, and
- transferring models,

with decreasing emphasis as we move down the list. (Ansley Coale chose to emphasize the first two of these, in an independent assessment.)

For the first purpose, description, we want to (a) make our fit to the diversity of the real world as good as we can, subject to (b) holding the number of parameters to a minimum. For this purpose, it should not be important whether we work with fertility in age ranges or with accumulated fertility. Equally it should not matter what re-expression proves to be useful.

The analysis suggested in the closing paragraphs of Breckenridge's paper, in which empirical higher-rank analysis would be applied to data from a wide variety of countries (and time periods), would be a natural first step in an EHR search for just what description would be most useful. To be fully effective, such an analysis ought to explore the advantages, not only of age-range vs. accumulation, but also of varied re-expressions of each.

The absence of effective, experience-tested techniques for guiding the exploration of re-expression in such situations is to be regretted, but we must start to learn somewhere.

Once we understand clearly both how we can do relatively very well with both broad-stock and narrow-stock fits, it will be time to ask how well the results serve our needs as separating and transferring models. Then we can sensibly consider what changes in the structure of the empirically best-fitting models it is wise or reasonable to make in order to do better in separating and transferring.

We ought, in Student's words, plan to "use all the allowed principles of witchcraft".

6. References.

P. W. Bridgman 1927. *The Logic of Modern Physics*, MacMillan, New York.

Harold Jeffreys 1929. Random and systematic arrangements. *Biometrika* **31**: 1-8.

R. D. Luce and J. W. Tukey 1964. Simultaneous and conjoint measurement. *J. Math. Psych.* 1-27.

F. Mosteller and J. W. Tukey 1977. *Data Analysis and Regression: a second course in statistics.* Addison-Wesley Publishing Company, Reading, Massachusetts. Especially page 302ff.

"Student" 1937. Comparison between balanced and random arrangements of field plots. *Biometrika* **29**: 363-379. See page 365.